A Concise Tutorial on Approximate Message Passing

Qiuyun Zou and Hongwen Yang

Abstract—High-dimensional signal recovery of standard linear regression is a key challenge in many engineering fields, such as, communications, compressed sensing, and image processing. The approximate message passing (AMP) algorithm proposed by Donoho *et al* is a computational efficient method to such problems, which can attain Bayes-optimal performance in independent identical distributed (IID) sub-Gaussian random matrices region. A significant feature of AMP is that the dynamical behavior of AMP can be fully predicted by a scalar equation termed station evolution (SE). Although AMP is optimal in IID sub-Gaussian random matrices, AMP may fail to converge when measurement matrix is beyond IID sub-Gaussian. To extend the region of random measurement matrix, an expectation propagation (EP)-related algorithm orthogonal AMP (OAMP) was proposed, which shares the same algorithm with EP, expectation consistent (EC), and vector AMP (VAMP). This paper aims at giving a review for those algorithms. We begin with the worst case, i.e., least absolute shrinkage and selection operator (LASSO) inference problem, and then give the detailed derivation of AMP derived from message passing. Also, in the Bayes-optimal setting, we give the Bayes-optimal AMP which has a slight difference from AMP for LASSO. In addition, we review some AMP-related algorithms: OAMP, VAMP, and Memory AMP (MAMP), which can be applied to more general random matrices.

Index Terms—Standard linear regression, message passing, expectation propagation, state evolution.

I. Introduction

We focus on the sparse signal recovery of the standard linear regression

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ is the sparse signal to be estimated, $\mathbf{H} \in \mathbb{R}^{M \times N}(M \ll N)$ is the measurement matrix which is perfectly known beforehand, $\mathbf{n}$ is the additive white Gaussian noise with zero mean and covariance σ_w^2, and $\mathbf{y} \in \mathbb{R}^M$ is the observation. In the existing works, the sparse signal can be divided into two kinds: one is that $\mathbf{x}$ is k-sparsity but without true distribution, i.e., only k elements of $\mathbf{x}$ being non-zero, and the other is that $\mathbf{x}$ is drawn from a specific distribution with sparsity pattern, such as Bernoulli-Gaussian (BG) distribution. Throughout, we focus on the large system limit, in which the dimensions of system tend to infinity $(M, N) \to \infty$ but the ratio $\alpha = \frac{M}{N}$ is fixed. At the worst case, where prior and likelihood function are both unknown, this sparse inference

Q. Zou and H. Yang are with Beijing University of Posts and Telecommunications, Beijing 100876, China (email: qiuyun.zou@bupt.edu.cn; yanghong@bupt.edu.cn).

The Matlab code of this paper is available in https://github.com/QiuyunZou/AMPTutorial.

problem can be formalized as a least absolute shrinkage and selection operator (LASSO) [1] inference problem

$$\hat{\mathbf{x}}_{\text{LASSO}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{2}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are ℓ_1, ℓ_2 norm, respectively, and $\lambda \geq 0$ is the parameter of regularization that balances the sparsity and error of solution. The inference problem above is also known as basis pursuit de-noising (BPDN) inference. Such problem has a mass of applications in many fields such as compressed sensing [1], [2], [3], [4], [5], [6], image processing [7], [8], and sparse channel estimation in wireless communications etc.

To solve the LASSO inference problem, there are many kinds of algorithms. For example,

- **Convex relaxation**. LASSO inference problem is a compound optimization problem involving a smooth function and a non-smooth function such as ℓ_1 norm regularization. There are a mass of algorithms for compound optimization problem such as sub-gradient method, proximal gradient descent, also known as iterative soft threshold algorithm (ISTA) [9], Newton acceleration algorithm, and alternating direction method of multiplies (ADMM) [10], etc. Among them, ADMM alternatively optimizes the objective function containing quadric error and the objective function involving ℓ_1 norm regularization.

- **Greedy algorithm**. A kind of alternative method refers to greedy algorithms [11] in compressed sensing, such as, match pursuit (MP), orthogonal match pursuit (OMP) [12], and subspace pursuit (SP) [13], etc. In those greedy algorithms, they make a 'hard' decision based upon some locally optimal optimization criterion. All of those methods can be regarded as a variant of least square. The basic ideal of them is to approximate the signal of interest by selecting the atom or sub-hyperplane from measurement matrix that best matches the residual error of each iteration. Among them, MP projects the residual error of each iteration onto a specific atom, while OMP projects the residual error of each iteration onto a sub-hyperplane from measurement matrix.

- **Bayesian estimation**. The Bayesian estimator [14, Chapter 10] is a kind of algorithm which aims at minimizing the Bayes loss function. According to different Bayes risk functions, the Bayesian estimator can be generally divided into minimum mean square error (MMSE) and maximum a posterior (MAP). In fact, the exact MMSE or MAP is NP-hard problem in general cases. However, there are some algorithms which implement the exact Bayesian estimator iteratively. Among them, the
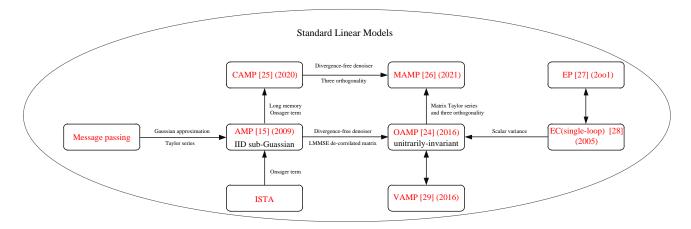
Fig. 1. The relations between the message passing based algorithms in standard linear regression inference problem.

approximate message passing (AMP) [15] algorithm, the main focus in this paper, is a celebrated implementation of Bayes estimation. By postulated posterior/MMSE, in which the postulated prior and likelihood function are different from true ones, AMP can provide the exact sparse solution to LASSO inference problem using Laplace method of integration. In general, we call the algorithm which relies on Bayesian formula as Bayesian algorithms.

On the other hand, in the *Bayes-optimal setting* (may $M \geq N$) where both prior and likelihood function are known, the MMSE and MAP give a much better performance than convex relaxation. However, due to high-dimensional integration, the exact MMSE is hard to obtain. Fortunately, some existing works [16] showed that AMP can achieve the Bayes-optimal MSE performance but with affordable complexity in independent identical distributed (IID) sub-Gaussian random measurement matrices region [17]. For convenience, we depict Fig. 1 to show the relations between AMP and its related algorithms. The AMP derives from the message passing [18] algorithm in coding theory, which is also known as belief propagation [19] in computer science or cavity method [20] in statistic mechanics. The AMP algorithm is closely related to the Thouless-Anderson-Palmer (TAP) [21] equations which is used to approximate marginal moments in large probabilistic models. In [22], the first AMP algorithm was proposed for the code division multiple access (CDMA) multi-user detection problem. A significant feature of AMP algorithm is that the dynamic of AMP can be fully predicted by a scalar equation termed state evolution (SE) [16], which is perfectly agree with the fixed point of the exact MMSE estimator using replica method [23]. The AMP algorithm is also related to ISTA, the difference between them is the *Onsager* term, which leads to AMP more faster than ISTA but it doesn't change its fixed points. As the measurement matrix is beyond IID sub-Gaussian region, AMP methods often fail to converge. Beyond IID sub-Gaussian region, the orthogonal AMP (OAMP) [24] can be applied to more general unitarily-invariant matrices via the LMMSE de-correlated matrix and divergence-free denoiser, but it should pay more computational complexity due to the matrix inversion. To balance the complexity and region of

random measurement matrix, recently, some long memory algorithms such as convolutional AMP (CAMP) [25], and memory AMP [26] were proposed. Different from OAMP, CAMP only modifies the Onsager term of AMP. The Onsager term of CAMP includes all proceeding messages to ensure the Gaussianity of input signal of denoiser. However, CAMP may fail to converge in the case of large condition number. Following CAMP and OAMP, the MAMP algorithm applies finite terms of matrix Taylor series to approximate matrix inversion of OAMP and involves all previous messages to ensure three orthogonality.

Another efficient algorithm related to AMP is called expectation propagation (EP) [27]. EP is earlier than AMP, which approximates the factorable factors by choosing a distribution from Gaussian family via minimizing Kullback-Leibler (KL) divergence. Some EP-related methods refer to expectation consistent (EC) [28, Appendix D] (single-loop), OAMP [24], and vector AMP (VAMP) [29]. They were proposed independently in different manners but share the same algorithm. Actually, EP/EC (single-loop) have a slight difference from OAMP/VAMP, since EP/EC has the element-wise variances and they can be reduced to OAMP/VAMP by taking the mean operation for element-wise variance. Among them, EC approximation is based on the minimum Gbiss free energy. It means that those methods can be regarded as an example of solving the fixed point of Gbiss free energy. Almost at the same time as OAMP, the VAMP was proposed using a EP-type message passing and the dynamic of VAMP was rigorously analyzed in [29]. Recently, [30] proved that VAMP and AMP have identical fixed points in their state evolutions in their overlapping random matrices. We also note that under the mismatch case [31], where the prior and likelihood function applied to the inference problem are different from the true prior and likelihood function, the AMP as well as its related algorithms may not converge although the corresponding SE converges to a fixed point predicted by replica method. Actually, AMP for LASSO is one case of mismatched model, but its convergence is guaranteed due to convex nature of LASSO [32]. The failure of AMP can occur when the mismatched models are defined by *non-convex* cost function [33].

Besides, there are some algorithms that extend AMP to more general models beyond standard liner model. In [34], a generalized AMP (GAMP) algorithm was proposed for generalized linear model which allows an arbitrary row-wise mapping. A concise derivation of GAMP using EP projection can be found in [35], [36]. Further, Park *et al* [37] developed bilinear GAMP (BiG-AMP) which extends the GAMP algorithm to bilinear model in which both the signal of interest and measurement matrix are unknown. Recent works showed that the BiG-AMP can be obtained by Plefka-Georges-Yedidia method [38], [39]. Following VAMP, [40], [41] developed a generalized linear model VAMP (GLM-VAMP) algorithm by constructing an equivalent linear model. Compared to GAMP, GLM-VAMP can be applied to more general random matrices but needs to pay more computational complexity. Similar to GLM-VAMP, a generalized version of MAMP was proposed in [42]. Beyond single-layer model, some extensions of AMP in multi-layer regions can be found in [43], [44], [45], [46]. However, those algorithms are out of the scope of this paper.

Although AMP and its related methods have attracted a lot of attention in many engineering fields, there still isn't a tutorial that gives a clear line to summarize them and provides concise derivations. That is the purpose of this paper. For that purpose, we begin with the LASSO inference problem, which is original goal of AMP. By Laplace method of integration, the LASSO inference problem can be converted into the limit of postulated MMSE estimator. Using factor graph representation and message passing, we give the detailed derivation of AMP for LASSO. And then we move to the Bayes-optimal setting, which is more attractive and common in some engineering fields, such as wireless communications. Beyond IID sub-Gaussian random matrices, we review several extensions of AMP: OAMP, VAMP, and MAMP, and illustrate their relations and differences.

Notations: Throughout, we use $\mathbf{x}$ and $\mathbf{X}$ to denote column vector and matrix, respectively. $(\cdot)^{\mathrm{T}}$ denotes transpose operator such as $\mathbf{X}^{\mathrm{T}}$. $\mathrm{Tr}(\mathbf{A})$ denotes the trace of square matrix $\mathbf{A}$. $\overset{\mathrm{a.s.}}{=}$ means equal almost sure. Given the original signal $\mathbf{x}$ and its estimator $\hat{\mathbf{x}}$, the normalized MSE (NMSE) is defined as $\mathrm{NMSE}(\mathbf{x}) = \frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2}$ with $\|\cdot\|_2$ being ℓ_2 norm. We apply $\mathcal{N}(x|a, A)$ to denote a Gaussian probability density function with mean a and variance A described by:

$$\mathcal{N}(x|a, A) = \frac{1}{\sqrt{2\pi A}} \exp\left[-\frac{(x-a)^2}{2A}\right].$$

$\mathcal{BG}(\mu, \rho)$ is a Bernoulli Gaussian distribution: $\mathcal{BG}(\mu, \rho) = \rho\mathcal{N}(x|\mu, \rho^{-1}) + (1-\rho)\delta(x)$.

II. APPROXIMATE MESSAGE PASSING

A. Iterative Soft Threshold Algorithm

Before introducing AMP algorithm, we first review a AMP related algorithm: ISTA. Recalling that the term $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y}-\mathbf{Hx}\|_2^2$ in (2) is continuous and derivative while the second term $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ is not differentiable at $\mathbf{x} = \mathbf{0}$. The minimization of $f(\mathbf{x})$ can be achieved by gradient descent

$$\hat{\mathbf{x}}^{(t)} = \arg\min_{\mathbf{x}} \frac{1}{2\alpha_t}\|\mathbf{x} - (\hat{\mathbf{x}}^{(t-1)} - \alpha_{t-1}\nabla f(\hat{\mathbf{x}}^{(t-1)}))\|_2^2, \quad (3)$$
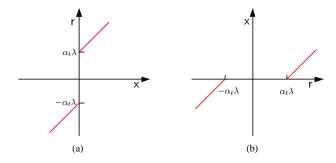

Fig. 2. (a) x-z coordinate axis; (b) z-x coordinate axis.

where α_t is step size and $\hat{\mathbf{x}}^{(t)}$ is the estimator of $\mathbf{x}$ at t-iteration. Adding ℓ_1 norm regularization, (3) becomes

$$\hat{\mathbf{x}}^{(t)} = \arg\min_{\mathbf{x}} \frac{1}{2\alpha_t}\|\mathbf{x} - (\hat{\mathbf{x}}^{(t-1)} - \alpha_{t-1}\nabla f(\hat{\mathbf{x}}^{(t-1)}))\|_2^2 + \lambda\|\mathbf{x}\|_1.$$
$$(4)$$

Defining $\mathbf{r}^{(t)} = \hat{\mathbf{x}}^{(t)} - \alpha_t\nabla f(\hat{\mathbf{x}}^{(t)}) = \hat{\mathbf{x}}^{(t)} + \alpha_t\mathbf{H}^{\mathrm{T}}(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)})$, the equation above becomes

$$\forall i: \quad \hat{x}_i^{(t)} = \arg\min_{x_i} \left\{ \frac{1}{2\alpha_t}(x_i - r_i^{(t-1)})^2 + \lambda|x_i| \right\}. \quad (5)$$

Zeroing the gradients w.r.t. x_i yields $r_i^{(t-1)} = \hat{x}_i^{(t)} + \alpha_t\lambda\mathrm{sign}(\hat{x}_i^{(t)})$. Then swapping the axes (see Fig 2) gets

$$\hat{x}_i^{(t)} = \mathrm{sign}(r_i^{(t-1)})\max(|r_i^{(t-1)}| - \alpha_t\lambda, 0). \quad (6)$$

Totally, the ISTA is summarized as

$$\mathbf{r}^{(t)} = \hat{\mathbf{x}}^{(t)} + \alpha_t\mathbf{H}^{\mathrm{T}}(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}), \quad (7a)$$
$$\hat{\mathbf{x}}^{(t+1)} = \mathrm{sign}(\mathbf{r}^{(t)})\max(|\mathbf{r}^{(t)}| - \alpha_t\lambda, 0). \quad (7b)$$

To in line with AMP, let's define $\mathbf{z}^{(t)} = \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}$ and $\eta(\mathbf{r}^{(t)}, \alpha_t\lambda) = \mathrm{sign}(\mathbf{r}^{(t)})\max(|\mathbf{r}^{(t)}|, \alpha_t\lambda)$. The ISTA algorithm can be written as

$$\mathbf{z}^{(t)} = \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}, \quad (8a)$$
$$\hat{\mathbf{x}}^{(t+1)} = \eta(\hat{\mathbf{x}}^{(t)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t)}, \lambda), \quad (8b)$$

where the step size α_t is set to $\alpha_t = 1$. However, in practical, α_t may cause the algorithm to diverge and actually $\alpha_t \in [0.1, 0.35]$ is appropriate in our simulation. The complexity of ISTA is dominated by the matrix multiplication with the cost of $\mathcal{O}(MN)$. However, the convergence speed of ISTA is too slow. To improve the convergence speed of ISTA, the fast ISTA (FISTA) [7] was proposed. The FISTA is beyond the scope of this paper. We only post it as below

$$\hat{\mathbf{x}}^{(t)} = \hat{\mathbf{x}}^{(t-1)} + \frac{t-2}{t+1}(\hat{\mathbf{x}}^{(t-1)} - \hat{\mathbf{x}}^{(t-2)}), \quad (9a)$$
$$\mathbf{z}^{(t)} = \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}, \quad (9b)$$
$$\hat{\mathbf{x}}^{(t+1)} = \eta(\hat{\mathbf{x}}^{(t)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t)}). \quad (9c)$$

Comparing FISTA in (9) with ISTA in (7), the difference between them is that the term $\hat{\mathbf{x}}^{(t)}$ is constructed from two previous results.

