A Simple Derivation of AMP and its State Evolution via First-Order Cancellation

Philip Schniter, *Fellow, IEEE*

***Abstract*—We consider the linear regression problem, where the goal is to recover the vector $x \in \mathbb{R}^n$ from measurements $y = Ax + w \in \mathbb{R}^m$ under known matrix A and unknown noise w. For large i.i.d. sub-Gaussian A, the approximate message passing (AMP) algorithm is precisely analyzable through a state-evolution (SE) formalism, which furthermore shows that AMP is Bayes optimal in certain regimes. The rigorous SE proof, however, is long and complicated. And, although the AMP algorithm can be derived as an approximation of loop belief propagation (LBP), this viewpoint provides little insight into why large i.i.d. A matrices are important for AMP, and why AMP has a state evolution. In this work, we provide a heuristic derivation of AMP and its state evolution, based on the idea of "first-order cancellation," that provides insights missing from the LBP derivation while being much shorter than the rigorous SE proof.**

***Index Terms*—Belief propagation, compressive sensing, inference algorithms, message passing, random matrices.**

I. Introduction

We consider the standard linear regression problem, where the goal is to recover the vector $\boldsymbol{x} \in \mathbb{R}^n$ from measurements

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} \in \mathbb{R}^m, \tag{1}$$

where $\boldsymbol{A}$ is a known matrix and $\boldsymbol{w}$ is an unknown disturbance. With high-dimensional random $\boldsymbol{A}$, the approximate message passing (AMP) algorithm [1] remains one of the most celebrated and best understood iterative algorithms. In particular, when the entries of $\boldsymbol{A}$ are drawn i.i.d. from a sub-Gaussian distribution and $m, n \to \infty$ with $m/n \to \delta \in (0, \infty)$, ensemble behaviors of AMP, such as the per-iteration mean-squared error (MSE), can be perfectly predicted using a state evolution (SE) formalism [2].[1] Furthermore, the SE formalism shows that, in certain regimes, AMP's MSE converges to the minimum MSE as predicted by the replica method [2], [3], which has been shown to coincide with the minimum MSE for linear regression under i.i.d. Gaussian $\boldsymbol{A}$ [4], [5] as $m, n \to \infty$ with $m/n \to \delta \in (0, \infty)$. More recently, it has been proven that the state-evolution accurately characterizes AMP's behavior for large but finite m, n [6].

The rigorous SE proofs in [2], [3], [6], however, are long and complicated, and thus remain out of reach for many readers. And, although the AMP algorithm can be heuristically derived from an approximation of loop belief propagation (LBP) [7], [8], as outlined in [9], [3, App.A], and [10], and expectation propagation (EP) [11], [12], as outlined in [13], the LBP/EP perspective is lacking in several respects. First, LBP and EP are heuristic, making it surprising that further approximations of these approaches can be optimal. Second, the LBP and EP derivations provide little insight into why large i.i.d. $\boldsymbol{A}$ matrices are important for AMP. For the LBP/EP derivations, it suffices that the entries of $\boldsymbol{A}$ have "roughly the same magnitude $O(1/\sqrt{m})$" [10], [13], suggesting that structured matrices (e.g., DCT, Hadamard, Fourier) should work as well as i.i.d. random ones. But, in practice, AMP often diverges with such structured matrices, and the reasons why are not evident from the LBP/EP viewpoint. Third, the LBP and EP derivations do not explain why AMP obeys a scalar state evolution when $\boldsymbol{A}$ is large and i.i.d. sub-Gaussian, nor do they describe how the estimation error variance can be predicted at each iteration.

In this work, we propose a heuristic derivation of AMP and its MSE state evolution that uses the simple idea of "first-order cancellation." This derivation provides insights missing from the LBP and EP derivations, while being much more accessible than the rigorous SE proofs.

II. Problem Setup

In our treatment of the linear regression problem (1), $\boldsymbol{y} = [y_1, \dots, y_m]^\top$, $\boldsymbol{x} = [x_1, \dots, x_n]^\top$, and $\boldsymbol{w} = [w_1, \dots, w_m]^\top$ are deterministic vectors and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a deterministic matrix. Importantly, however, we assume that the components $\{a_{ij}\}$ of $\boldsymbol{A}$ are realizations of i.i.d. Bernoulli[2] random variables $A_{ij} \in \pm\frac{1}{\sqrt{m}}$ that are drawn independently of $\boldsymbol{x}$ and $\boldsymbol{w}$. Our model for $\boldsymbol{A}$ is a special case of that considered in [2].

Throughout, we will focus on the following large-system limit.

Definition 1: The "large system limit" is defined as $m, n \to \infty$ with $m/n \to \delta$ for some fixed sampling ratio $\delta \in (0, \infty)$.

We will assume that the components of $\boldsymbol{x}$, $\boldsymbol{w}$, and $\boldsymbol{y}$ scale as $O(1)$ in the large-system limit.

Manuscript received July 18, 2019; revised February 10, 2020; accepted June 25, 2020. Date of publication July 1, 2020; date of current version August 18, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Remy Boyer. This work was supported by the National Science Foundation under Grant CCF-1716388.

The author is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: schniter.1@osu.edu).

Digital Object Identifier 10.1109/TSP.2020.3005545

[1] See also [3] for an earlier proof of AMP's state evolution under i.i.d. Gaussian entries.