B. AMP for LASSO

The AMP algorithm [15] posted below is related to ISTA

$$\mathbf{z}^{(t)} = \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}$$
$$+ \frac{1}{\alpha}\mathbf{z}^{(t-1)}\left\langle \eta'_{t-1}(\hat{\mathbf{x}}^{(t-1)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t-1)})\right\rangle, \quad (10a)$$
$$\hat{\mathbf{x}}^{(t+1)} = \eta_t(\hat{\mathbf{x}}^{(t)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t)}), \quad (10b)$$

where $\langle\cdot\rangle$ is empirical mean such as $\langle\mathbf{x}\rangle = \frac{1}{N}\sum_{i=1}^{N}x_i$ and $\eta'_{t-1}(\mathbf{r})$ is the partial derivation of $\eta_{t-1}(\cdot)$ w.r.t. $\mathbf{r}$.

Compared to ISTA algorithm in (8), the key difference between AMP and ISTA is the *Onsager* term $\frac{1}{\alpha}\mathbf{z}^{(t-1)}\left\langle\eta'_{t-1}(\mathbf{x}^{(t-1)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t-1)})\right\rangle$. This term can improve the convergence speed of ISTA but does not change its fixed point. Essentially, this term ensures that the input $\mathbf{r}^{(t)}$ of denoiser $\eta(\cdot)$ can be expressed as the original signal adding an additive Gaussian noise (Gaussianity, see Fig. 3) and it leads to faster convergence than ISTA. As shown in Fig. 4, we compare per-iteration NMSE behavior of the AMP with ISTA and FISTA. From Fig. 4, we can see that AMP converges with $t = 18$ iterations which is far small than FISTA ($t = 108$) and ISTA ($t = 235$). Be aware, in ISTA, one should adjust the step size α_t to ensure the convergence but the step size is unnecessary to AMP. In addition, an appropriate step size ensures the algorithm to converge but does not change the fixed point. The below is the detailed derivation to obtain AMP for LASSO inference problem.

As shown in [47, Appendix D], [48], the LASSO inference problem can be expressed as the limit of the postulated MMSE estimator using Laplace method of integration

$$\hat{\mathbf{x}} = \lim_{\beta\to\infty}\int \mathbf{x}\,\underbrace{\frac{1}{Z_\beta^{\mathrm{pos}}}\exp\left[-\beta\left(\frac{1}{2}\|\mathbf{y}-\mathbf{Hx}\|_2^2 + \lambda\|\mathbf{x}\|_1\right)\right]}_{q(\mathbf{x}|\mathbf{y})}\,\mathrm{d}\mathbf{x}$$
$$= \arg\min_{\mathbf{x}}\left\{\frac{1}{2}\|\mathbf{y}-\mathbf{Hx}\|_2^2 + \lambda\|\mathbf{x}\|_1\right\}, \quad (11)$$

where Z_β^{pos} is the normalization constant. Using Bayes' rules, the postulated posterior $q(\mathbf{x}|\mathbf{y})$ in (11) is expressed as

$$q(\mathbf{x}|\mathbf{y}) = \frac{1}{q(\mathbf{y})}q(\mathbf{x})q(\mathbf{y}|\mathbf{x}),$$
$$q(\mathbf{x}) = \frac{1}{Z_\beta^{\mathrm{pri}}}\exp(-\beta\lambda\|\mathbf{x}\|_1),$$
$$q(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\beta^{\mathrm{lik}}}\exp\left(-\frac{\beta}{2}\|\mathbf{y}-\mathbf{Hx}\|_2^2\right),$$

where $q(\mathbf{y})$, Z_β^{pri}, and Z_β^{lik} are normalization constants, and $q(\mathbf{x})$ is the postulated prior while $q(\mathbf{y}|\mathbf{x})$ is the postulated likelihood function. The postulated likelihood function can also be formalized as $q(\mathbf{y}|\mathbf{x}) = \prod_{a=1}^{M}\mathcal{N}(y_a|\sum_{i=1}^{N}h_{ai}x_i, \frac{1}{\beta})$.

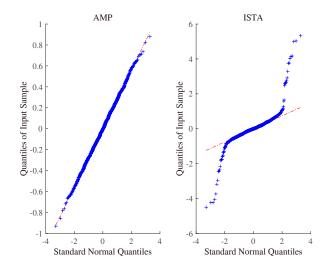The factor graph of postulated posterior defined in (11) is depicted in Fig. 5. For basis of factor graph and message

Fig. 3. QQplot comparing the distribution of input error $\mathbf{r}^{(t)} - \mathbf{x}$ of AMP and ISTA at $t = 5$. The system setups are similar to that of Fig. 4. The blue points match the red line better, the closer the input error is to the Gaussian distribution. Notice that the input error of AMP remains Gaussianity due to Onsager term.
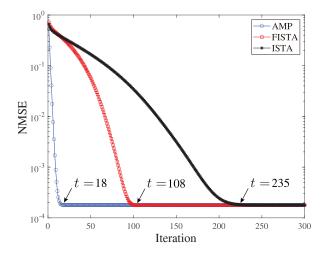

Fig. 4. Comparison of AMP, FISTA, and ISTA for LASSO inference problem. $\mathbf{H}$ has IID Gaussian entries with zero mean and $1/M$ variance. $N = 1024$, $M = 512$, $\alpha = \frac{M}{N} = \frac{1}{2}$, and $\lambda = 0.05$. SNR $= 1/\sigma_w^2 = 50$dB. $\mathbf{x}$ has IID BG entries following $\mathcal{BG}(0, 0.05)$. The step size $\alpha_t = 0.35$ and $\alpha_t = 0.2$ are applied to ISTA and FISTA, respectively.

passing, we suggest [49, Chapter 2] for more details. From this figure, the messages are addressed as

$$\mu_{i\to a}^{(t+1)}(x_i) \propto e^{-\beta\lambda|x_i|}\prod_{b\neq a}^{M}\mu_{i\leftarrow b}^{(t)}(x_i), \quad (12a)$$

$$\mu_{i\leftarrow a}^{(t)}(x_i) \propto \int q(y_a|\mathbf{x})\prod_{j\neq i}^{N}\mu_{j\to a}^{(t)}(x_j)\mathrm{d}\mathbf{x}_{\setminus i}, \quad (12b)$$

where $\mathbf{x}_{\setminus i}$ is $\mathbf{x}$ expect x_i, $\mu_{i\to a}^{(t+1)}(x_i)$ is the message from variable node x_i to factor node $q(y_a|\mathbf{x})$, $\mu_{i\leftarrow a}^{(t)}(x_i)$ is the message in opposite direction at t-iteration, and superscript t denotes the number of iteration. It is worth noting that at t-iteration, the marginal posterior $\mathcal{P}(x_i|\mathbf{y})$ can be approximated
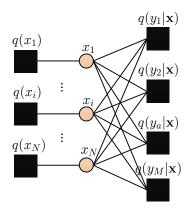
Fig. 5. Factor graph of postulated posterior $q(\mathbf{x}|\mathbf{y})$ defined in (11), where $q(x_i) \propto e^{-\beta|x_i|}$ and $q(y_a|\mathbf{x}) = \mathcal{N}(y_a|\sum_{i=1}^{N} h_{ai}x_i, 1/\beta)$. The square denotes the factor node (e.g. $q(x_i)$) while the circle denotes the variable node (e.g. x_i). The messages delivers between factor nodes and variable nodes via their edges.

by

$$\hat{q}^{(t+1)}(x_i|\mathbf{y}) = \frac{e^{-\beta\lambda|x_i|} \prod_{a=1}^{M} \mu_{i\leftarrow a}^{(t)}(x_i)}{\int e^{-\beta\lambda|x_i|} \prod_{a=1}^{M} \mu_{i\leftarrow a}^{(t)}(x_i)\mathrm{d}x_i}, \quad (13)$$

while the mean of the approximated posterior $\hat{q}^{(t+1)}(x_i|\mathbf{y})$ will serve as an approximation of MMSE estimator.

To reduce the complexity of sum-product message passing shown in (12), we first simplify the message $\mu_{i\leftarrow a}^{(t)}(x_i)$ as below

$$\mu_{i\leftarrow a}^{(t)}(x_i)$$
$$\propto \int_{\mathbf{x}\backslash i} \int_{z_a} q(y_a|z_a)\delta(z_a - \sum_{k=1}^{N} h_{ak}x_k)\mathrm{d}z_a \prod_{j\neq i}^{N} \mu_{j\rightarrow a}^{(t)}(x_j)\mathrm{d}\mathbf{x}\backslash i$$
$$\propto \int_{z_a} q(y_a|z_a)\mathbb{E}\left\{\delta\left(z_a - \sum_{j\neq i} h_{aj}x_j - h_{ai}x_i\right)\right\}\mathrm{d}z_a, \quad (14)$$

where the expectation is over $\prod_{j\neq i}^{N} \mu_{j\rightarrow a}^{(t)}(x_j)$. We define random variable (RV) $\zeta_{i\leftarrow a}^{(t)}$ associated with z_a and $\xi_{j\rightarrow a}^{(t)}$ following $\mu_{j\rightarrow a}^{(t)}(x_j)$ associated with x_j. Denote the mean and variance of $\xi_{j\rightarrow a}^{(t)}$ as $\hat{x}_{j\rightarrow a}^{(t)}$ and $\hat{v}_{j\rightarrow a}^{(t)}/\beta$, respectively. From (14), as the dimension N tends to infinity, using central limit (CLT) theorem the RV $\zeta_{i\leftarrow a}^{(t)}$ converges to a Gaussian RV with mean and variance

$$\mathbb{E}\{\zeta_{i\leftarrow a}^{(t)}\} = Z_{i\leftarrow a}^{(t)} + h_{ai}x_i, \quad \mathrm{Var}\{\zeta_{i\leftarrow a}^{(t)}\} = \frac{1}{\beta}V_{i\leftarrow a}^{(t)}, \quad (15)$$

where

$$Z_{i\leftarrow a}^t = \sum_{j\neq i} h_{aj}\hat{x}_{j\rightarrow a}^t, \quad V_{i\leftarrow a}^t = \sum_{j\neq i} |h_{aj}|^2 \hat{v}_{j\rightarrow a}^t. \quad (16)$$

Based on this Gaussian approximation, the term $\mathbb{E}\{\delta(z_a - \sum_{j\neq i} h_{aj}x_j - h_{ai}x_i)\}$ in (14) is replaced by $\mathcal{N}(z_a|h_{ai}x_i +$

$Z_{i\leftarrow a}^{(t)}, \frac{1}{\beta}V_{i\leftarrow a}^{(t)})$. By Gaussian reproduction lemma[1], the message $\mu_{i\leftarrow a}^{(t)}(x_i)$ is approximated as

$$\mu_{i\leftarrow a}^{(t)}(x_i) \propto \mathcal{N}\left(0|y_a - h_{ai}x_i - Z_{i\leftarrow a}^{(t)}, \frac{1}{\beta}(1 + V_{i\leftarrow a}^{(t)})\right)$$
$$\propto \mathcal{N}\left(x_i|\frac{y_a - Z_{i\leftarrow a}^t}{h_{ai}}, \frac{1 + V_{i\leftarrow a}^{(t)}}{\beta|h_{ai}|^2}\right). \quad (17)$$

In the sequel, the mean and variance of $\mu_{i\leftarrow a}^{(t)}(x_i)$ are defined and evaluated as

$$\hat{x}_{i\leftarrow a}^{(t)} = \frac{y_a - Z_{i\leftarrow a}^{(t)}}{h_{ai}}, \quad \hat{v}_{i\leftarrow a}^{(t)} = \frac{1 + V_{i\leftarrow a}^{(t)}}{\beta|h_{ai}|^2}. \quad (18)$$

Be aware the equation (17) is mathematically invalid as $h_{ai} = 0$. However, in the rest of this section we will show that several zero elements in $\mathbf{H}$ has no effect on the final result.

Let's move to calculate the message $\mu_{i\rightarrow a}^{(t+1)}(x_i)$ in (12) based on the approximated result above. Applying Gaussian reproduction property, the term $\prod_{b\neq a}^{M} \mu_{i\leftarrow b}^{(t)}(x_i)$ in $\mu_{i\rightarrow a}^{(t+1)}(x_i)$ is proportion to

$$\prod_{b\neq a}^{M} \mu_{i\leftarrow b}^{(t)}(x_i) \propto \mathcal{N}(x_i|r_{i\rightarrow a}^{(t)}, \Sigma_{i\rightarrow a}^{(t)}), \quad (19)$$

where

$$\Sigma_{i\rightarrow a}^{(t)} = \left(\sum_{b\neq a} \frac{|h_{bi}|^2}{1 + V_{i\leftarrow b}^{(t)}}\right)^{-1}, \quad (20)$$

$$r_{i\rightarrow a}^{(t)} = \Sigma_{i\rightarrow a}^t \sum_{b\neq a} \frac{h_{bi}^*(y_b - Z_{i\leftarrow b}^{(t)})}{1 + V_{i\leftarrow b}^{(t)}}. \quad (21)$$

Note that several zero value elements in $\mathbf{H}$ have no effect on $\Sigma_{i\rightarrow a}^{(t)}$, $r_{i\rightarrow a}^{(t)}$ as well as rest parameters in the derivation of AMP.

As a result, the message $\mu_{i\rightarrow a}^{(t+1)}(x_i)$ is approximated as the product of a Laplace prior and a Gaussian likelihood function

$$\mu_{i\rightarrow a}^{(t+1)}(x_i) = \frac{1}{Z_\beta} e^{-\beta\lambda|x_i|} \mathcal{N}(x_i|r_{i\rightarrow a}^{(t)}, \Sigma_{i\rightarrow a}^{(t)}), \quad (22)$$

where Z_β is normalized constant.

For convenience, define a distribution

$$f_\beta(x; r, \Sigma) = \frac{1}{Z_\beta} \exp\left[-\beta\left(\lambda|x| + \frac{1}{2\Sigma}(x - r)^2\right)\right], \quad (23)$$

and its mean and variance

$$\mathsf{F}_\beta(x; r, \Sigma) = \int x f_\beta(x; r, \Sigma)\mathrm{d}x, \quad (24)$$

$$\mathsf{G}_\beta(x; r, \Sigma) = \int x^2 f_\beta(x; r, \Sigma)\mathrm{d}x - |\mathsf{F}_\beta(x; r, \Sigma)|^2. \quad (25)$$

The mean and variance of the message $\mu_{i\rightarrow a}^{(t+1)}(x_i)$ are represented as

$$\hat{x}_{i\rightarrow a}^{(t+1)} = \mathsf{F}_\beta(x_i; r_{i\rightarrow a}^{(t)}, \Sigma_{i\rightarrow a}^{(t)}), \quad (26)$$

$$\hat{v}_{i\rightarrow a}^{(t+1)} = \beta\mathsf{G}_\beta(x_i; r_{i\rightarrow a}^{(t)}, \Sigma_{i\rightarrow a}^{(t)}). \quad (27)$$

[1] $\mathcal{N}(x|a, A)\mathcal{N}(x|b, B) = \mathcal{N}(x|c, C)\mathcal{N}(0|a - b, A + B)$ with $C = (A^{-1} + B^{-1})^{-1}$ and $c = C\left(\frac{a}{A} + \frac{b}{B}\right)$

Recalling the approximated posterior $\hat{q}^{(t+1)}(x_i|\mathbf{y})$ in (13), we define

$$\Sigma_i^{(t)} = \left(\sum_{a=1}^{M} \frac{|h_{ai}|^2}{1+V_{i\leftarrow a}^{(t)}}\right)^{-1}, \tag{28}$$

$$r_i^{(t)} = \Sigma_i^{(t)} \sum_{a=1}^{M} \frac{h_{ai}^*(y_a - Z_{i\leftarrow a}^{(t)})}{1+V_{i\leftarrow a}^{(t)}}. \tag{29}$$

The term $\prod_{a=1}^{M} \mu_{i\leftarrow a}^{(t)}(x_i)$ is proportion to $\mathcal{N}(x_i|r_i^{(t)}, \Sigma_i^{(t)})$. Accordingly, the mean and variance of approximated posterior $\hat{q}^{(t+1)}(x_i|\mathbf{y})$ can be denoted as

$$\hat{x}_i^{(t+1)} = \mathsf{F}_\beta(x_i; r_i^{(t)}, \Sigma_i^{(t)}), \tag{30}$$

$$\hat{v}_i^{(t+1)} = \beta\mathsf{G}_\beta(x_i; r_i^{(t)}, \Sigma_i^{(t)}). \tag{31}$$

Also define

$$Z_a^{(t)} = \sum_{i=1}^{N} h_{ai}\hat{x}_{i\rightarrow a}^{(t)} \tag{32}$$

$$V_a^{(t)} = \sum_{i=1}^{N} |h_{ai}|^2 \hat{v}_{i\rightarrow a}^{t} \approx V_{i\leftarrow a}^{(t)} \tag{33}$$

where $V_a^{(t)} = V_{i\leftarrow a}^{(t)}$ holds by ignoring infinitesimal.

Applying first-order Taylor series[2] to $\hat{x}_{i\rightarrow a}^{(t+1)}$ in (26), we have

$$\hat{x}_{i\rightarrow a}^{(t+1)} \approx \hat{x}_i^{(t+1)} + \triangle r \frac{\partial}{\partial r}\mathsf{F}_\beta(x_i; r_i^{(t)}, \Sigma_i^{(t)})$$
$$+ \triangle\Sigma \frac{\partial}{\partial\Sigma}\mathsf{F}_\beta(x_i; r_i^{(t)}, \Sigma_i^{(t)}), \tag{34}$$

where

$$\triangle\Sigma = \Sigma_{i\rightarrow a}^{(t)} - \Sigma_i^{(t)}$$
$$= \frac{\frac{|h_{ai}|^2}{1+V_a^{(t)}}}{\left(\sum_{a=1}^{M}\frac{|h_{ai}|^2}{1+V_{i\leftarrow a}^{(t)}}\right)\left(\sum_{b\neq a}^{M}\frac{|h_{bi}|^2}{1+V_{i\leftarrow b}^{(t)}}\right)}$$
$$\approx 0, \tag{35}$$

$$\triangle r = r_{i\rightarrow a}^{(t)} - r_i^{(t)}$$
$$\approx -\Sigma_i^{(t)}\frac{h_{ai}^*(y_a - Z_{i\leftarrow a}^{(t)})}{1+V_a^{(t)}}, \tag{36}$$

where we use the approximations $V_a^{(t)} = V_{i\leftarrow a}^{(t)} + O(1/N)$ and $\Sigma_i^{(t)} = \Sigma_{i\rightarrow a}^{t} + O(1/N)$ to obtain $\triangle r$. Applying the fact[3] $\frac{\partial}{\partial r}\mathsf{F}_\beta(x_i; r, \Sigma_i^{(t)})|_{r=r_i^{(t)}} = \frac{\beta}{\Sigma_i^{(t)}}\mathsf{G}_\beta(x; r_i^{(t)}, \Sigma_i^{(t)}) = \frac{\hat{v}_i^{(t+1)}}{\Sigma_i^{(t)}}$, (34) can be simplified as

$$\hat{x}_{i\rightarrow a}^{(t+1)} \approx \hat{x}_i^{(t+1)} - \frac{h_{ai}^*(y_a - Z_{i\leftarrow a}^{(t)})}{1+V_a^{(t)}}\hat{v}_i^{(t+1)}. \tag{37}$$

[2] $f(x+\triangle x, y+\triangle y) = f(x,y) + \triangle x f_x'(x,y) + \triangle y f_y'(x,y)$, where f_x' and f_y' are the partial derivation of $f(x,y)$ w.r.t. x and y, respectively.
[3] Provided that $f(x)$ is an arbitrary bounded and non-negative function and define a distribution $\mathcal{P}(x) = \frac{f(x)\mathcal{N}(x|m,v)}{\int f(x)\mathcal{N}(x|m,v)dx}$. Denote its mean and variance as $\mathbb{E}\{x\} = \int x\mathcal{P}(x)dx$ and $\text{Var}\{x\} = \int(x - \mathbb{E}\{x\})^2\mathcal{P}(x)dx$. We have $\frac{\partial\int x\mathcal{P}(x)dx}{\partial m} = \frac{\int x\frac{x-m}{v}f(x)\mathcal{N}(x|m,v)dx\cdot\int f(x)\mathcal{N}(x|m,v)dx}{[\int f(x)\mathcal{N}(x|m,v)dx]^2} - \frac{\int xf(x)\mathcal{N}(x|m,v)dx\cdot\int\frac{x-m}{v}f(x)\mathcal{N}(x|m,v)dx}{[\int f(x)\mathcal{N}(x|m,v)dx]^2} = \frac{\text{Var}\{x\}}{v}$.

Applying Taylor series to $\hat{v}_{i\rightarrow a}^{(t+1)}$ in (27), we have

$$\hat{v}_{i\rightarrow a}^{(t+1)} \approx \hat{v}_i^{(t+1)} + \triangle r\frac{\partial}{\partial r}\beta\mathsf{G}_\beta(x_i; r_i^{(t)}, \Sigma_i^{(t)}). \tag{38}$$

Combining (36) with (38) into (33) obtains

$$V_a^{(t)} = \sum_{i=1}^{N} |h_{ai}|^2\left(\hat{v}_i^{(t)} - \Sigma_i^{(t)}\frac{h_{ai}^*(y_a - Z_{i\leftarrow a}^{(t)})}{1+V_a^{(t)}}\right.$$
$$\left.\times\frac{\partial}{\partial r}\beta\mathsf{G}_\beta(x_i; r, \Sigma_i^{(t)})\right)$$
$$\approx \sum_{i=1}^{N} |h_{ai}|^2\hat{v}_i^{(t)} - \sum_{i=1}^{N}\frac{|h_{ai}|^3(y_a - Z_{i\leftarrow a}^{(t)})}{\sum_{a=1}^{M}|h_{ai}|^2}$$
$$\times\frac{\partial}{\partial r}\beta\mathsf{G}_\beta(x_i; r, \Sigma_i^{(t)})$$
$$= \sum_{i=1}^{N} |h_{ai}|^2\hat{v}_i^{(t)} + O(1/\sqrt{N})$$
$$\approx \sum_{i=1}^{N} |h_{ai}|^2\hat{v}_i^{(t)}. \tag{39}$$

Substituting (37) into (32) gets

$$Z_a^{(t)} \approx \sum_{i=1}^{N} h_{ai}\hat{x}_i^{(t)} - \sum_{i=1}^{N}\frac{|h_{ai}|^2(y_a - Z_{i\leftarrow a}^{(t-1)})}{1+V_a^{(t-1)}}\hat{v}_i^{(t)}$$
$$= \sum_{i=1}^{N} h_{ai}\hat{x}_i^{(t)} - \sum_{i=1}^{N}\frac{|h_{ai}|^2\hat{v}_i^{(t)}(y_a - Z_a^{(t-1)} + h_{ai}\hat{x}_i^{(t-1)})}{1+V_a^{(t-1)}}$$
$$\approx \sum_{i=1}^{N} h_{ai}\hat{x}_i^{(t)} - \frac{V_a^{(t)}(y_a - Z_a^{(t-1)})}{1+V_a^{(t-1)}}. \tag{40}$$

Inserting (37) into (29) yields

$$r_i^{(t)} \approx \Sigma_i^{(t)}\sum_{a=1}^{M}\frac{h_{ai}^*(y_a - Z_a^{(t)} + h_{ai}\hat{x}_i^{(t)})}{1+V_a^{(t)}}$$
$$= \hat{x}_i^{(t)} + \Sigma_i^{(t)}\sum_{a=1}^{M}\frac{h_{ai}^*(y_a - Z_a^{(t)})}{1+V_a^{(t)}}. \tag{41}$$

Up to now, the derivation of AMP for LASSO is complete. The AMP algorithm is shown in Algorithm 1.

To in line with Donoho's AMP, we still need to carry out the following simplifications using the fact $|h_{ai}|^2 = O(1/M)$

$$V_a^{(t)} = \frac{1}{M}\sum_{i=1}^{N}\hat{v}_i^{t} \triangleq V^{(t)}, \tag{43a}$$

$$Z_a^{(t)} = \sum_{i=1}^{N} h_{ai}\hat{x}_i^{(t)} - \frac{V^{(t)}(y_a - Z_a^{(t-1)})}{1+V^{(t-1)}}, \tag{43b}$$

$$\Sigma_i^{(t)} = 1 + V^{(t)} \triangleq \Sigma^{(t)}, \tag{43c}$$

$$r_i^{(t)} = \hat{x}_i^{(t)} + \sum_{a=1}^{M} h_{ai}^*(y_a - Z_a^{(t)}), \tag{43d}$$

$$\hat{x}_i^{(t+1)} = \mathsf{F}_\beta(x_i; r_i^{(t)}, \Sigma^{(t)}), \tag{43e}$$

$$\hat{v}_i^{(t+1)} = \Sigma^{(t)}\mathsf{F}_\beta'(x_i; r_i^{(t)}, \Sigma^{(t)}), \tag{43f}$$

Algorithm 1: AMP for LASSO

---

1. **Input: y, H.**
2. **Initialization:** $\hat{x}_i^{(1)} = 0$, $\hat{v}_i^{(1)} = 1$, $Z_a^{(0)} = y_a$.
3. **Output:** $\hat{\mathbf{x}}^{(T)}$.
4. **Iteration:**
for $t = 1, \cdots, T$ **do**

$$V_a^{(t)} = \sum_{i=1}^{N} |h_{ai}|^2 \hat{v}_i^{(t)} \tag{42a}$$

$$Z_a^{(t)} = \sum_{i=1}^{N} h_{ai}\hat{x}_i^{(t)} - \frac{V_a^{(t)}(y_a - Z_a^{(t-1)})}{1 + V_a^{(t-1)}} \tag{42b}$$

$$\Sigma_i^{(t)} = \left(\sum_{a=1}^{M} \frac{|h_{ai}|^2}{1 + V_a^{(t)}} \right)^{-1} \tag{42c}$$

$$r_i^{(t)} = \hat{x}_i^{(t)} + \Sigma_i^{(t)} \sum_{a=1}^{M} \frac{h_{ai}^*(y_a - Z_a^{(t)})}{1 + V_a^{(t)}} \tag{42d}$$

$$\hat{x}_i^{(t+1)} = \mathsf{F}_\beta(x_i; r_i^{(t)}, \Sigma_i^{(t)}) \tag{42e}$$

$$\hat{v}_i^{(t+1)} = \beta\mathsf{G}_\beta(x_i; r_i^{(t)}, \Sigma_i^{(t)}) \tag{42f}$$

end

Algorithm 2: Bayes-Optimal AMP

---

1. **Input: y, H**, σ_w^2, $\mathcal{P}(\mathbf{x})$.
2. **Initialization:** $\hat{x}_i^{(1)} = 0$, $\hat{v}_i^{(1)} = 1$, $Z_a^{(0)} = y_a$.
3. **Output:** $\hat{\mathbf{x}}^{(T)}$.
4. **Iteration:**
for $t = 1, \cdots, T$ **do**

$$V_a^{(t)} = \sum_{i=1}^{N} |h_{ai}|^2 \hat{v}_i^{(t)} \tag{49a}$$

$$Z_a^{(t)} = \sum_{i=1}^{N} h_{ai}\hat{x}_i^{(t)} - \frac{V_a^{(t)}(y_a - Z_a^{(t-1)})}{\sigma_w^2 + V_a^{(t-1)}} \tag{49b}$$

$$\Sigma_i^{(t)} = \left(\sum_{a=1}^{M} \frac{|h_{ai}|^2}{\sigma_w^2 + V_a^{(t)}} \right)^{-1} \tag{49c}$$

$$r_i^{(t)} = \hat{x}_i^{(t)} + \Sigma_i^{(t)} \sum_{a=1}^{M} \frac{h_{ai}^*(y_a - Z_a^{(t)})}{\sigma_w^2 + V_a^{(t)}} \tag{49d}$$

$$\hat{x}_i^{(t+1)} = \mathbb{E}\{x_i | r_i^{(t)}, \Sigma_i^{(t)}\} \tag{49e}$$

$$\hat{v}_i^{(t+1)} = \mathrm{Var}\{x_i | r_i^{(t)}, \Sigma_i^{(t)}\} \tag{49f}$$

end

---

where $\mathsf{F}'_\beta(x_i; r_i^{(t)}, \Sigma^{(t)})$ is the partial derivation of $\mathsf{F}_\beta(x_i; r_i^{(t)}, \Sigma^{(t)})$ w.r.t. $r_i^{(t)}$.

Defining $\mathbf{z}^{(t)} = \mathbf{y} - \mathbf{Z}^{(t)}$ with $\mathbf{Z}^{(t)} \triangleq \{Z_a^{(t)}, \forall a\}$, we have

$$\mathbf{z}^{(t)} = \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}$$
$$+ \frac{1}{\alpha} \mathbf{z}^{(t-1)} \left\langle \mathsf{F}'_\beta(\mathbf{x}; \hat{\mathbf{x}}^{(t-1)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t-1)}, \Sigma^{(t-1)}) \right\rangle, \tag{44a}$$

$$\hat{\mathbf{x}}^{(t+1)} = \mathsf{F}_\beta(\mathbf{x}; \hat{\mathbf{x}}^{(t)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t)}), \Sigma^{(t)}), \tag{44b}$$

$$\Sigma^{(t+1)} = \Sigma^{(t)} \left\langle \mathsf{F}'_\beta(\mathbf{x}; \hat{\mathbf{x}}^{(t)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t)}, \Sigma^{(t)}) \right\rangle. \tag{44c}$$

In large β, by Laplace method of integration we have

$$\lim_{\beta \to \infty} \mathsf{F}_\beta(x_i; r_i^{(t)}, \Sigma^{(t)})$$
$$= \lim_{\beta \to \infty} \int x_i \frac{1}{Z^{\mathrm{pos}}} \exp\left[-\beta\left(\lambda|x_i| + \frac{1}{2\Sigma^{(t)}}(x_i - r_i^{(t)}) \right) \right] \mathrm{d}x_i$$
$$= \underset{x_i}{\arg\min} \frac{1}{2\Sigma^{(t)}}(x_i - r_i^{(t)})^2 + \lambda|x_i|. \tag{45}$$

Similar to (5)-(6), we get

$$\lim_{\beta \to \infty} \mathsf{F}_\beta(x_i; r_i^{(t)}, \Sigma^t) = \mathrm{sign}(r_i^{(t)}) \max(|r_i^{(t)}|, \lambda\Sigma^{(t)}), \tag{46}$$

$$\lim_{\beta \to \infty} \mathsf{F}'_\beta(x_i; r_i^t, \Sigma^{(t)}) = \begin{cases} 1 & |r_i^{(t)}| \geq \Sigma^{(t)} \\ 0 & \text{otherwise} \end{cases}. \tag{47}$$

Defining $\eta(r, \gamma) = \mathrm{sign}(r)\max(|r|, \gamma)$ and $\hat{\tau}^{(t)} = \lambda V^{(t)}$, we have

$$\mathbf{z}^{(t)} = \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}$$
$$+ \frac{1}{\alpha} \mathbf{z}^{(t-1)} \left\langle \eta'(\hat{\mathbf{x}}^{(t-1)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t-1)}, \lambda + \hat{\tau}^{(t-1)}) \right\rangle, \tag{48a}$$

$$\hat{\mathbf{x}}^{(t+1)} = \eta(\hat{\mathbf{x}}^{(t)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t)}, \lambda + \hat{\tau}^{(t)}), \tag{48b}$$

$$\hat{\tau}^{(t+1)} = \frac{\lambda + \hat{\tau}^{(t)}}{\alpha} \left\langle \eta'(\hat{\mathbf{x}}^{(t)} + \mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t)}, \lambda + \hat{\tau}^{(t)}) \right\rangle. \tag{48c}$$

By abusing η, we get the original AMP (10) for LASSO inference problem.