[2] With additional work, our derivation can be extended to i.i.d. Gaussian A_{ij}, but doing so lengthens the derivation and provides little additional insight.

We consider a family of algorithms that, starting with $\boldsymbol{x}^{(0)} = \mathbf{0}$, iterates the following over iteration index $t = 0, 1, 2, \ldots$:

$$\boldsymbol{v}^{(t)} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{(t)} + \boldsymbol{\mu}^{(t)} \tag{2a}$$

$$\boldsymbol{x}^{(t+1)} = \eta^{(t)}(\underbrace{\boldsymbol{x}^{(t)} + \boldsymbol{A}^\top \boldsymbol{v}^{(t)}}_{\triangleq \boldsymbol{r}^{(t)}}), \tag{2b}$$

where $\boldsymbol{\mu}^{(t)}$ is a correction term and $\eta^{(t)}(\cdot)$ is a component-wise separable function. That is, $[\eta^{(t)}(\boldsymbol{r})]_j = \eta^{(t)}(r_j)\,\forall j$, where, with some abuse of notation, we use the same notation for the function $\eta^{(t)} : \mathbb{R}^n \to \mathbb{R}^n$ and its component maps $\eta^{(t)} : \mathbb{R} \to \mathbb{R}$. The quantity $\boldsymbol{x}^{(t)}$ is iteration-t estimate of the unknown vector $\boldsymbol{x}$. We refer to $\eta^{(t)}(\cdot)$ as a "denoiser" for reasons that will become clear in the sequel. For technical reasons, we will assume that $\eta^{(t)}(\cdot)$ is a polynomial function of bounded degree, similar to the assumption in [2].

The classical iterative shrinkage/thresholding (IST) algorithm [14] uses no correction, i.e.,

$$\boldsymbol{\mu}^{(t)} = \mathbf{0}, \tag{3}$$

for all iterations t, whereas the AMP algorithm [1] uses the "Onsager" correction

$$\boldsymbol{\mu}^{(t)} = \frac{1}{m}\boldsymbol{v}^{(t-1)} \sum_{j=1}^{n} \eta^{(t-1)\prime}(r_j^{(t-1)}), \tag{4}$$

initialized with $\boldsymbol{\mu}^{(0)} = \mathbf{0}$. In (4), $\eta^{(t)\prime}$ refers to the derivative of $\eta^{(t)}$. Our goal is to analyze the effect of $\boldsymbol{\mu}^{(t)}$ on the behavior of algorithm (2) in the large-system limit, and in particular to understand how and why the Onsager correction (4) is a good choice. To do this, we will analyze the errors on $\boldsymbol{r}^{(t)}$ and $\boldsymbol{x}^{(t)}$ in (2) and drop terms that vanish in the large-system limit.

It has been shown [3], [10] that IST has an predictable and desirable behavior in the case that $\boldsymbol{A}$ is a large i.i.d. Gaussian matrix that is *re-drawn* at each iteration t (with a corresponding update of $\boldsymbol{y}$). But this desirable behavior vanishes in the practical case that $\boldsymbol{A}$ is fixed over the iterations. In some sense, the goal of the Onsager correction (4) is to restore this desirable behavior when $\boldsymbol{A}$ is *fixed* over the iterations.

III. AMP Derivation

We will now analyze the error $\boldsymbol{e}^{(t)}$ on the input to the denoiser $\boldsymbol{r}^{(t)}$, i.e.,

$$\boldsymbol{e}^{(t)} \triangleq \boldsymbol{r}^{(t)} - \boldsymbol{x}. \tag{5}$$

From (2) and (5) we have that

$$\boldsymbol{e}^{(t)} = \boldsymbol{x}^{(t)} + \boldsymbol{A}^\top(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{(t)} + \boldsymbol{\mu}^{(t)}) - \boldsymbol{x} \tag{6}$$

$$= (\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}^{(t)} + \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} + \boldsymbol{\mu}^{(t)}) - \boldsymbol{x} \tag{7}$$

$$= (\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}^{(t)} - (\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x} + \boldsymbol{A}^\top(\boldsymbol{w} + \boldsymbol{\mu}^{(t)}). \tag{8}$$

Let us examine the jth component of $\boldsymbol{e}^{(t)}$ when $t \geq 1$. We have that

$$[(\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}^{(t)}]_j$$

$$= x_j^{(t)} - \sum_i a_{ij} \sum_l a_{il} x_l^{(t)} \tag{9}$$

$$= \left(1 - \sum_{i=1}^{m} a_{ij}^2\right) x_j^{(t)} - \sum_i a_{ij} \sum_{l \neq j} a_{il} x_l^{(t)} \tag{10}$$

$$= -\sum_i a_{ij} \sum_{l \neq j} a_{il} x_l^{(t)} \tag{11}$$

since $a_{ij}^2 = 1/m\,\forall ij$. Continuing,

$$[(\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}^{(t)}]_j$$

$$= -\sum_i a_{ij} \sum_{l \neq j} a_{il} \eta^{(t-1)}(r_l^{(t-1)}) \tag{12}$$

$$= -\sum_i a_{ij} \sum_{l \neq j} a_{il} \eta^{(t-1)}\left(\underbrace{x_l^{(t-1)} + \sum_{k \neq i} a_{kl} v_k^{(t-1)}}_{\triangleq r_{il}^{(t-1)}} + a_{il} v_i^{(t-1)}\right), \tag{13}$$

where $r_{il}^{(t-1)}$ omits the direct contribution of a_{il} from $r_l^{(t-1)}$ and thus is only weakly dependent on $\{a_{ij}\}_{j=1}^n$. We formalize this weak dependence through Assumption 1, which is admittedly an approximation. In fact, the approximate nature of Assumption 1 is the main reason that our derivation is heuristic.