C. Bayes-optimal AMP

In LASSO inference problem, both the prior and likelihood are unknown. However, in the Bayes-optimal setting, where both prior and likelihood function are perfectly given, the MMSE estimator can achieve Bayes-optimal error. Actually, this situation is common in communications. In those cases, it is assumed that each element of $\mathbf{x}$ follows IID distribution $\mathcal{P}_{\mathsf{X}}$. The joint distribution is then represented as

$$\mathcal{P}(\mathbf{x}, \mathbf{y}) = \mathcal{P}(\mathbf{y}|\mathbf{x})\mathcal{P}(\mathbf{x})$$
$$= \prod_{a=1}^{M} \mathcal{P}(y_a|\mathbf{x}) \prod_{i=1}^{N} \mathcal{P}_{\mathsf{X}}(x_i). \tag{50}$$

Similar to the derivation of AMP for LASSO, we get the Bayes-optimal AMP as depicted in Algorithm 2, where the expectation in (49e) and (49f) is taken over

$$\hat{\mathcal{P}}^{(t)}(x_i|\mathbf{y}) = \frac{\mathcal{P}_{\mathsf{X}}(x_i)\mathcal{N}(x_i|r_i^{(t)}, \Sigma_i^{(t)})}{\int \mathcal{P}_{\mathsf{X}}(x)\mathcal{N}(x|r_i^{(t)}, \Sigma_i^{(t)})\mathrm{d}x}. \tag{51}$$

This form of AMP is widely applied to many engineering regions. We call it as Bayes-optimal AMP since (1) this algorithm is based on Bayes-optimal setting; (2) the SE of this algorithm perfectly matches the fixed point of the exact MMSE estimator predicted by replica method. Similar to AMP for LASSO, the form of Bayes-optimal AMP can also be written as (48) with $\eta(\cdot)$ being MMSE denoiser.

D. State Evolution

In this subsection, we only give a sketch of proving AMP's SE in [16]. Let's introduce the following general iterations.

$$\mathbf{h}^{(t+1)} = \mathbf{H}^{\mathbf{T}}\mathbf{m}^{(t)} - \xi_t\mathbf{q}^{(t)}, \tag{52a}$$

$$\mathbf{b}^{(t)} = \mathbf{H}\mathbf{q}^{(t)} - \lambda_t\mathbf{m}^{(t-1)}, \tag{52b}$$

where $\mathbf{m}^{(t)} = g_t(\mathbf{b}^{(t)}, \mathbf{n})$, $\mathbf{q}^{(t)} = f_t(\mathbf{h}^{(t)}, \mathbf{x})$, $\xi_t = \langle g_t'(\mathbf{b}^{(t)}, \mathbf{n})\rangle$, and $\lambda_t = \frac{1}{\alpha}\langle f_t'(\mathbf{h}^{(t)}, \mathbf{x})\rangle$.

Pertaining to this general iterations, the following conclusions can be established. In the large system limit, for any pseudo-Lipschitz function $\varphi : \mathbb{R}^2 \mapsto \mathbb{R}$ of order k and all $t \geq 0$,

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \varphi(h_i^{(t+1)}, x_i) \overset{\text{a.s.}}{=} \mathbb{E}_{\mathsf{Z},\mathsf{X}}\{\varphi(\tau_t\mathsf{Z}, \mathsf{X})\}, \tag{53a}$$

$$\lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \varphi(b_i^{(t)}, n_i) \overset{\text{a.s.}}{=} \mathbb{E}_{\mathsf{Z},\mathsf{N}}\{\varphi(\sigma_t\mathsf{Z}, \mathsf{N})\}, \tag{53b}$$

where

$$\tau_t^2 = \mathbb{E}\{g_t(\sigma_t\mathsf{Z}, \mathsf{N})^2\}, \tag{54}$$

$$\sigma_t^2 = \frac{1}{\alpha}\mathbb{E}\{f_t(\tau_{t-1}\mathsf{Z}, \mathsf{X})^2\}, \tag{55}$$

where $\mathsf{N} \sim \mathcal{P}_\mathsf{N}$ and $\mathsf{X} \sim \mathcal{P}_\mathsf{X}$ are independent of $\mathsf{Z} \sim \mathcal{N}(0,1)$. Specially, $\sigma_0^2 = \lim_{N\to\infty} \frac{1}{N\alpha}\|\mathbf{q}^{(0)}\|^2$.

Define

$$g_t(\mathbf{b}^{(t)}, \mathbf{n}) = \mathbf{b}^{(t)} - \mathbf{n}, \tag{56}$$

$$f_t(\mathbf{h}^{(t)}, \mathbf{x}) = \eta_{t-1}(\mathbf{x} - \mathbf{h}^{(t)}) - \mathbf{x}. \tag{57}$$

Then $\xi_t = 1$ and $\lambda_t = -\frac{1}{\alpha}\langle \eta_{t-1}'(\mathbf{x} - \mathbf{h}^{(t)})\rangle$. To coincide with AMP (Donoho) in (10), it implies that $\mathbf{x} - \mathbf{h}^{(t+1)} = \mathbf{H}^{\mathbf{T}}\mathbf{z}^{(t)} + \mathbf{x}^{(t)}$. We thus have

$$\mathbf{h}^{(t+1)} = \mathbf{x} - (\mathbf{H}^{\mathbf{T}}\mathbf{z}^{(t)} + \mathbf{x}^{(t)}) \tag{58a}$$

$$\mathbf{q}^{(t)} = \hat{\mathbf{x}}^{(t)} - \mathbf{x}, \tag{58b}$$

$$\mathbf{b}^{(t)} = \mathbf{n} - \mathbf{z}^{(t)}, \tag{58c}$$

$$\mathbf{m}^{(t)} = -\mathbf{z}^{(t)}. \tag{58d}$$

Using (56)-(57) and (58), the general iterative equations (52) reduce to the original AMP.

From (54)-(57), we get the SE of AMP

$$\tau_{t+1}^2 = \sigma_w^2 + \sigma_{t+1}^2$$
$$= \sigma_w^2 + \frac{1}{\alpha}\mathbb{E}\{(\eta_t(\mathsf{X} + \tau_t\mathsf{Z}) - \mathsf{X})^2\}. \tag{59}$$

For the proof of AMP's SE, we have the following remarks:

Remark 1. *Conditional distribution.* To prove the equations (53), the so-called condition technique is applied. Let's consider a linear constrain $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{A}$ follows $\mathcal{P}_{\mathbf{A}}(\mathbf{A})$. Let G denote the event that $\mathbf{A}$ satisfies the linear constrain $\mathbf{Y} = \mathbf{AX}$. Then we say that $\mathbf{A}$ under G is distributed as $\mathbf{B}$ following

$$\mathcal{P}_{\mathbf{A}|G}(\mathbf{B}) = \frac{1}{Z}\mathcal{P}_{\mathbf{A}}(\mathbf{B}) \cdot \mathbb{1}_{\mathbf{B}\in\mathcal{L}}, \tag{60}$$

where Z is normalized constant and $\mathcal{L}$ denotes the set of $\mathbf{A}$ that satisfies the linear constrain $\mathbf{Y} = \mathbf{AX}$. We write it as

$\mathbf{A}|_G \overset{\text{d}}{=} \mathbf{B}$.

Gaussianity. The equations (53) shows that in the large system limit, each entry of $\mathbf{h}^{(t+1)}$ and $\mathbf{b}^{(t)}$ tends to Gaussian RV. Regarding $\mathbf{h}^{(t)}$ and $\mathbf{b}^{(t)}$ as column vectors, then for $t \geq 0$, from (52), we have

$$\underbrace{\left[\mathbf{h}^{(1)} + \xi_0\mathbf{q}^{(0)}, \cdots, \mathbf{h}^{(t)} + \xi_{t-1}\mathbf{q}^{(t-1)}\right]}_{\triangleq\mathbf{X}_t}$$
$$= \mathbf{H}^{\mathbf{T}}\underbrace{\left[\mathbf{m}^{(0)}, \cdots, \mathbf{m}^{(t-1)}\right]}_{\triangleq\mathbf{M}_t}, \tag{61}$$

$$\underbrace{\left[\mathbf{b}^{(0)}, \mathbf{b}^{(1)} + \lambda_1\mathbf{m}^{(0)}, \cdots, \mathbf{b}^{(t-1)} + \lambda_{t-1}\mathbf{m}^{(t-2)}\right]}_{\triangleq\mathbf{Y}_t}$$
$$= \mathbf{H}\underbrace{\left[\mathbf{q}^{(0)}, \cdots, \mathbf{q}^{(t-1)}\right]}_{\triangleq\mathbf{Q}_t}. \tag{62}$$

Let G_{t_1,t_2} denote the event that $\mathbf{H}$ satisfies the linear constrains $\mathbf{X}_{t_1} = \mathbf{H}^{\mathbf{T}}\mathbf{M}_{t_1}$ and $\mathbf{Y}_{t_2} = \mathbf{H}\mathbf{Q}_{t_2}$. Then the conditional distribution of $\mathbf{h}^{(t+1)}$ and $\mathbf{b}^{(t)}$ can be expressed as

$$\mathbf{h}^{(t+1)}|_{G_{t+1,t}} \overset{\text{d}}{=} \mathbf{H}|_{G_{t+1,t}}\mathbf{m}^{(t)} - \xi_t\mathbf{q}^{(t)}, \tag{63}$$

$$\mathbf{b}^{(t)}|_{G_{t,t}} \overset{\text{d}}{=} \mathbf{H}|_{G_{t,t}}\mathbf{q}^{(t)} - \lambda_t\mathbf{m}^{(t-1)}. \tag{64}$$

The approximated expressions are shown in [16, Lemma 1], where t-iteration $\mathbf{h}^{(t+1)}$ (or $\mathbf{b}^{(t)}$) on the conditions $G_{t+1,t}$ (or $G_{t,t}$) is expressed as a combination of all preceding $\{\mathbf{h}^{(\tau)}, \forall\tau \leq t\}$ (or $\{\mathbf{b}^{(\tau)}, \tau < t\}$). The proof of Lemma 1 is rigorous since the induction on t is rigorous. Be aware, during the proof of Lemma 1, the fact that $\mathbf{H}$ has IID Gaussian entries is applied to derive the Gaussianity of $\mathbf{h}^{(t+1)}$ and $\mathbf{b}^{(t)}$.

E. Numeric Simulations

In Fig. 6, we show the comparison of Bayes-optimal AMP and its SE in the application of wireless communications. As can be seen from Fig. 6, firstly, AMP matches the SE curve very well; secondly, the performance of AMP becomes better as SNR increases; finally, in small SNR, the measurement ratio α has the effect on convergence speed and fixed point while in large SNR, the effect of α on fixed point can be ignored in the case of QPSK prior. Specially, as SNR=12dB, the curves of $\alpha = 1$ and $\alpha = 4$ converge to the same fixed point almost sure.

In Fig 7, we show the comparison of Bayes-optimal AMP and its SE in compressed sense. As can be observed from Fig. 7, AMP matches the SE curves in all settings. We also see that similar to application in wireless communications, as SNR increases, the gap between the difference measurement ratios will be decreased. Besides, as measurement ratio increases, the convergence speed of AMP will be faster.

III. FROM AMP TO OAMP

Although AMP can achieve the Bayes-optimal MSE performance in IID sub-Gaussian region, the AMP algorithm may fail to converge when $\mathbf{H}$ is ill-conditioned (e.g. large
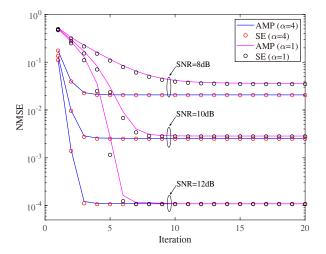
Fig. 6. Iterative behavior of Bayes-optimal AMP and its SE in wireless communications. $\mathbf{H}$ has IID Gaussian entries with zero mean and $1/M$ variance. $M = 1024$, $N = \frac{M}{\alpha}$ and SNR$=1/\sigma_w^2$. The signal of interest $\mathbf{x}$ has IID entries from the set $\{\pm\frac{1}{\sqrt{2}} \pm \mathbb{J}\frac{1}{\sqrt{2}}\}$ with $\mathbb{J}^2 = -1$.
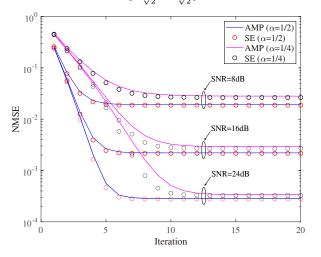

Fig. 7. Iterative behavior of Bayes-optimal AMP and its SE in compressed sensing. $\mathbf{H}$ has IID Gaussian entries with zero mean and $1/M$ variance. $M = \alpha N$, $N = 1024$ and SNR$=1/\sigma_w^2$. The signal of interest $\mathbf{x}$ has IID entries following $\mathcal{BG}(0, 0.05)$.

conditional number, non-zero mean). To extend the scope of AMP to more general random matrices (unitarily-invariant matrix[4]), a modified AMP algorithm termed OAMP [24] was proposed. Different from AMP, the denoiser of OAMP is divergence-free so that the Onsager term vanishes and the LMMSE de-correlated matrix is applied to ensure the orthogonality[5] of input and output errors of denoiser.

A. Orthogonality of input and output errors

Let's consider the following general iterations containing a linear estimation (LE) and a nonlinear estimation (NLE):

$$\text{LE}: \quad \mathbf{r}^{(t)} = \hat{\mathbf{x}}^{(t)} + \mathbf{W}_t(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}) + \mathbf{r}_{\text{Onsager}}^{(t)}, \quad (65a)$$

[4]We say $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\text{T}}$ is unitarily-invariant if $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Sigma}$ are mutually independent, and $\mathbf{U}$, $\mathbf{V}$ are Haar-distributed.

[5]Given two random variables X, Y, we say X is orthogonal to Y if $\mathbb{E}\{\mathsf{XY}\} = 0$. Provided that $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$ are generated by X and Y, respectively, then $\frac{1}{N}\mathbf{x}^{\text{T}}\mathbf{y} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i \overset{\text{a.s.}}{=} \mathbb{E}\{\mathsf{XY}\} = 0$.

$$\text{NLE}: \quad \hat{\mathbf{x}}^{(t+1)} = \tilde{\eta}_t(\mathbf{r}^{(t)}). \quad (65b)$$

where $\mathbf{W}_t$ is a linear transform matrix that maps residual error $\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}$ onto $\mathbb{R}^N$, $\mathbf{r}_{\text{Onsager}}^{(t)}$ is the Onsager term, and $\tilde{\eta}_t$ is the denoiser. Specially, as $\mathbf{r}_{\text{Onsager}}^{(t)} = \frac{\mathbf{r}^{(t-1)} - \hat{\mathbf{x}}^{(t-1)}}{\alpha}\langle \eta'_{t-1}(\mathbf{r}^{(t-1)})\rangle$, $\mathbf{W}_t = \mathbf{H}^{\text{T}}$, and $\tilde{\eta}_t(\mathbf{r}^{(t)}) = \eta_t(\mathbf{r}^{(t)})$, the above general iterations reduce to Donoho's AMP. In AMP algorithm, the Onsager term $\mathbf{r}_{\text{Onsager}}^{(t)}$ ensures the Gaussianity of input signal $\mathbf{r}^{(t)}$ and AMP can achieve Bayes-optimal performance in IID sub-Gaussian random measurement matrix. A significant disadvantage of AMP is that AMP may diverge when the random measurement matrix is beyond IID sub-Gaussian. To extend the scope of AMP to more general case, [24] proposed a modified AMP algorithm called OAMP.