Assumption 1: The matrix entry a_{ij} is a realization of an equiprobable Bernoulli random variable $A_{ij} \in \pm\frac{1}{\sqrt{m}}$, where $\{A_{ij}\}$ are mutually independent and, A_{ij} is independent of $\{r_{il}^{(t-1)}\}_{l=1}^n$, $\{x_l\}_{l=1}^n$, and $\{w_k\}_{k=1}^m$.

We now say a few words about Assumption 1. The assumption that $\{A_{ij}\}$ are i.i.d. and independent of $\boldsymbol{x}$ and $\boldsymbol{w}$ is rather common in the compressive sensing literature. For example, these assumptions are used in the rigorous AMP analyses [2], [3], [6]. The assumption of Bernoulli A_{ij} is a bit stronger, but it is not critical, in that our analysis could be extended to handle other sub-Gaussian distributions on A_{ij} with additional steps. Doing so, however, would complicate the derivation without providing much additional insight, and so we have elected not to take this path. The assumption that A_{ij} is independent of $\{r_{il}^{(t-1)}\}_{l=1}^n$ is far stronger. In reality, there is a weak dependence between these quantities, but properly accounting for it seems to require completely different analysis methods, such as those in [2], [3], [6].

Assumption 1 will be used often when analyzing summations, as in the following lemma.

Lemma 1: Consider the quantity $z_i = \sum_{j=1}^n a_{ij}u_j$, where a_{ij} are realizations of i.i.d. random variables A_{ij} with zero mean and $\mathbb{E}[A_{ij}^2] = 1/m$. If $\{A_{ij}\}$ are drawn independently of $\{u_j\}$, and $\{u_j\}$ scale as $O(1)$ in the large-system limit, then z_i also scales as $O(1)$.

Proof: First, note that z_i is a realization of the random variable $Z_i \triangleq \sum_{j=1}^n A_{ij}u_j$. Furthermore, $\mathbb{E}[Z_i^2] = \mathbb{E}[(\sum_{j=1}^n A_{ij}u_j)^2] = \sum_{j=1}^n \sum_{l=1}^n \mathbb{E}[A_{ij}A_{il}]u_j u_l = \frac{1}{m}\sum_{j=1}^n u_j^2 = \frac{n}{m}\frac{1}{n}\sum_{j=1}^n u_j^2$, since $\mathbb{E}[A_{ij}A_{il}] = 1/m$ if $j = l$ and $\mathbb{E}[A_{ij}A_{il}] = \mathbb{E}[A_{ij}]\,\mathbb{E}[A_{il}] = 0$ if $j \neq l$. Clearly m/n and $\frac{1}{n}\sum_{j=1}^n u_j^2$ are both $O(1)$ in the large-system limit. Thus we conclude that $\mathbb{E}[Z_i^2]$ is $O(1)$. Finally, since z_i is a realization

of a random variable Z_i whose second moment is $O(1)$, we conclude that z_i scales as $O(1)$ in the large-system limit. ■

We now emphasize the importance of the independence properties on $\{A_{ij}\}$ used in Assumption 1 and Lemma 1. For example, consider the case where $A_{ij} = \mathrm{sgn}(u_j)/\sqrt{m}$, so that $\mathbb{E}[A_{ij}^2] = 1/m$ (as in Lemma 1) but where A_{ij} depends on $\{u_j\}$. In this case, $z_i = \sum_{j=1}^n a_{ij}u_j = \frac{1}{\sqrt{m}}\sum_{j=1}^n |u_j|$, which scales as $O(\sqrt{m})$ in the large-system limit, and thus is fundamentally different from the $O(1)$ scaling observed in Lemma 1.

The distinction between $\{a_{ij}\}$ being independent of other quantities, versus $\{a_{ij}\}$ simply being the right size, is one of the major differences between our derivation and other simple derivations based on LBP and EP. For example, both [10] and [13] require only that "$\{a_{ij}\}$ are roughly the same magnitude $O(1/\sqrt{m})$," which suggests that properly normalized structured matrices (e.g, Hadamard, Fourier, Discrete Cosine Transform, Radon Transform) should work well with AMP. But when used to recover natural images/signals with these matrices, AMP often diverges. And the LBP/EP viewpoint does not explain why.

In the sequel, we will make use of the following lemma, whose proof is postponed because it is a bit long and does not provide much insight.

Lemma 2: Under Assumption 1 and the Onsager choice of $\boldsymbol{\mu}^{(t)}$ from (4), the elements of $\boldsymbol{v}^{(t)}$, $\boldsymbol{r}^{(t)}$, $\boldsymbol{x}^{(t)}$, and $\boldsymbol{\mu}^{(t)}$ scale as $O(1)$ in the large-system limit for all iterations t.

Proof: See the appendix. ■

We now perform a Taylor series expansion of the $\eta^{(t-1)}$ term in (13) about $r_{il}^{(t-1)}$:

$$\begin{aligned}&\eta^{(t-1)}(r_{il}^{(t-1)} + a_{il}v_i^{(t-1)})\\ &\quad= \eta^{(t-1)}(r_{il}^{(t-1)}) + a_{il}v_i^{(t-1)}\eta^{(t-1)\prime}(r_{il}^{(t-1)})\\ &\quad\quad+ \underbrace{\frac{1}{2}a_{il}^2(v_i^{(t-1)})^2\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)}) + \text{H.O.T.}}_{O(1/m)},\end{aligned} \tag{14}$$

where the $O(1/m)$ scaling follows from the fact that $a_{il}^2 = 1/m\,\forall il$, that both $v_i^{(t-1)}$ and $r_{il}^{(t-1)}$ scale as $O(1)$ via Lemma 1, and $\eta^{(t-1)}(\cdot)$ is polynomial of bounded degree, which implies that $\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)})$ also scales as $O(1)$. For similar reasons, the second term in (14) scales as $O(1/\sqrt{m})$. We will ignore the $O(1/m)$ term in (14) since it vanishes relative to the $O(1/\sqrt{m})$ component in the large-system limit. Thus we have

$$\begin{aligned}&[(\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}^{(t)}]_j\\ &\approx -\sum_i a_{ij}\sum_{l\neq j} a_{il}\left[\eta^{(t-1)}(r_{il}^{(t-1)}) + a_{il}v_i^{(t-1)}\eta^{(t-1)\prime}(r_{il}^{(t-1)})\right]\end{aligned} \tag{15}$$

$$\begin{aligned}&= -\sum_i a_{ij}\sum_{l\neq j} a_{il}\eta^{(t-1)}(r_{il}^{(t-1)})\\ &\quad- \frac{1}{m}\sum_i a_{ij}v_i^{(t-1)}\sum_{l\neq j}\eta^{(t-1)\prime}(r_{il}^{(t-1)})\end{aligned} \tag{16}$$

using $a_{il}^2 = 1/m\,\forall il$.

Similar to (11), we have

$$[(\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}]_j = -\sum_i a_{ij}\sum_{l\neq j} a_{il}x_l, \tag{17}$$

which, combined with (8) and (16), yields

$$\begin{aligned}e_j^{(t)} &\approx \sum_i a_{ij}\sum_{l\neq j} a_{il}\left[x_l - \eta^{(t-1)}(r_{il}^{(t-1)})\right]\\ &\quad- \frac{1}{m}\sum_i a_{ij}v_i^{(t-1)}\sum_{l\neq j}\eta^{(t-1)\prime}(r_{il}^{(t-1)}) + \sum_i a_{ij}(w_i + \mu_i^{(t)})\end{aligned} \tag{18}$$

$$\begin{aligned}&= \sum_i a_{ij}\sum_{l\neq j} a_{il}\left[x_l - \eta^{(t-1)}(r_{il}^{(t-1)})\right]\\ &\quad+ \sum_i a_{ij}w_i + \sum_i a_{ij}\left[\mu_i^{(t)} - v_i^{(t-1)}\frac{1}{m}\sum_{l\neq j}\eta^{(t-1)\prime}(r_{il}^{(t-1)})\right].\end{aligned} \tag{19}$$

We are now in a position to observe the principal mechanism of AMP. As we argue below (using the central limit theorem), the first and second terms in (19) behave like realizations of zero-mean Gaussians in the large-system limit, because $\{a_{il}\}$ are realizations of i.i.d. zero-mean random variables $\{A_{il}\}$ that are independent of x_l, w_i, and $\{r_{il}^{(t-1)}\}$ under Assumption 1. But the same cannot be said in general about the third term in (19), because $v_i^{(t-1)}$ is strongly coupled to a_{ij}. Consequently, the denoiser input-error $e_j^{(t)}$ is difficult to characterize for general $\mu_i^{(t)}$.

With AMP's choice of $\mu_i^{(t)}$, however, the 3 rd term in (19) vanishes in the large-system limit. In particular, with the Onsager choice (4), the 3rd term in (19) takes the form

$$\begin{aligned}&\sum_i a_{ij}\left[\frac{v_i^{(t-1)}}{m}\sum_l \eta^{(t-1)\prime}(r_l^{(t-1)}) - \frac{v_i^{(t-1)}}{m}\sum_{l\neq j}\eta^{(t-1)\prime}(r_{il}^{(t-1)})\right]\\ &= \frac{1}{m}\sum_i a_{ij}v_i^{(t-1)}\left[\eta^{(t-1)\prime}(r_j^{(t-1)})\right.\\ &\quad\left.+\sum_{l\neq j}\left(\eta^{(t-1)\prime}(r_l^{(t-1)}) - \eta^{(t-1)\prime}(r_{il}^{(t-1)})\right)\right]\end{aligned} \tag{20}$$

$$\begin{aligned}&\approx \frac{1}{m}\sum_i a_{ij}v_i^{(t-1)}\left[\eta^{(t-1)\prime}(r_j^{(t-1)})\right.\\ &\quad\left.+\sum_{l\neq j} a_{il}v_i^{(t-1)}\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)})\right],\end{aligned} \tag{21}$$