The main ideal of OAMP is to design a linear transform matrix $\mathbf{W}_t$ and denoiser $\tilde{\eta}_t$ so that

- **Divergence-free**[6]. The modified algorithm does not dependent on the Onsager term so that the Onsager term vanishes;
- **Orthogonality**. The modified algorithm maintains the orthogonality of the input and output errors of denoiser $\tilde{\eta}_t(\cdot)$.

For the first issue, a divergence-free denoiser $\tilde{\eta}_t$ can be constructed as

$$\tilde{\eta}_t(\mathbf{r}^{(t)}) = C\left[\eta_t(\mathbf{r}^{(t)}) - \mathbf{r}^{(t)}\left\langle \eta'_t(\mathbf{r}^{(t)})\right\rangle\right], \quad (66)$$

where $\eta_t(\cdot)$ can be an arbitrary pseudo-Lipschitz function and C is a constant. In this case, we have $\langle \tilde{\eta}'_t(\mathbf{r}^{(t)})\rangle = 0$.

For convenience, we define the input and output errors

$$\mathbf{q}^{(t)} = \hat{\mathbf{x}}^{(t)} - \mathbf{x}, \quad (67)$$

$$\mathbf{h}^{(t)} = \mathbf{r}^{(t)} - \mathbf{x}. \quad (68)$$

Substituting the system model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$ and (65) into equations above, we have

$$\text{LE}: \quad \mathbf{h}^{(t)} = (\mathbf{I} - \mathbf{W}_t\mathbf{H})\mathbf{q}^{(t)} + \mathbf{W}_t\mathbf{n}, \quad (69a)$$

$$\text{NLE}: \quad \mathbf{q}^{(t+1)} = \tilde{\eta}_t(\mathbf{x} + \mathbf{h}^{(t)}) - \mathbf{x}. \quad (69b)$$

Also, we define error-related parameters

$$\hat{v}^{(t)} = \lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{q}}^{(t)}\|_2^2, \quad \tau_t^2 = \lim_{N\to\infty}\frac{1}{N}\|\mathbf{h}^{(t)}\|_2^2. \quad (70)$$

Similar to AMP, we assume that the following assumptions hold

- **Assumption 1**: the input error $\mathbf{h}^{(t)}$ consists of IID zero-mean Gaussian entries independent of $\mathbf{x}$, i.e., $\mathsf{R}^{(t)} = \mathsf{X} + \tau_t\mathsf{Z}$ with Z being a standard Gaussian RV.
- **Assumption 2**: the output error $\mathbf{q}^{(t+1)}$ consists of IID entries independent of $\mathbf{H}$ and noise $\mathbf{n}$.

We will show that based on the assumptions above, the de-correlated matrix $\mathbf{W}_t$ and divergence-free imply the orthogonality between input error $\mathbf{h}^{(t)}$ and output error $\mathbf{q}^{(t+1)}$. We say LE is de-correlated one if $\text{Tr}(\mathbf{I} - \mathbf{W}_t\mathbf{H}) = 0$, which implies

$$\mathbf{W}_t = \frac{N}{\text{Tr}(\hat{\mathbf{W}}_t\mathbf{H})}\hat{\mathbf{W}}_t, \quad (71)$$

[6]We say $\eta : \mathbb{R} \mapsto \mathbb{R}$ is divergence-free if $\mathbb{E}\{\eta'(R)\} = 0$.

where $\hat{\mathbf{W}}_t$ can be chosen from:

$$\hat{\mathbf{W}}_t = \begin{cases} \mathbf{H}^{\mathrm{T}} & \text{matched filter (MF)} \\ \hat{\mathbf{W}}_t^{\mathrm{pinv}} & \text{pseudo-inverse} \\ \mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{H}^{\mathrm{T}} + \frac{\sigma^2}{\hat{v}^{(t)}}\mathbf{I}\right)^{-1} & \text{LMMSE} \end{cases} \quad (72)$$

where $\hat{\mathbf{W}}_t^{\mathrm{pinv}} = \mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{H}^{\mathrm{T}})^{-1}$ for $M < N$ and $\hat{\mathbf{W}}_t^{\mathrm{pinv}} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}$ for $M \geq N$. Considering LMMSE de-correlated matrix, one should need to determine $\hat{v}^{(t)}$. We consider η_t as MMSE denoiser and denote it as η_t^{mmse} to distinguish $\tilde{\eta}_t$. Based on Assumption 1, from (66), we have

$$\tilde{\eta}_t(\mathbf{r}^{(t)}) = C\hat{v}_{\mathrm{mmse}}^{(t)}\left(\frac{\eta_t^{\mathrm{mmse}}(\mathbf{r}^{(t)})}{\hat{v}_{\mathrm{mmse}}^{(t)}} - \frac{\mathbf{r}^{(t)}}{\tau_t^2}\right), \quad (73)$$

where the relation $\langle \eta_t^{\mathrm{mmse}\prime}(\mathbf{r}^{(t)})\rangle = \frac{\hat{v}_{\mathrm{mmse}}^{(t)}}{\tau_t^2}$ is applied. The MMSE estimator and its variance are defined as

$$\eta_t^{\mathrm{mmse}}(r_i^{(t)}) = \mathbb{E}\{x_i|r_i^{(t)} = x + \tau_t z\}, \quad (74)$$

$$\hat{v}_{\mathrm{mmse}}^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}\{x_i|r_i^{(t)} = x_i + \tau_t z\}, \quad (75)$$

where the expectation is taken over $\frac{\mathcal{P}_{\mathsf{X}}(x_i)\mathcal{N}(x_i|r_i^{(t)},\tau_t^2)}{\int \mathcal{P}_{\mathsf{X}}(x)\mathcal{N}(x|r_i^{(t)},\tau_t^2)\mathrm{d}x}$.
Then

$$\hat{v}^{(t)} = \lim_{N\to\infty}\frac{1}{N}\|\mathbf{q}^{(t)}\|^2$$

$$\stackrel{\mathrm{a.s.}}{=} \mathbb{E}_{\mathsf{Z},\mathsf{X}}\left\{\left(C\hat{v}_{\mathrm{mmse}}^{(t)}\left(\frac{\eta_t^{\mathrm{mmse}}(\mathsf{X} + \tau_t\mathsf{Z})}{\hat{v}_{\mathrm{mmse}}^{(t)}} - \frac{\mathsf{X} + \tau_t\mathsf{Z}}{\tau_t^2}\right) - \mathsf{X}\right)^2\right\}$$

$$= \mathbb{E}_{\mathsf{Z},\mathsf{X}}\left\{\left(C\eta_t^{\mathrm{mmse}}(\mathsf{X} + \tau_t\mathsf{Z}) - \frac{C\hat{v}_{\mathrm{mmse}}^{(t)} + \tau_t^2}{\tau_t^2}\mathsf{X} - \frac{C\hat{v}_{\mathrm{mmse}}^{(t)}}{\tau_t}\mathsf{Z}\right)^2\right\}$$
$$(76)$$

The coefficients of η_t^{mmse} and X should be equal, i.e., $C = \frac{C\hat{v}_{\mathrm{mmse}}^{(t)} + \tau_t^2}{\tau_t^2}$, and it leads to

$$C = \frac{\tau_t^2}{\tau_t^2 - \hat{v}_{\mathrm{mmse}}^{(t)}}. \quad (77)$$

Substituting this fact into $\hat{v}^{(t)}$ obtains

$$\hat{v}^{(t)} = \mathbb{E}_{\mathsf{Z},\mathsf{X}}\left\{\left[\frac{\tau_t^2}{\tau_t^2 - \hat{v}_{\mathrm{mmse}}^{(t)}}\left(\eta_t^{\mathrm{mmse}}(\mathsf{X} + \tau_t\mathsf{Z}) - \mathsf{X}\right)\right.\right.$$
$$\left.\left. - \frac{\tau_t\hat{v}_{\mathrm{mmse}}^{(t)}}{\tau_t^2 - \hat{v}_{\mathrm{mmse}}^{(t)}}\mathsf{Z}\right]^2\right\}$$

$$= \left(\frac{\tau_t^2}{\tau_t^2 - \hat{v}_{\mathrm{mmse}}^{(t)}}\right)^2\hat{v}_{\mathrm{mmse}}^{(t)} + \left(\frac{\tau_t\hat{v}_{\mathrm{mmse}}^{(t)}}{\tau_t^2 - \hat{v}_{\mathrm{mmse}}^{(t)}}\right)^2$$

$$= \left(\frac{1}{\hat{v}_{\mathrm{mmse}}^{(t)}} - \frac{1}{\tau_t^2}\right)^{-1}, \quad (78)$$

where the facts $\hat{v}_{\mathrm{mmse}}^{(t)} \stackrel{\mathrm{a.s.}}{=} \mathbb{E}_{\mathsf{X},\mathsf{Z}}\{(\eta_t^{\mathrm{mmse}}(\mathsf{X} + \tau_t\mathsf{Z}) - \mathsf{X})^2\}$ and $(\eta_t^{\mathrm{mmse}}(\mathsf{X} + \tau_t\mathsf{Z}) - \mathsf{X})$ independent of Z are applied.

Inserting (77) into (73) obtains

$$\tilde{\eta}_t(\mathbf{r}^{(t)}) = \hat{v}^{(t)}\left(\frac{\eta_t^{\mathrm{mmse}}(\mathbf{r}^{(t)})}{\hat{v}_{\mathrm{mmse}}^{(t)}} - \frac{\mathbf{r}^{(t)}}{\tau_t^2}\right). \quad (79)$$

Algorithm 3: Bayes-Optimal Orthogonal AMP

1. Input: $\mathbf{y}$, $\mathbf{H}$, σ_w^2, and $\mathcal{P}(\mathbf{x})$.
1.Initialization: $\hat{x}_i^{(1)} = 0$, $\hat{v}_i^{(1)} = 1$, $Z_a^{(0)} = y_a$.
2.Output: $\hat{\mathbf{x}}^{(T)}$.
3.Iteration:
for $t = 1, \cdots, T$ **do**

$$\hat{\mathbf{W}}_t = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H} + \frac{\sigma_w^2}{\hat{v}^{(t)}}\mathbf{I}\right)^{-1}\mathbf{H}^{\mathrm{T}} \quad (81\mathrm{a})$$

$$\tau_t^2 = \hat{v}^{(t)}\left(\frac{N}{\mathrm{Tr}(\hat{\mathbf{W}}_t\mathbf{H})} - 1\right) \quad (81\mathrm{b})$$

$$\mathbf{r}^{(t)} = \hat{\mathbf{x}}^{(t)} + \frac{N}{\mathrm{Tr}(\hat{\mathbf{W}}_t\mathbf{H})}\hat{\mathbf{W}}_t(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)})$$
$$(81\mathrm{c})$$

$$\hat{\mathbf{x}}_{\mathrm{mmse}}^{(t+1)} = \eta_t^{\mathrm{mmse}}(\mathbf{r}^{(t)}, \tau_t) \quad (81\mathrm{d})$$

$$\hat{v}_{\mathrm{mmse}}^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}\{x_i|r_i^{(t)}, \tau_t\} \quad (81\mathrm{e})$$

$$\hat{v}^{(t+1)} = \left(\frac{1}{\hat{v}_{\mathrm{mmse}}^{(t+1)}} - \frac{1}{\tau_t^2}\right)^{-1} \quad (81\mathrm{f})$$

$$\hat{\mathbf{x}}^{(t+1)} = \hat{v}^{(t+1)}\left(\frac{\hat{\mathbf{x}}_{\mathrm{mmse}}^{(t+1)}}{\hat{v}_{\mathrm{mmse}}^{(t+1)}} - \frac{\mathbf{r}^{(t)}}{\tau_t^2}\right) \quad (81\mathrm{g})$$

end

Be aware, there is an unknown noise-related parameter τ_t^2, which is expressed as

$$\tau_t^2 = \lim_{N\to\infty}\frac{1}{N}\|\mathbf{h}^{(t)}\|^2$$

$$= \frac{1}{N}\mathrm{Tr}\left((\mathbf{I} - \mathbf{W}_t\mathbf{H})(\mathbf{I} - \mathbf{W}_t\mathbf{H})^{\mathrm{T}}\right)\hat{v}^{(t)}$$
$$+ \frac{1}{N}\mathrm{Tr}(\mathbf{W}_t\mathbf{W}_t^{\mathrm{T}})\sigma_w^2$$

$$\stackrel{(a)}{=} \hat{v}^{(t)}\left[\frac{N\mathrm{Tr}(\hat{\mathbf{W}}_t\mathbf{H}\mathbf{H}^{\mathrm{T}}\hat{\mathbf{W}}_t^{\mathrm{T}})}{\mathrm{Tr}^2\left(\hat{\mathbf{W}}_t\mathbf{H}\right)} - \right] + \frac{N\mathrm{Tr}(\hat{\mathbf{W}}_t\hat{\mathbf{W}}_t^{\mathrm{T}})}{\mathrm{Tr}^2(\hat{\mathbf{W}}_t\mathbf{H})}\sigma_w^2$$

$$= \hat{v}^{(t)}\left[\frac{N\mathrm{Tr}(\hat{\mathbf{W}}_t(\mathbf{H}\mathbf{H}^{\mathrm{T}} + \frac{\sigma_w^2}{\hat{v}^{(t)}}\mathbf{I})\hat{\mathbf{W}}_t^{\mathrm{T}})}{\mathrm{Tr}^2(\hat{\mathbf{W}}_t\mathbf{H})} - 1\right]$$

$$\stackrel{(b)}{=} \hat{v}^{(t)}\left(\frac{N}{\mathrm{Tr}(\hat{\mathbf{W}}_t\mathbf{H})} - 1\right), \quad (80)$$

where the fact $\mathbf{W}_t = \frac{N}{\mathrm{Tr}(\hat{\mathbf{W}}_t\mathbf{A})}\hat{\mathbf{W}}_t$ is used to obtain (a) and the LMMSE de-correlated matrix $\hat{\mathbf{W}}_t = \mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{H}^{\mathrm{T}} + \frac{\sigma^2}{\hat{v}^{(t)}}\mathbf{I}\right)^{-1}$ is applied to obtain (b). This completes the derivation of orthogonal AMP and we post OAMP algorithm in Algorithm 3. Be aware, we here use LMMSE de-correlated $\mathbf{W}_t = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H} + \frac{\sigma^2}{\hat{v}^{(t)}}\mathbf{I}\right)^{-1}\mathbf{H}^{\mathrm{T}}$ and it can be verified that this form is equal to $\mathbf{W}_t = \mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{H}^{\mathrm{T}} + \frac{\sigma^2}{\hat{v}^{(t)}}\mathbf{I}\right)^{-1}$ via SVD.

The below is to prove orthogonality of input and output errors. Define $\mathbf{B}_t = \mathbf{I} - \mathbf{W}_t\mathbf{H}$, we have

$$\mathbb{E}\{\mathbf{h}^{(t)}(\mathbf{q}^{(t)})^{\mathrm{T}}\} = \mathbb{E}\{\mathbf{B}_t\}\mathbb{E}\{\mathbf{q}^{(t)}(\mathbf{q}^{(t)})^{\mathrm{T}}\}$$

$$+ \mathbb{E}\{\mathbf{W}_t\}\mathbb{E}\{\mathbf{n}(\mathbf{q}^{(t)})^{\mathrm{T}}\}$$
$$= \mathbb{E}\{\mathbf{B}_t\}\mathbb{E}\{\mathbf{q}^{(t)}(\mathbf{q}^{(t)})^{\mathrm{T}}\}, \qquad (82)$$

where the last equation holds by Assumption 2. By the SVDs $\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$ and $\mathbf{W}_t = \mathbf{V}\mathbf{G}_t\mathbf{U}^{\mathrm{T}}$, we have

$$\mathbb{E}\{(\mathbf{B}_t)_{ij}\} = \sum_{m=1}^{M} \mathbb{E}\{V_{im}V_{jm}\}(1 - g_m\sigma_m), \qquad (83)$$

where g_m is m-th element of $\mathbf{G}_t$ and σ_m is m-th element of $\boldsymbol{\Sigma}$. Since $\mathbf{V}$ is Haar distribution, we have

$$\mathbb{E}\{(\mathbf{B}_t)_{ij}\} = \begin{cases} 0 & i \neq j \\ \frac{\mathrm{Tr}(\mathbf{B}_t)}{N} & i = j \end{cases}. \qquad (84)$$

Since $\mathrm{Tr}(\mathbf{B}_t) = 0$, we then have $\mathbb{E}\{\mathbf{B}_t\} = \mathbf{0}$ and further

$$\mathbb{E}\{\mathbf{h}^{(t)}(\mathbf{q}^{(t)})^{\mathrm{T}}\} = \mathbf{0}. \qquad (85)$$

This completes the proof of orthogonality of the input and output errors.