where for the last step we used the Taylor-series expansion

$$\begin{aligned}&\eta^{(t-1)\prime}(r_l^{(t-1)})\\ &= \eta^{(t-1)\prime}(r_{il}^{(t-1)} + a_{il}v_i^{(t-1)})\end{aligned} \tag{22}$$

$$= \eta^{(t-1)\prime}(r_{il}^{(t-1)}) + a_{il}v_i^{(t-1)}\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)}) + O(1/m) \tag{23}$$

and dropped the $O(1/m)$ term, since it will vanish relative to the $a_{il}v_i^{(t-1)}\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)})$ term in the large-system limit. Looking at (21), the first term is

$$\frac{1}{m}\sum_{i=1}^{m}\underbrace{a_{ij}v_i^{(t-1)}\eta^{(t-1)\prime}(r_j^{(t-1)})}_{O(1/\sqrt{m})} = O(1/\sqrt{m}) \tag{24}$$

since $a_{ij} \in \pm 1/\sqrt{m}$ and $v_i^{(t-1)}\eta^{(t-1)\prime}(r_j^{(t-1)})$ is $O(1)$ due to Lemma 1. Thus the first term in (21) will vanish in the large-system limit. The second term in (21) is

$$\underbrace{\frac{1}{m}\sum_{i=1}^{m}a_{ij}(v_i^{(t-1)})^2\underbrace{\sum_{l\neq j}a_{il}\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)})}_{O(1)}}_{O(1/\sqrt{m})} = O(1/\sqrt{m}), \tag{25}$$

which will also vanish in the large-system limit. The $O(1)$ scaling in (25) follows from Lemma 1 under Assumption 1, and the $O(1/\sqrt{m})$ scaling follows from the fact that $a_{il} \in \pm 1/\sqrt{m}$ and $(v_i^{(t-1)})^2 = O(1)$.

Thus, for large m and the AMP choice of $\mu_i^{(t)}$, equation (19) becomes

$$e_j^{(t)} \approx \sum_i a_{ij}\sum_{l\neq j}a_{il}\underbrace{[x_l - \eta^{(t-1)}(r_{il}^{(t-1)})]}_{\triangleq\, \epsilon_{il}^{(t)}} + \sum_i a_{ij}w_i. \tag{26}$$

Recall that, under Assumption 1, a_{ij} is a realization of equiprobable $A_{ij} \in \pm\frac{1}{\sqrt{m}}$ that is independent of $\{A_{il}\}_{l\neq j}$, $\{x_l\}_{l=1}^n$, and $\{r_{il}^{(t-1)}\}_{l=1}^n$. This implies that A_{ij} is also independent of $\epsilon_{il}^{(t)} = x_l - \eta^{(t-1)}(r_{il}^{(t-1)})$ for any l. Thus we can apply the central limit theorem to say that, for any fixed $\{\epsilon_{il}^{(t)}\}$, the first term converges to a Gaussian with mean and variance

$$\mathbb{E}\left[\sum_i A_{ij}\sum_{l\neq j}A_{il}\epsilon_{il}^{(t)}\right] = \sum_i \mathbb{E}[A_{ij}]\sum_{l\neq j}\mathbb{E}[A_{il}]\epsilon_{il}^{(t)} \tag{27}$$

$$= 0 \tag{28}$$

$$\mathbb{E}\left[\left(\sum_i A_{ij}\sum_{l\neq j}A_{il}\epsilon_{il}^{(t)}\right)^2\right] = \sum_i \mathbb{E}[A_{ij}^2]\sum_{l\neq j}\mathbb{E}[A_{il}^2](\epsilon_{il}^{(t)})^2 \tag{29}$$

$$= \frac{1}{m^2}\sum_i\sum_{l\neq j}(\epsilon_{il}^{(t)})^2. \tag{30}$$

From the Taylor expansion (14), we have

$$\epsilon_{il}^{(t)} = x_l - \eta^{(t-1)}(r_{il}^{(t-1)}) \tag{31}$$

$$= \underbrace{x_l - \eta^{(t-1)}(r_l^{(t-1)})}_{\triangleq\, \epsilon_l^{(t)}} + \underbrace{a_{il}v_i^{(t-1)}\eta^{(t-1)\prime}(r_{il}^{(t-1)}) + O(1/m)}_{O(1/\sqrt{m})}, \tag{32}$$

where the $O(1/\sqrt{m})$ scaling follows from the facts that $a_{il} \in \pm 1/\sqrt{m}$ and $v_i^{(t-1)}\eta^{(t-1)\prime}(r_{il}^{(t-1)})$ is $O(1)$. Notice that $\epsilon_l^{(t)}$ is the denoiser output error. Because both x_l and $r_l^{(t-1)}$ are $O(1)$, it follows that $\epsilon_l^{(t)}$ is also $O(1)$. Because the $O(1/\sqrt{m})$ term in (32) vanishes in the large-system limit, we see that (30) becomes

$$\frac{1}{m^2}\sum_i\sum_{l\neq j}(\epsilon_{il}^{(t)})^2 \approx \frac{1}{m^2}\sum_{i=1}^{m}\sum_{l\neq j}(\epsilon_l^{(t)})^2 = \frac{1}{m}\sum_{l\neq j}(\epsilon_l^{(t)})^2 \tag{33}$$