B. Relation to Vector AMP

In this subsection, we will show that OAMP shares the same algorithm as Vector AMP (VAMP) [29]. For convenience, we post the VAMP algorithm by omitting iteration as below

$$\hat{\mathbf{x}}_1 = \left(\sigma_w^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{H} + \frac{1}{\gamma_2}\mathbf{I}\right)^{-1}\left(\sigma_w^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{y} + \frac{\mathbf{r}_2}{\gamma_2}\right), \qquad (86a)$$

$$\hat{v}_1 = \frac{1}{N}\mathrm{Tr}\left[\left(\sigma_w^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{H} + \frac{1}{\gamma_2}\mathbf{I}\right)^{-1}\right], \qquad (86b)$$

$$\gamma_1 = \left(\frac{1}{\hat{v}_1} - \frac{1}{\gamma_2}\right)^{-1}, \qquad (86c)$$

$$\mathbf{r}_1 = \gamma_1\left(\frac{\hat{\mathbf{x}}_1}{\hat{v}_1} - \frac{\mathbf{r}_2}{\gamma_2}\right), \qquad (86d)$$

$$\hat{\mathbf{x}}_2 = \mathbb{E}\{\mathbf{x}|\mathbf{r}_1, \gamma_1\}, \qquad (86e)$$

$$\hat{v}_2 = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}\{x_i|r_{1i}, \gamma_1\}, \qquad (86f)$$

$$\gamma_2 = \left(\frac{1}{\hat{v}_2} - \frac{1}{\gamma_1}\right)^{-1}, \qquad (86g)$$

$$\mathbf{r}_2 = \gamma_2\left(\frac{\hat{\mathbf{x}}_2}{\hat{v}_2} - \frac{\mathbf{r}_1}{\gamma_1}\right). \qquad (86h)$$

Comparing VAMP (86) with OAMP in Algorithm 3, it can be found that equations (81d)-(81g) of OAMP are equal to equations (86d)-(86h) of VAMP. To show the equivalence of OAMP and VAMP, one should prove the equivalence of (81b)-(81c) and (86c)-(86d). From (86c), we have

$$\gamma_1 = \frac{\gamma_2\hat{v}_1}{\gamma_2 - \hat{v}_1}$$
$$= \frac{\gamma_2\mathrm{Tr}\left[\left(\mathbf{H}^{\mathrm{T}}\mathbf{H} + \frac{\sigma_w^2}{\gamma_2}\mathbf{I}\right)^{-1}\right]}{N\frac{\gamma_2}{\sigma_w^2} - \mathrm{Tr}\left[\left(\mathbf{H}^{\mathrm{T}}\mathbf{H} + \frac{\sigma_w^2}{\gamma_2}\mathbf{I}\right)^{-1}\right]}, \qquad (87)$$

where by SVD $\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$

$$\mathrm{Tr}\left[\left(\mathbf{H}^{\mathrm{T}}\mathbf{H} + \frac{\sigma_w^2}{\gamma_2}\mathbf{I}\right)^{-1}\right] = \mathrm{Tr}\left[\left(\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma} + \frac{\sigma_w^2}{\gamma_2}\right)^{-1}\right]$$
$$= \sum_{i=1}^{N}\frac{\gamma_2}{\lambda_i\gamma_2 + \sigma_w^2}, \qquad (88)$$

where λ_i is the i-th eigenvalue of $\mathbf{H}^{\mathrm{T}}\mathbf{H}$. Note that if we assume that $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ only has K non-zero eigenvalues then $\lambda_i = 0$ for $i > K$.

From (87), we have

$$\gamma_1 = \frac{\gamma_2\sum_{i=1}^{N}\frac{\gamma_2}{\lambda_i\gamma_2 + \sigma_w^2}}{N\frac{\gamma_2}{\sigma_w^2} - \sum_{i=1}^{N}\frac{\gamma_2}{\lambda_i\gamma_2 + \sigma_w^2}}$$
$$= \frac{\gamma_2\sum_{i=1}^{N}\frac{1}{\lambda_i\gamma_2 + \sigma_w^2}}{\sum_{i=1}^{N}\left(\frac{1}{\sigma_w^2} - \frac{1}{\lambda_i\gamma_2 + \sigma_w^2}\right)}$$
$$= \frac{\gamma_2 N - \gamma_2\sum_{i=1}^{N}\frac{\lambda_i\gamma_2}{\lambda_i\gamma_2 + \sigma_w^2}}{\sum_{i=1}^{N}\frac{\lambda_i\gamma_2}{\lambda_i\gamma_2 + \sigma_w^2}}. \qquad (89)$$

On the other hand, from (81b), we have

$$\tau_t^2 = \hat{v}^{(t)}\left(\frac{N}{\sum_{i=1}^{N}\frac{\lambda_i}{\lambda_i + \frac{\sigma_w^2}{\hat{v}^{(t)}}}} - 1\right)$$
$$= \frac{N\hat{v}^{(t)} - \hat{v}^{(t)}\sum_{i=1}^{N}\frac{\lambda_i\hat{v}^{(t)}}{\hat{v}^{(t)}\lambda_i + \sigma_w^2}}{\sum_{i=1}^{N}\frac{\lambda_i\hat{v}^{(t)}}{\hat{v}^{(t)}\lambda_i + \sigma_w^2}}. \qquad (90)$$

This completes the proof of the equivalence of γ_1 in VAMP and τ_t^2 in OAMP. In addition, (91) and (92) complete the proof of the equivalence of $\mathbf{r}^{(t)}$ of OAMP and $\mathbf{r}_1$ of VAMP. Besides, one could find that VAMP algorithm (86a)-(86h) is same as the diagonal expectation propagation (EP) [27] and expectation consistent (EC) [28, Appendix D] (single-loop). They were proposed independently in different manners but shares the same form. Actually, EP/EC (single-loop) with element-wise variance has a slight difference from OAMP/VAMP, where EP/EC (single-loop) is reduced to OAMP/VAMP by taking the mean operation for element-wise variance. The EP was proposed by modifying the assumed density filter, the EC was proposed by minimizing the Gibbs free energy, OAMP was proposed by extending AMP to more general measurement matrix region, and VAMP was proposed using EP-type message passing. The order of them is EP (2001) by Minka, EC (2005) by Opper, OAMP (2016) by Ma, and VAMP (2016) by Rangan.

C. State Evolution

The asymptotic MSE of OAMP is defined as

$$\mathsf{mse}(\mathbf{x}, t+1) = \lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{x}}_{\mathrm{mmse}}^{(t+1)} - \mathbf{x}\|_2^2$$
$$= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_{\mathrm{mmse},i}^{(t+1)} - x_i)^2$$
$$\stackrel{\mathrm{a.s.}}{=} \mathbb{E}_{\mathsf{X},\mathsf{Z}}\left\{\left(\eta_t^{\mathrm{mmse}}(\mathsf{X} + \tau_t\mathsf{Z}) - \mathsf{X}\right)^2\right\}. \qquad (93)$$

$$\mathbf{r}^{(t)} = \hat{\mathbf{x}}^{(t)} + \frac{N}{\text{Tr}(\hat{\mathbf{W}}_t \mathbf{H})} \left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\hat{v}^{(t)}} \mathbf{I} \right)^{-1} \mathbf{H}^T (\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)})$$

$$= \frac{N}{\text{Tr}(\hat{\mathbf{W}}_t \mathbf{H})} \left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\hat{v}^{(t)}} \mathbf{I} \right)^{-1} \mathbf{H}^T \mathbf{y} + \frac{N}{\text{Tr}(\hat{\mathbf{W}}_t \mathbf{H})} \left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\hat{v}^{(t)}} \mathbf{I} \right)^{-1} \left[\frac{\text{Tr}(\hat{\mathbf{W}}_t \mathbf{H})}{N} \left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\hat{v}^{(t)}} \mathbf{I} \right) \hat{\mathbf{x}}^{(t)} - \mathbf{H}^T \mathbf{H}\hat{\mathbf{x}}^{(t)} \right]$$

$$= \frac{\left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\hat{v}^{(t)}} \mathbf{I} \right)^{-1} \mathbf{H}^T \mathbf{y}}{\frac{1}{N} \sum_{i=1}^N \frac{\lambda_i \hat{v}^{(t)}}{\hat{v}^{(t)} \lambda_i + \sigma_w^2}} + \frac{\left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\hat{v}^{(t)}} \mathbf{I} \right)^{-1}}{\frac{1}{N} \sum_{i=1}^N \frac{\lambda_i \hat{v}^{(t)}}{\hat{v}^{(t)} \lambda_i + \sigma_w^2}} \left[\left(\frac{1}{N} \sum_{i=1}^N \frac{\lambda_i \hat{v}^{(t)}}{\hat{v}^{(t)} \lambda_i + \sigma_w^2} - 1 \right) \mathbf{H}^T \mathbf{H}\hat{\mathbf{x}}^{(t)} + \frac{1}{N} \sum_{i=1}^N \frac{\lambda_i \sigma_w^2}{\hat{v}^{(t)} \lambda_i + \sigma_w^2} \hat{\mathbf{x}}^{(t)} \right]$$

$$= \frac{\left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\hat{v}^{(t)}} \mathbf{I} \right)^{-1} \mathbf{H}^T \mathbf{y}}{\frac{1}{N} \sum_{i=1}^N \frac{\lambda_i \hat{v}^{(t)}}{\hat{v}^{(t)} \lambda_i + \sigma_w^2}} + \left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\hat{v}^{(t)}} \mathbf{I} \right)^{-1} \left(\frac{\sigma_w^2}{\hat{v}^{(t)}} \hat{\mathbf{x}}^{(t)} - \frac{\sigma_w^2 \frac{1}{N} \sum_{i=1}^N \frac{1}{\hat{v}^{(t)} \lambda_i + \sigma_w^2}}{\hat{v}^{(t)} \frac{1}{N} \sum_{i=1}^N \frac{\lambda_i}{\hat{v}^{(t)} \lambda_i + \sigma_w^2}} \mathbf{H}^T \mathbf{H}\hat{\mathbf{x}}^{(t)} \right). \tag{91}$$

$$\mathbf{r}_1 = \frac{\gamma_2 \hat{\mathbf{x}}_1}{\gamma_2 - \hat{v}_1} - \frac{\hat{v}_1 \mathbf{r}_2}{\gamma_2 - \hat{v}_1}$$

$$= \frac{\gamma_2 \left(\sigma_w^{-2} \mathbf{H}^T \mathbf{H} + \frac{1}{\gamma_2} \mathbf{I} \right)^{-1} \left(\sigma_w^{-2} \mathbf{H}^T \mathbf{y} + \frac{\mathbf{r}_2}{\gamma_2} \right)}{\gamma_2 - \frac{1}{N} \text{Tr} \left[\left(\sigma_w^{-2} \mathbf{H}^T \mathbf{H} + \frac{1}{\gamma_2} \mathbf{I} \right)^{-1} \right]} - \frac{\frac{1}{N} \text{Tr} \left[\left(\sigma_w^{-2} \mathbf{H}^T \mathbf{H} + \frac{1}{\gamma_2} \mathbf{I} \right)^{-1} \right] \mathbf{r}_2}{\gamma_2 - \frac{1}{N} \text{Tr} \left[\left(\sigma_w^{-2} \mathbf{H}^T \mathbf{H} + \frac{1}{\gamma_2} \mathbf{I} \right)^{-1} \right]}$$

$$= \frac{\gamma_2 \left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\gamma_2} \mathbf{I} \right)^{-1} \left(\mathbf{H}^T \mathbf{y} + \frac{\sigma_w^2}{\gamma_2} \mathbf{r}_2 \right)}{\gamma_2 - \frac{\sigma_w^2}{N} \sum_{i=1}^N \frac{\gamma_2}{\lambda_i \gamma_2 + \sigma_w^2}} - \frac{\frac{\sigma_w^2}{N} \sum_{i=1}^N \frac{\gamma_2}{\lambda_i \gamma_2 + \sigma_w^2} \mathbf{r}_2}{\gamma_2 - \frac{\sigma_w^2}{N} \sum_{i=1}^N \frac{\gamma_2}{\lambda_i \gamma_2 + \sigma_w^2}}$$

$$= \frac{\left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\gamma_2} \mathbf{I} \right)^{-1} \mathbf{H}^T \mathbf{y}}{\frac{1}{N} \sum_{i=1}^N \frac{\lambda_i \gamma_2}{\lambda_i \gamma_2 + \sigma_w^2}} + \left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\gamma_2} \mathbf{I} \right)^{-1} \left[\frac{\frac{\sigma_w^2}{\gamma_2}}{\frac{1}{N} \sum_{i=1}^N \frac{\lambda_i \gamma_2}{\lambda_i \gamma_2 + \sigma_w^2}} \mathbf{r}_2 - \left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\gamma_2} \mathbf{I} \right) \frac{\frac{1}{N} \sum_{i=1}^N \frac{\sigma_w^2}{\lambda_i \gamma_2 + \sigma_w^2}}{\frac{1}{N} \sum_{i=1}^N \frac{\lambda_i \gamma_2}{\gamma_2 \lambda_i + \sigma_w^2}} \mathbf{r}_2 \right]$$

$$= \frac{\left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\gamma_2} \mathbf{I} \right)^{-1} \mathbf{H}^T \mathbf{y}}{\frac{1}{N} \sum_{i=1}^N \frac{\lambda_i \gamma_2}{\lambda_i \gamma_2 + \sigma_w^2}} + \left(\mathbf{H}^T \mathbf{H} + \frac{\sigma_w^2}{\gamma_2} \mathbf{I} \right)^{-1} \left(\frac{\sigma_w^2}{\gamma_2} \mathbf{r}_2 - \frac{\sigma_w^2 \frac{1}{N} \sum_{i=1}^N \frac{1}{\lambda_i \gamma_2 + \sigma_w^2}}{\gamma_2 \frac{1}{N} \sum_{i=1}^N \frac{\lambda_i}{\gamma_2 \lambda_i + \sigma_w^2}} \mathbf{H}^T \mathbf{H} \mathbf{r}_2 \right). \tag{92}$$

---

where the last equation holds by Assumption 1. As we observed from OAMP in Algorithm 3, in the large system limit, the variance of OAMP estimator can be written as

$$\hat{v}_{\text{mmse}}^{(t+1)} = \frac{1}{N} \sum_{i=1}^N \text{Var}\{x_i | r_i^{(t)}, \tau_t\}$$
$$\overset{\text{a.s.}}{=} \mathbb{E}_{\mathsf{X},\mathsf{Z}} \left\{ (\eta_t^{\text{mmse}}(\mathsf{X} + \tau_t \mathsf{Z}) - \mathsf{X})^2 \right\}. \tag{94}$$

Combining (93) and (94) proves that the variance of OAMP estimator is equal to asymptotic MSE of OAMP almost sure, i.e., $\hat{v}_{\text{mmse}}^{(t+1)} \overset{\text{a.s.}}{=} \mathsf{mse}(\mathbf{x}, t+1)$. Note that $\hat{v}_{\text{mmse}}^{(t+1)}$ in (94) only relies on the parameter τ_t^2 and this parameter can be obtained by

$$\hat{v}^{(t)} = \left(\frac{1}{\hat{v}_{\text{mmse}}^{(t)}} - \frac{1}{\tau_{t-1}^2} \right)^{-1}, \tag{95}$$

$$\tau_t^2 = \hat{v}^{(t)} \left(\frac{N}{\text{Tr}(\hat{\mathbf{W}}_t \mathbf{H})} - 1 \right), \tag{96}$$

where by SVD $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$

$$\frac{1}{N} \text{Tr}(\hat{\mathbf{W}}_t \mathbf{H}) = \frac{1}{N} \text{Tr} \left(\mathbf{H}^T \left(\mathbf{H}\mathbf{H}^T + \frac{\sigma^2}{\hat{v}^{(t)}} \mathbf{I} \right)^{-1} \mathbf{H} \right)$$

$$= \frac{1}{N} \text{Tr} \left(\mathbf{\Sigma}^T \left(\mathbf{\Sigma}\mathbf{\Sigma}^T + \frac{\sigma^2}{\hat{v}^{(t)}} \mathbf{I} \right)^{-1} \mathbf{\Sigma} \right)$$

$$= \frac{1}{N} \sum_{i=1}^N \frac{\sigma_i^2}{\sigma_i^2 + \frac{\sigma_w^2}{\hat{v}^{(t)}}}$$

$$\overset{\text{a.s.}}{=} \mathbb{E} \left\{ \frac{\lambda}{\lambda + \frac{\sigma_w^2}{\hat{v}^{(t)}}} \right\}, \tag{97}$$

where σ_i is the i-th diagonal element of $\mathbf{\Sigma}$, and the expectation in $\mathbb{E}\{\lambda\}$ is taken over the asymptotic eigenvalue distribution of $\mathbf{H}^T \mathbf{H}$.