$$= \underbrace{\frac{n}{m}\frac{1}{n}\sum_{l=1}^{n}(\epsilon_l^{(t)})^2}_{O(1)} - \underbrace{\frac{1}{m}(\epsilon_j^{(t)})^2}_{O(1/m)} \approx \delta^{-1}\mathcal{E}^{(t)}, \tag{34}$$

where

$$\mathcal{E}^{(t)} \triangleq \lim_{n\to\infty}\frac{1}{n}\sum_{l=1}^{n}(\epsilon_l^{(t)})^2 \tag{35}$$

is the average squared error on the denoiser output $\boldsymbol{x}^{(t)}$. We have thus deduced that, in the large-system limit, the first term in (26) behaves like a zero-mean Gaussian with variance $\delta^{-1}\mathcal{E}^{(t)}$. For the second term in (26), we can again use the central limit theorem to say that, for any fixed $\{w_i\}$, the second term converges to a Gaussian with mean and variance

$$\mathbb{E}\left[\sum_i A_{ij}w_i\right] = \sum_i \mathbb{E}[A_{ij}]w_i = 0 \tag{36}$$

$$\mathbb{E}\left[\left(\sum_i A_{ij}w_i\right)^2\right] = \sum_i \mathbb{E}[A_{ij}^2]w_i^2 = \frac{1}{m}\sum_{i=1}^{m}w_i^2 \approx \tau_w, \tag{37}$$

where τ_w denotes the empirical second moment of the noise:

$$\tau_w \triangleq \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}w_i^2. \tag{38}$$

To summarize, with AMP's choice of $\boldsymbol{\mu}^{(t)}$ from (4), the jth component of the denoiser input-error behaves like

$$e_j^{(t)} \sim \mathcal{N}(0, \underbrace{\delta^{-1}\mathcal{E}^{(t)} + \tau_w}_{\triangleq\, \tau_r^{(t)}}) \tag{39}$$

in the large-system limit, where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian random variable with mean μ and variance σ^2. With other choices of $\boldsymbol{\mu}^{(t)}$ (e.g., IST's choice of $\boldsymbol{\mu}^{(t)} = \mathbf{0}\,\forall t$), it is difficult to characterize the denoiser input-error $\boldsymbol{e}^{(t)}$ and in general it will not be Gaussian.

IV. AMP State Evolution

In Section III, we used Assumption 1 to argue that the AMP algorithm yields a denoiser input-error $\boldsymbol{e}^{(t)}$ whose components are $\mathcal{N}(0, \tau_r^{(t)})$ in the large system limit. Here, $\tau_r^{(t)} = \delta^{-1}\mathcal{E}^{(t)} + \tau_w$ where $\mathcal{E}^{(t)}$ is the average squared-error at the denoiser output in the large-system limit.

Recalling the definition of $\mathcal{E}^{(t)}$ from (35), we can write

$$\frac{1}{n}\sum_{l=1}^{n}(\epsilon_l^{(t)})^2 \approx \frac{1}{n}\sum_{l=1}^{n}\left[\eta^{(t-1)}(x_l + \mathcal{N}(0, \tau_r^{(t-1)})) - x_l\right]^2 \tag{40}$$

$$= \mathbb{E}\left[\eta^{(t-1)}(X + \mathcal{N}(0, \tau_r^{(t-1)})) - X\right]^2 \tag{41}$$

where X is a scalar random variable defined from the empirical distribution

$$X \sim p(x) = \frac{1}{n}\sum_{l=1}^{n}\delta(x - x_l), \tag{42}$$

with $\delta(\cdot)$ denoting the Dirac delta function. Thus we can argue that, in the large-system limit,

$$\mathcal{E}^{(t)} = \mathbb{E}\left[\eta^{(t-1)}\left(X + \mathcal{N}(0, \tau_r^{(t-1)})\right) - X\right]^2, \tag{43}$$

where X now is distributed according to the $n \to \infty$ limit of the empirical distribution. Combining (43) with the update equation for $\tau_r^{(t)}$ gives the following recursion for $t = 0, 1, 2, \ldots$:

$$\tau_r^{(t)} = \delta^{-1}\mathcal{E}^{(t)} + \tau_w \tag{44a}$$

$$\mathcal{E}^{(t+1)} = \mathbb{E}\left[\eta^{(t)}\left(X + \mathcal{N}(0, \tau_r^{(t)})\right) - X\right]^2, \tag{44b}$$

initialized with $\mathcal{E}^{(0)} = \mathbb{E}[X^2]$. The recursion (44) is known as AMP's "state evolution" for the mean-squared error [1]–[3].

The reason that we call $\eta^{(t)}(\cdot)$ a "denoiser" should now be clear. To minimize the mean-squared error $\mathcal{E}^{(t+1)}$, the function $\eta^{(t)}(\cdot)$ should remove as much of the noise from its input as possible. The smaller that $\mathcal{E}^{(t+1)}$ is, the smaller the input-noise variance $\tau_r^{(t+1)}$ will be during the next iteration.