In the sequel, we obtain the SE of OAMP as below

LE: $$\tau_t^2 = \hat{v}^{(t)} \left(\mathbb{E} \left\{ \frac{\lambda^2}{\lambda^2 + \frac{\sigma_w^2}{\hat{v}^{(t)}}} \right\}^{-1} - 1 \right) \tag{98}$$

NLE: $$\begin{cases} \hat{v}_{\text{mmse}}^{(t+1)} = \mathbb{E}_{\mathsf{X},\mathsf{Z}} \left\{ (\eta_t^{\text{mmse}}(\mathsf{X} + \tau_t \mathsf{Z}) - \mathsf{X})^2 \right\} \\ \hat{v}^{(t+1)} = \left(\frac{1}{\hat{v}_{\text{mmse}}^{(t+1)}} - \frac{1}{\tau_t^2} \right)^{-1} \end{cases} \tag{99}$$

Be aware, in the NLE part, the $\hat{v}_{\text{mmse}}^{(t+1)}$ is output MSE rather than $\hat{v}^{(t+1)}$.
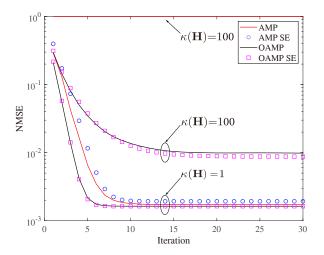
Fig. 8. Iterative behavior of OAMP, AMP and their SEs in compressed sensing. $M = 512$, $N = 1024$ and SNR$=1/\sigma_w^2=$20dB. $\mathbf{x}$ has IID BG entries following $\mathcal{BG}(0, 0.05)$. The measurement matrix is generated by $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ where $\mathbf{U}$ and $\mathbf{V}$ are both Haar distribution and $\mathbf{\Sigma}$ is rectangular matrix whose diagonal is $\sigma_1 \cdots \sigma_M$ with $\sigma_i/\sigma_{i+1} = \kappa(\mathbf{H})^{1/M}$ and $\sum_{i=1}^{M} \sigma_i^2 = N$ such that $|h_{ai}|^2 = O(1/M)$. The condition number is defined as $\kappa(\mathbf{H}) = \frac{\sigma_{\max}(\mathbf{H})}{\sigma_{\min}(\mathbf{H})}$ where $\sigma_{\max}(\mathbf{H})$ and $\sigma_{\min}(\mathbf{H})$ denotes maximum and minimum singular values of $\mathbf{H}$, respectively.

D. Numeric Simulations

In Fig 8, we present the comparison of OAMP, AMP and their SEs in compressed sensing. In $\kappa(\mathbf{H}) = 1$, the SE curves match AMP or OAMP well and OAMP converges faster than AMP. In this case, the gap of the fixed point of AMP and OAMP can be ignored. On the other hand, in $\kappa(\mathbf{H}) = 100$, AMP fails to converge while OAMP and its SE converge to the same fixed point. Besides, the MSE performance of OAMP in $\kappa(\mathbf{H}) = 1$ is better than that in $\kappa(\mathbf{H}) = 100$ (ill-conditioned matrix). Note that since the SE of AMP is obtained under the Gaussian random matrix and thus the condition number has no effect on the performance of SE of AMP.

IV. LONG MEMORY AMP

Although OAMP can be applied to more general random matrices, its complexity with roughly $\mathcal{O}(N^3)$ is larger than AMP with roughly $\mathcal{O}(N^2)$. To balance the computational complexity and region of random measurement matrix, several long memory algorithms have been proposed, such as convolution AMP (CAMP) [25] and memory AMP (MAMP) [26]. CAMP only adjusts the Onsager term where all preceding messages are involved to ensure the Gaussianity of input error. However, CAMP may fail to convergence in ill-conditioned measurement matrix such as large conditional number. Following CAMP and OAMP, MAMP applies a few terms of matrix Taylor series to carry out the matrix inversion in OAMP and modifies the structure of input signal of denoiser to ensure (a) the orthogonality of all preceding input errors and t-th output error, (b) the orthogonality of t-th input error and original signal $\mathbf{x}$, (c) the orthogonality of t-th input error and all preceding output errors.

Recalling the OAMP iterations in Algorithm 3, the complexity of OAMP is dominated by the matrix inversion in (81a). Let's define $\varsigma_t = \frac{\sigma_w^2}{\hat{v}^{(t)}}$ and a relaxation parameter θ_t. Then

$$\left(\theta_t\left(\mathbf{H}\mathbf{H}^{\mathrm{T}} + \varsigma_t\mathbf{I}\right)\right)^{-1} = \left(\mathbf{I} - \left(\mathbf{I} - \theta_t(\mathbf{H}\mathbf{H}^{\mathrm{T}} + \varsigma_t\mathbf{I})\right)\right)^{-1}. \tag{100}$$

Defining $\mathbf{C}_t = \mathbf{I} - \theta_t(\mathbf{H}\mathbf{H}^{\mathrm{T}} + \varsigma_t\mathbf{I})$, we have

$$\left(\theta_t\left(\mathbf{H}\mathbf{H}^{\mathrm{T}} + \varsigma_t\mathbf{I}\right)\right)^{-1} = (\mathbf{I} - \mathbf{C}_t)^{-1}. \tag{101}$$

As spectral radius of $\mathbf{C}_t$ satisfies $\rho(\mathbf{C}_t) < 1$, applying matrix Taylor series gets

$$\left(\theta_t\left(\mathbf{H}\mathbf{H}^{\mathrm{T}} + \varsigma_t\mathbf{I}\right)\right)^{-1} = \sum_{k=0}^{\infty} \mathbf{C}_t^k. \tag{102}$$

It can be verified that $\theta_t = (\lambda^\dagger + \varsigma_t)^{-1}$ with $\lambda^\dagger = \frac{\lambda_{\max}+\lambda_{\min}}{2}$ satisfies $\rho(\mathbf{C}_t) < 1$, where $\lambda_{\max}$ and $\lambda_{\min}$ denote the maximum and minimum eigenvalue of $\mathbf{H}\mathbf{H}^{\mathrm{T}}$, respectively. For convenience, defining $\mathbf{B} = \lambda^\dagger\mathbf{I} - \mathbf{H}\mathbf{H}^{\mathrm{T}}$ yields

$$\hat{\mathbf{W}}_t(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}) = \mathbf{H}^{\mathrm{T}} \sum_{k=0}^{\infty} (\theta_t\mathbf{B})^k(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}). \tag{103}$$

However, the complexity of the exact approximation is still huge. The MAMP applies a few terms of matrix series to represent matrix inversion and use all preceding terms to ensure three orthogonality.

The MAMP considers the following structure:

$$\text{LE}: \begin{cases} \mathbf{z}^{(t)} = \theta_t\mathbf{B}\mathbf{z}^{(t-1)} + \xi_t(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(t)}) \\ \mathbf{r}^{(t)} = \frac{1}{\varepsilon_t}\left(\mathbf{H}^{\mathrm{T}}\mathbf{z}^{(t)} + \sum_{i=1}^{t} p_{t,i}\hat{\mathbf{x}}^{(i)}\right) \end{cases}, \tag{104a}$$

$$\text{NLE}: \begin{cases} \hat{\mathbf{x}}_{\mathrm{mmse}}^{(t+1)} = \eta_t^{\mathrm{mmse}}(\mathbf{r}^{(t)}, \tau_{t,t}) \\ \hat{v}_{\mathrm{mmse}}^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}\{x_i|r_i^{(t)}, \tau_{t,t}\} \\ \hat{v}_{t+1,t+1} = \left(\frac{1}{\hat{v}_{\mathrm{mmse}}^{(t+1)}} - \frac{1}{\tau_{t,t}^2}\right)^{-1} \\ \hat{\mathbf{x}}^{(t+1)} = \hat{v}_{t+1,t+1}\left(\frac{\hat{\mathbf{x}}_{\mathrm{mmse}}^{(t+1)}}{\hat{v}_{\mathrm{mmse}}^{(t+1)}} - \frac{\mathbf{r}^{(t)}}{\tau_{t,t}^2}\right)^{-1} \end{cases}, \tag{104b}$$

where $\mathbf{B} = \lambda^\dagger\mathbf{I} - \mathbf{H}\mathbf{H}^{\mathrm{T}}$ and $\theta_t = (\lambda^\dagger + \varsigma_t)^{-1}$. Note that $\hat{\mathbf{x}}_{\mathrm{mmse}}^{(t+1)}$ is the output estimator rather than $\hat{\mathbf{x}}^{(t+1)}$.

Remark 2. As can be seen from the MAMP algorithm in (104a)-(104b), the parameter $\lambda^\dagger = \frac{\lambda_{\min}+\lambda_{\max}}{2}$ of MAMP algorithm relies on the eigenvalue of $\mathbf{H}\mathbf{H}^{\mathrm{T}}$ which is roughly with cost of $\mathcal{O}(N^3)$. Although some works give the approximations to the maximum or minimum singular value of $\mathbf{H}$, its complexity is still huge. We also note that in the long version [50], a simple bound of maximum eigenvalue and minimum eigenvalue is applied to provide a close performance of perfect eigenvalues, especially in low condition number. In the case of given eigenvalues of $\mathbf{H}\mathbf{H}^{\mathrm{T}}$, the MAMP balances the computational complexity and random measurement region well.

A. Derivation of MAMP

Similar to OAMP, the following assumptions are applied

- **Assumption 3**: the input error $\mathbf{h}^{(t)}$ consists of IID zero-mean Gaussian entries independent of $\mathbf{x}$, i.e., $\mathsf{R}^{(t)} = \mathsf{X} +$

$\tau_{t,t}Z_t$ with Z being a standard Gaussian RV. Let's define $\eta_t = \tau_{t,t}Z_t$. Different from OAMP, MAMP assumes that $[\eta_1, \cdots, \eta_t]^T$ follows joint Gaussian $\mathcal{N}(0, V_t)$ with $V_t = [\tau_{i,j}^2]_{t \times t}$.

- **Assumption 4**: the output error $q^{(t+1)}$ consists of IID entries independent of H and noise n.

Using initial conditions $z^{(0)} = \hat{x}^{(0)} = 0$, from (104a)

$$z^{(t)} = \sum_{i=1}^{t} \xi_i \left(\prod_{j=i+1}^{t} \theta_j \right) B^{t-i}(y - H\hat{x}^{(i)}). \quad (105)$$

Defining $\overline{\theta}_{t,i} = \prod_{j=i+1}^{t} \theta_j$ ($\overline{\theta}_{t,i} = 1$ for $i \geq t$), we have

$$r^{(t)} = \frac{1}{\varepsilon_t} \left(Q_t y + \sum_{i=1}^{t} H_i^t \hat{x}^{(i)} \right), \quad (106)$$

where

$$Q_t = \sum_{i=1}^{t} \xi_i \overline{\theta}_{t,i} H^T B^{t-i}, \quad (107)$$

$$H_i^t = p_{t,i} I - \xi_i \overline{\theta}_{t,i} H^T B^{t-i} H. \quad (108)$$

From the orthogonality of input error and original signal i.e., $\frac{1}{N}(h^{(t)})^T x \overset{\text{a.s.}}{=} 0$, we have

$$\frac{1}{N} \left(h^{(t)} \right)^T x$$
$$= \frac{1}{N} \left(\frac{1}{\varepsilon_t} Q_t (Hx + n) + \frac{1}{\varepsilon_t} \sum_{i=1}^{t} H_i^t (q^{(i)} + x) - x \right)^T x$$
$$= \frac{1}{N\varepsilon_t} x^T ((Q_t H)^T + \sum_{i=1}^{t} (H_i^t)^T) x - \frac{1}{N} x^T x, \quad (109)$$

where $\frac{1}{N}(q^{(i)})^T x \overset{\text{a.s.}}{=} 0$ is applied. Then we get

$$\frac{1}{N\varepsilon_t} \text{Tr} \left\{ Q_t H + \sum_{i=1}^{t} H_i^t \right\} = 1. \quad (110)$$

From the orthogonality of input error and output errors, i.e., $\frac{1}{N}(h^{(t)})^T q^{(i)} \overset{\text{a.s.}}{=} 0$, we have

$$\text{Tr}\{H_i^t\} = 0. \quad (111)$$

Combining (110) and (111), we have

$$p_{t,i} = \frac{1}{N} \xi_i \overline{\theta}_{t,i} \text{Tr} \left\{ H^T B^{t-i} H \right\} \quad (112)$$

$$\varepsilon_t = \sum_{i=1}^{t} p_{t,i} \quad (113)$$

where the parameters $p_{t,i}$ and ε_t can be determined once the parameters $\{\theta_t\}$ and $\{\xi_t\}$ are determined, where ξ_t is obtained by minimizing the averaged input error

$$\tau_{t,t}^2 = \lim_{N \to \infty} \frac{1}{N} \|r^{(t)} - x\|_2^2. \quad (114)$$

Using the facts $\frac{1}{N}(q^{(i)})^T x$ and independence of n and x, we have

$$\tau_{t,t}^2 = \frac{1}{N\varepsilon_t^2} \left\| Q_t n + \sum_{i=1}^{N} H_i^t q^{(i)} \right\|_2^2$$
$$= \frac{1}{N\varepsilon_t^2} \left(\sum_{i=1}^{t} \sum_{j=1}^{t} \xi_i \xi_j \theta_{t,i} \theta_{t,j} \sigma_w^2 \text{Tr} \left\{ H^T B^{2t-i-j} H \right\} \right.$$
$$\left. - \sum_{i=1}^{t} \sum_{j=1}^{t} \hat{v}_{i,j} \text{Tr} \left\{ (H_i^t)^T H_i^t \right\} \right), \quad (115)$$

where $\hat{v}_{i,j} = \frac{1}{N}(q^{(i)})^T q^{(j)}$ and $\hat{v}_{i,j} = \hat{v}_{j,i}$. Defining

$$\vartheta_{t,i} = \xi_i \overline{\theta}_{t,i}, \quad (116)$$

$$W_t = H^T B^t H, \quad w_t = \frac{1}{N} \text{Tr}(W_t), \quad (117)$$

$$N_{i,j} = W_i W_j, \quad \overline{w}_{i,j} = \frac{1}{N} \text{Tr} \left\{ N_{i,j} \right\} - w_i w_j, \quad (118)$$

we get $p_{t,i} = \vartheta_{t,i} w_{t-i}$ and

$$\tau_{t,t}^2 = \frac{1}{\varepsilon_t^2} \sum_{i=1}^{t} \sum_{j=1}^{t} \vartheta_{t,i} \vartheta_{t,j} \left(\sigma_w^2 w_{2t-i-j} + \hat{v}_{i,j} \overline{w}_{t-i,t-j} \right)$$
$$= \frac{c_{t,1} \xi_t^2 - 2c_{t,2} \xi_t + c_{t,3}}{w_0^2 (\xi_t + c_{t,0})^2}, \quad (119)$$

where

$$c_{t,0} = \sum_{i=1}^{t-1} \frac{p_{t,i}}{w_0},$$
$$c_{t,1} = \sigma_w^2 w_0 + \hat{v}_{t,t} \overline{w}_{0,0},$$
$$c_{t,2} = -\sum_{i=1}^{t-1} \vartheta_{t,i} (\sigma_w^2 w_{t-i} + \hat{v}_{t,i} \overline{w}_{0,t-i}),$$
$$c_{t,3} = \sum_{i=1}^{t-1} \sum_{j=1}^{t-1} \vartheta_{t,i} \vartheta_{t,j} (\sigma_w^2 w_{2t-i-j} + \hat{v}_{i,j} \overline{w}_{t-i,t-j}).$$

The parameter ξ_t is obtained by minimizing $\tau_{t,t}^2$. Zeroing $\frac{\partial \tau_{t,t}^2}{\partial \xi_t}$ gets two points $\xi_t = -c_{t,0}$ and $\xi_t = \frac{c_{t,2}c_{t,0}+c_{t,3}}{c_{t,1}c_{t,0}+c_{t,2}}$, where $\xi_t = -c_{t,0}$ is maximum value point while

$$\xi_t^\star = \begin{cases} \frac{c_{t,2}c_{t,0}+c_{t,3}}{c_{t,1}c_{t,0}+c_{t,2}} & c_{t,1}c_{t,0} + c_{t,2} \neq 0 \\ +\infty & \text{otherwise} \end{cases}. \quad (120)$$

Defining the residual error $\tilde{z}^{(t)} = y - H\hat{x}^{(t)}$, the crossed variance $\hat{v}_{i,j}$ can be provided by

$$\frac{1}{N}(\tilde{z}^{(i)})^T \tilde{z}^{(j)} = \frac{1}{N} \left[H(x - \hat{x}^{(i)}) + n \right]^T \left[H(x - \hat{x}^{(j)}) + n \right]$$
$$= \frac{1}{N} \left(-Hq^{(i)} + n \right)^T \left(-Hq^{(j)} + n \right)$$
$$= \frac{1}{N} \text{Tr} \left\{ H^T H \right\} \hat{v}_{i,j} + \alpha \sigma_w^2. \quad (121)$$

It implies $\hat{v}_{i,j} = (\frac{1}{N}(\tilde{z}^{(i)})^T \tilde{z}^{(j)} - \alpha \sigma_w^2)/w_0$.

From (116), we get

$$\vartheta_{t,i} = \begin{cases} \theta_t \vartheta_{t-1,i} & 0 \leq i < t-1 \\ \xi_{t-1} \theta_t & i = t-1 \\ \xi_t & i = t \end{cases}. \quad (122)$$

Totally, the MAMP is run in the following steps: (a) calculating parameters: θ_t, $\vartheta_{t,i}$, $p_{t,i}$ $(i < t)$; (b) calculating parameters: $c_{t,0}$, $c_{t,1}$, $c_{t,2}$, and $c_{t,3}$, and applying them to get $\xi_t = \vartheta_{t,t}$, $p_{t,t}$, and ε_t; (c) calculating $\tau_{t,t}^2$ and carrying out LE; (d) carrying out NLE and calculating $[\hat{v}_{i,j}]_{(t+1)\times(t+1)}$.

In fact, the MAMP is easy to fail to converge without damping, especially in the case of large condition number (e.g., $\kappa(\mathbf{H}) > 10^2$). To ensure the convergence of MAMP, the damping factor is applied to the parameters $\hat{\mathbf{x}}^{(t+1)}$, $\hat{v}_{t+1,t+1}$, and $\tilde{\mathbf{z}}^{(t+1)}$

$$\hat{\mathbf{x}}^{(t+1)} = \beta_1^{(t)}\hat{\mathbf{x}}^{(t+1)} + (1 - \beta_1^{(t)})\hat{\mathbf{x}}^{(t)},$$

$$\tilde{\mathbf{z}}^{(t+1)} = \beta_1^{(t)}\tilde{\mathbf{z}}^{(t+1)} + (1 - \beta_1^{(t)})\tilde{\mathbf{z}}^{(t+1)},$$

$$\hat{v}_{t+1,i} = \beta_2^{(t)}\hat{v}_{t+1,i} + (1 - \beta_2^{(t)})\hat{v}_{t+1,i-1},$$

for $1 < i < t + 1$. Different from the damping presented here, [26] shows another kind damping. But, in fact, the damping factor only has the effect on the convergence speed if algorithm converges.

B. State Evolution

Similar to other AMP-like algorithms, the MSE of MAMP can also be predicted by its SE. The asymptotic MSE of MAMP is defined as

$$\begin{aligned}
\mathsf{mse}(\mathbf{x}, t+1) &= \frac{1}{N}\|\hat{\mathbf{x}}_{\mathrm{mmse}}^{(t+1)} - \mathbf{x}\|_2^2 \\
&\overset{\mathrm{a.s.}}{=} \mathbb{E}\left\{(\eta_t^{\mathrm{mmse}}(\mathsf{X} + \tau_{t,t}\mathsf{Z}) - \mathsf{X})^2\right\} \\
&\overset{\mathrm{a.s.}}{=} \hat{v}_{t+1,t+1}.
\end{aligned} \tag{123}$$

This term only relies on the parameter $\tau_{t,t}^2$, which can be obtained by (119). In $\tau_{t,t}^2$, the parameter $\hat{v}_{i,j} \overset{\mathrm{a.s.}}{=} \frac{1}{N}(\mathbf{q}^{(i)})^\mathsf{T}\mathbf{q}^{(j)}$ is obtained numerically by generating x following $\mathcal{P}_\mathsf{X}(x) = \rho\mathcal{N}(x|0, \rho^{-1}) + (1 - \rho)\delta(x)$ and $r^{(t)} = x + n_t$ with $[n_1, \cdots, n_t] \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Xi}_t)$ where $\boldsymbol{\Xi}_t = [\tau_{i,j}^2]_{t\times t}$ and

$$\begin{aligned}
\tau_{t,\tau}^2 &= \lim_{N\to\infty}\frac{1}{N}(\mathbf{r}^{(t)} - \mathbf{x})^\mathsf{T}(\mathbf{r}^{(\tau)} - \mathbf{x}) \\
&= \frac{1}{\varepsilon_t\varepsilon_\tau}\sum_{i=1}^{t}\sum_{j=1}^{\tau}\vartheta_{t,i}\vartheta_{\tau,j}\left(\sigma_w^2 w_{t+\tau-i-j} + \hat{v}_{ij}\overline{w}_{t-i,\tau-j}\right),
\end{aligned}$$

with $\tau_{t,\tau} = \tau_{\tau,t}$. Then,

$$\forall \tau < t: \quad \hat{v}_{t,\tau} = \mathbb{E}\left\{(\hat{x}^{(t)} - x)(\hat{x}^{(\tau)} - x)\right\}.$$

C. Numeric Simulation

In Fig. 9, we show the pre-iteration behavior of MAMP and OAMP by varying the condition number in application of compressed sensing. As can be observed from this figure, MAMP and OAMP converge to the same fixed point. In $\kappa(\mathbf{H}) = 1$, MAMP has the comparable convergence speed as OAMP. However, as the $\kappa(\mathbf{H})$ increases, MAMP need to pay more iteration times to converge the same fixed point as OAMP. Also, we note that the convergence speed and NMSE performance of MAMP and OAMP tend to worse in large condition number.
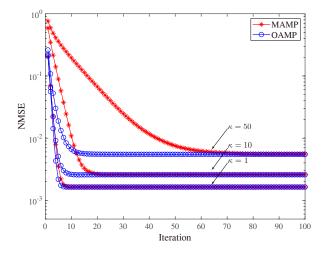

Fig. 9. Comparison of MAMP and OAMP in different condition numbers. Each entry of $\mathbf{x}$ is generated from BG distribution $\mathcal{BG}(0, 0.1)$. $(M, N) = (1024, 512)$ and SNR $= 1/\sigma_w^2 = 20$dB. The measurement matrix is generated by $\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}$ where both $\mathbf{U}$ and $\mathbf{V}$ are Haar-distributed and $\boldsymbol{\Sigma}$ is rectangular matrix whose diagonal elements are $\sigma_1, \cdots, \sigma_M$ with $\frac{\sigma_i}{\sigma_{i+1}} = \kappa(\mathbf{H})^{1/M}$ and $\sum_{i=1}^{M}\sigma_i^2 = N$, where $\kappa(\mathbf{H}) = \frac{\sigma_{\max}(\mathbf{H})}{\sigma_{\min}(\mathbf{H})}$ with $\sigma_{\max}(\mathbf{H})$ and $\sigma_{\min}(\mathbf{H})$ being maximum and minimum singular values of $\mathbf{H}$, respectively. The damping factors $\beta_1^{(t)} = 0.7$ and $\beta_2^{(t)} = 0.8$ are applied to the cases of $\kappa(\mathbf{H}) = 1$ and $\kappa(\mathbf{H}) = 10$, while $\beta_1^{(t)} = \beta_2^{(t)} = 0.4$ are applied to the case of $\kappa(\mathbf{H}) = 50$.

V. Conclusions

In this paper, we reviewed several AMP-like algorithms: AMP, OAMP, VAMP, and MAMP. We began at introducing AMP algorithm, which is originally proposed for providing a sparse solution to LASSO inference problem but is widely applied to a lot of engineering fields under Bayes-optimal setting. In IID sub-Gaussian random measurement matrices region, the AMP algorithm can achieve Bayes-optimal MSE performance, but it may fail to converge if random measurement is beyond IID sub-Gaussian. Following AMP, we introduced a modified AMP algorithm termed OAMP, which modified AMP in two aspects: LMMSE de-correlated matrix and divergence-free denoiser. The OAMP algorithm can be applied to more general region: unitarily-invariant matrix, but it should be payed more computational complexity due to matrix inversion. To balance the computational complexity and random measurement region, the MAMP algorithm applies several terms of matrix Taylor series to approximate matrix inversion and applies all preceding messages to ensure three orthogonality. The MAMP algorithm relies on the given spectral of sample of random measurement matrix. Although, several works gave some approximations to it, the complexity is still huge. In addition, the convergence speed of MAMP is slower than OAMP especially in the case of large condition number. On the other hand, a significant feature of AMP-like algorithms is that their asymptotic MSE performance can be fully predicted by their SEs. We also gave a brief derivation of their SEs.

VI. Acknowledgements

REFERENCES

[1] D. L. Donoho, "For most large underdetermined systems of equations, the minimal L1-norm near-solution approximates the sparsest near-solution," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 7, pp. 907–934, 2006.

[2] ——, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] Y. Kabashima, T. Wadayama, and T. Tanaka, "A typical reconstruction limit for compressed sensing based on lp-norm minimization," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 09, p. L09003, 2009.

[4] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *2010 IEEE information theory workshop on information theory (ITW 2010, Cairo)*. IEEE, 2010, pp. 1–5.

[5] ——, "Message passing algorithms for compressed sensing: II. analysis and validation," in *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*. IEEE, 2010, pp. 1–5.

[6] ——, "How to design message passing algorithms for compressed sensing," *preprint*, 2011.

[7] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[8] C. A. Metzler, A. Maleki, and R. G. Baraniuk, "From denoising to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 62, no. 9, pp. 5117–5144, 2016.

[9] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, 2009.

[10] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[11] T. Blumensath, M. E. Davies, G. Rilling, Y. Eldar, and G. Kutyniok, "Greedy algorithms for compressed sensing." 2012.

[12] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[13] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[14] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

[15] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[16] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, 2011.

[17] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and message passing algorithms," *The Annals of Applied Probability*, vol. 25, no. 2, pp. 753–822, 2015.

[18] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[19] J. Kim and J. Pearl, "A computational model for causal and diagnostic reasoning in inference systems," in *International Joint Conference on Artificial Intelligence*, 1983, pp. 0–0.

[20] M. Mezard and A. Montanari, *Information, physics, and computation*. Oxford University Press, 2009.

[21] D. J. Thouless, P. W. Anderson, and R. G. Palmer, "Solution of 'solvable model of a spin glass'," *Philosophical Magazine*, vol. 35, no. 3, pp. 593–601, 1977.

[22] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," *Journal of Physics A: Mathematical and General*, vol. 36, no. 43, pp. 11 111–11 121, 2003.

[23] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983–2010, 2005.

[24] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[25] K. Takeuchi, "Bayes-optimal convolutional AMP," *IEEE Trans. Inf. Theory*, 2021.

[26] L. Liu, S. Huang, and B. M. Kurkoski, "Memory approximate message passing," in *2021 IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, 2021, pp. 1379–1384.

[27] T. P. Minka, "A family of algorithms for approximate bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[28] M. Opper, O. Winther, and M. J. Jordan, "Expectation consistent approximate inference." *Journal of Machine Learning Research*, vol. 6, no. 12, 2005.

[29] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

[30] H. Zhang, "Identical fixed points in state evolutions of AMP and VAMP," *Signal Processing*, vol. 173, p. 107601, 2020.

[31] T. Takahashi and Y. Kabashima, "Macroscopic analysis of vector approximate message passing in a model mismatch setting," in *2020 IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, 2020, pp. 1403–1408.

[32] C. Gerbelot, A. Abbara, and F. Krzakala, "Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula)," *arXiv preprint arXiv:2006.06581*, 2020.

[33] T. Obuchi and A. Sakata, "Cross validation in sparse linear regression with piecewise continuous nonconvex penalties and its acceleration," *Journal of Physics A: Mathematical and Theoretical*, vol. 52, no. 41, p. 414003, 2019.

[34] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *2011 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2011, pp. 2168–2172.

[35] X. Meng, S. Wu, L. Kuang, and J. Lu, "An expectation propagation perspective on approximate message passing," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1194–1197, 2015.

[36] Q. Zou, H. Zhang, C.-K. Wen, S. Jin, and R. Yu, "Concise derivation for generalized approximate message passing using expectation propagation," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1835–1839, 2018.

[37] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing¡ªpart i: Derivation," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5839–5853, 2014.

[38] A. Maillard, L. Foini, A. L. Castellanos, F. Krzakala, M. Mézard, and L. Zdeborová, "High-temperature expansions and message passing algorithms," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 11, p. 113301, 2019.

[39] A. Maillard, F. Krzakala, M. Mézard, and L. Zdeborová, "Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising," *arXiv preprint arXiv:2110.08775*, 2021.

[40] P. Schniter, S. Rangan, and A. K. Fletcher, "Vector approximate message passing for the generalized linear model," in *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 1525–1529.

[41] H. He, C.-K. Wen, and S. Jin, "Generalized expectation consistent signal recovery for nonlinear measurements," in *2017 IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, 2017, pp. 2333–2337.

[42] F. Tian, L. Liu, and X. Chen, "Generalized memory approximate message passing," *arXiv preprint arXiv:2110.06069*, 2021.

[43] A. Manoel, F. Krzakala, M. Mézard, and L. Zdeborová, "Multi-layer generalized linear estimation," in *2017 IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, 2017, pp. 2098–2102.

[44] A. K. Fletcher, S. Rangan, and P. Schniter, "Inference in deep networks in high dimensions," in *2018 IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, 2018, pp. 1884–1888.

[45] P. Pandit, M. Sahraee, S. Rangan, and A. K. Fletcher, "Asymptotics of MAP inference in deep networks," in *2019 IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, 2019, pp. 842–846.

[46] Q. Zou, H. Zhang, and H. Yang, "Multi-layer bilinear generalized approximate message passing," *IEEE Trans. Signal Process.*, vol. 69, pp. 4529–4543, 2021.

[47] S. Rangan, A. K. Fletcher, and V. K. Goyal, "Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1902–1923, 2012.

[48] N. Merhav, *Statistical physics and information theory*. Now Publishers Inc, 2010.

[49] T. Richardson and R. Urbanke, *Modern coding theory*. Cambridge university press, 2008.

[50] L. Liu, S. Huang, and B. M. Kurkoski, "Memory approximate message passing," *arXiv preprint arXiv:2012.10861*, 2020.