V. AMP Variance Estimation

For best performance, the iteration-t denoiser $\eta^{(t)}(\cdot)$ should be designed in accordance with the iteration-t input noise variance $\tau_r^{(t)}$. With the AMP algorithm, there is an easy way to estimate the value of $\tau_r^{(t)}$ at each iteration t from the $\boldsymbol{v}^{(t)}$ vector, i.e., $\tau_r^{(t)} \approx \|\boldsymbol{v}^{(t)}\|^2/m$ [10]. Below, we explain this approach using arguments similar to those used above.

Equation (62) shows that

$$v_i^{(t)} = y_i - \sum_{l=1}^{n} a_{il}\eta^{(t-1)}(r_{il}^{(t-1)}) - v_i^{(t-1)}\frac{1}{m}\sum_{l=1}^{n}\eta^{(t-1)\prime}(r_{il}^{(t-1)}) + \mu_i^{(t)} + O(1/m). \tag{45}$$

Ignoring the $O(1/m)$ term and plugging in AMP's choice of $\mu_i^{(t)}$ from (4) yields

$$\begin{aligned} v_i^{(t)} &\approx y_i - \sum_{l=1}^{n} a_{il}\eta^{(t-1)}(r_{il}^{(t-1)}) + v_i^{(t-1)}\frac{1}{m}\sum_{l=1}^{n}\left[\eta^{(t-1)\prime}(r_l^{(t-1)}) - \eta^{(t-1)\prime}(r_{il}^{(t-1)})\right] \qquad (46)\\ &= y_i - \sum_{l=1}^{n} a_{il}\eta^{(t-1)}(r_{il}^{(t-1)}) + v_i^{(t-1)}\frac{n}{m}\frac{1}{n}\sum_{l=1}^{n}\underbrace{\left[a_{il}v_i^{(t-1)}\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)}) + O(1/m)\right]}_{O(1/\sqrt{m})}, \qquad (47)\end{aligned}$$

where we used the Taylor series (23) in the second step and $a_{il} \in \pm 1/\sqrt{m}$ to justify the $O(1/\sqrt{m})$ scaling. Since the last term in (47) is the scaled average of $O(1/\sqrt{m})$ terms, with $O(1)$ scaling, the entire term is $O(1/\sqrt{m})$. We can thus drop it since it will vanish relative to the others in the large-system limit. Doing this and plugging in $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ yields

$$v_i^{(t)} \approx w_i + \sum_{l=1}^{n} a_{il}\underbrace{[x_l - \eta^{(t-1)}(r_{il}^{(t-1)})]}_{=\epsilon_{il}^{(t)}}, \tag{48}$$

recalling the definition of $\epsilon_{il}^{(t)}$ from (26). Squaring the result and averaging over i yields

$$\frac{1}{m}\sum_{i=1}^{m}(v_i^{(t)})^2 \approx \frac{1}{m}\sum_{i=1}^{m}w_i^2 + \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{l=1}^{n}a_{il}\epsilon_{il}^{(t)}\right)^2 + \frac{2}{m}\sum_{i=1}^{m}\left(w_i\sum_{l=1}^{n}a_{il}\left[x_l - \eta^{(t-1)}(r_{il}^{(t-1)})\right]\right). \tag{49}$$

We now examine the components of (49) in the large-system limit. By definition, the first term in (49) converges to τ_w. By the law of large numbers, the second term converges to

$$\lim_{n\to\infty}\mathbb{E}\left[\left(\sum_{l=1}^{n}A_{il}\epsilon_{il}^{(t)}\right)^2\right] = \lim_{n\to\infty}\sum_{l=1}^{n}\sum_{j=1}^{n}\mathbb{E}[A_{il}A_{ij}]\epsilon_{il}^{(t)}\epsilon_{ij}^{(t)} \tag{50}$$

$$= \lim_{n\to\infty}\frac{1}{m}\sum_{l=1}^{n}(\epsilon_{il}^{(t)})^2, \tag{51}$$

since $\mathbb{E}[A_{il}A_{ij}] = 1/m$ when $l = j$ and $\mathbb{E}[A_{il}A_{ij}] = 0$ when $l \neq j$. Using the relationship between $\epsilon_{il}^{(t)}$ and $\epsilon_l^{(t)}$ from (32), it can be seen that

$$\lim_{n\to\infty}\frac{1}{m}\sum_{l=1}^{n}(\epsilon_{il}^{(t)})^2 = \lim_{n\to\infty}\frac{n}{m}\frac{1}{n}\sum_{l=1}^{n}(\epsilon_l^{(t)})^2 = \delta^{-1}\mathcal{E}^{(t)} \tag{52}$$

where m is implicitly a function of n because $m/n = O(1)$. In summary,

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}(v_i^{(t)})^2 = \tau_w + \delta^{-1}\mathcal{E}^{(t)} = \tau_r^{(t)}, \tag{53}$$

which shows that $\tau_r^{(t)}$ is well estimated by $\|\boldsymbol{v}^{(t)}\|^2/m$ in the large-system limit.

VI. Numerical Experiments

We now present numerical experiments that demonstrate the AMP behaviors discussed above. In all experiments, we used a sampling ratio of $\delta = 0.5$, $\{A_{ij}\}$ drawn i.i.d. zero-mean Gaussian with variance $1/m$, $\{x_j\}$ drawn i.i.d. from the Bernoulli-Gaussian distribution with sparsity rate $\beta = 0.1$ (i.e., $p_X(x_j) = (1-\beta)\delta(x_j) + \beta\mathcal{N}(x_j; 0, 1)\,\forall j$, where $\delta(\cdot)$ denotes the Dirac delta distribution), and $\{w_i\}$ drawn i.i.d. zero-mean Gaussian with variance $\beta 10^{-\text{SNRdB}/10}$ and SNRdB $= 20$, so that $\mathbb{E}[\|\boldsymbol{A}\boldsymbol{x}\|^2]/\mathbb{E}[\|\boldsymbol{w}\|^2] \approx 20$ dB. We experimented with two denoisers: the MMSE denoiser $\eta^{(t)}(r_j) = \mathbb{E}[X \mid r_j = X + \mathcal{N}(0, \tau_r^{(t)})]$ and the soft-thresholding denoiser $\eta^{(t)}(r_j) = \text{sgn}(r_j)\max\{0, |r_j| - \alpha\sqrt{\tau_r^{(t)}}\}$ with $\alpha = 1.14$, which is the

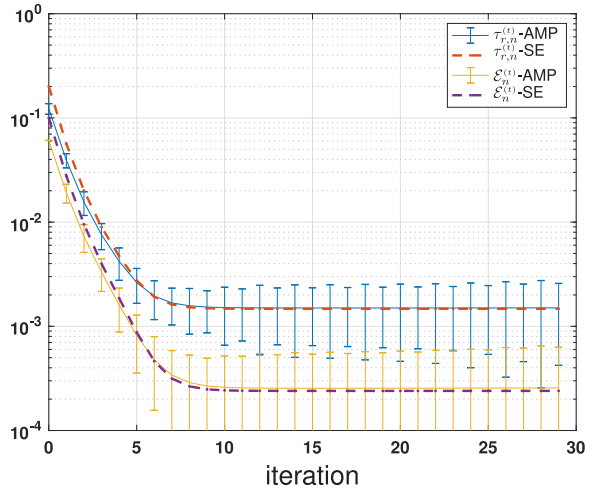

Fig. 1. Denoiser output error $\mathcal{E}_n^{(t)}$ and denoiser input-error variance $\tau_{r,n}^{(t)}$ versus iteration for AMP and its state evolution with MMSE denoising and $n = 300$. Dashed lines show the empirical average over 10000 random draws of $\boldsymbol{A}$ and error bars show the empirical standard deviation.

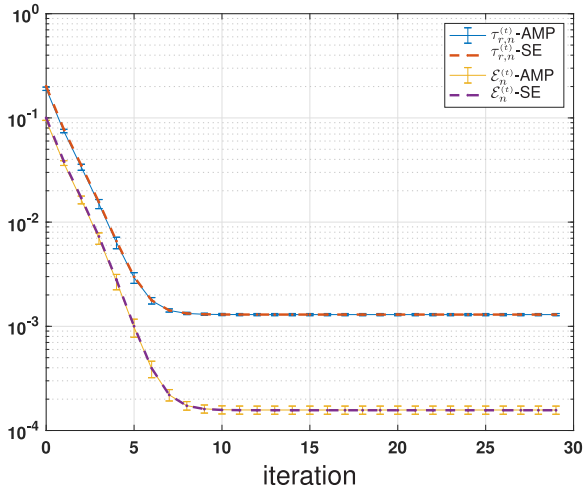

Fig. 2. Denoiser output MSE $\mathcal{E}_n^{(t)}$ and denoiser input-error variance $\tau_{r,n}^{(t)}$ versus iteration for AMP and its state evolution with MMSE denoising and $n = 3000$. Dashed lines show the empirical average over 1000 random draws of $\boldsymbol{A}$ and error bars show the empirical standard deviation.

minimax choice, i.e., the value of α that minimizes the maximum MSE over all 0.1-sparse signals (see [10] for more details). With the soft-thresholding denoiser, AMP solves the LASSO problem "$\arg\min_{\boldsymbol{x}}\{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|^2+\lambda\|\boldsymbol{x}\|_1\}$" for some value of λ [1], [10].

Figs. 1–6 plot finite-dimensional versions of the denoiser output MSE $\mathcal{E}^{(t)}$ and the denoiser input-error variance $\tau_r^{(t)}$ versus iteration t for both the AMP algorithm (2) and the AMP state evolution (44). For the AMP algorithm, the iteration-t denoiser output MSE was computed as $\mathcal{E}_n^{(t)} = \frac{1}{n}\sum_{j=1}^{n}(x_j - x_j^{(t)})^2$ and the denoiser input-error variance was computed as $\tau_{r,n}^{(t)} = \|\boldsymbol{v}^{(t)}\|^2/m$, where the subscript n indicates the dimensional dependence of these quantities. For the AMP state evolution, the denoiser output MSE was computed as

$$\mathcal{E}_n^{(t)} = \begin{cases} \mathbb{E}\left[\eta^{(t-1)}(X+\mathcal{N}(0,\tau_{r,n}^{(t-1)})) - X\right]^2 & t > 0 \\ \mathbb{E}\left[X^2\right] & t = 0, \end{cases} \tag{54}$$

with the expectation evaluated using the n-term empirical distribution for X, and the iteration-t denoiser input-error variance was computed as $\tau_{r,n}^{(t)} = \delta^{-1}\mathcal{E}_n^{(t)} + \tau_{w,n}$ using the empirical noise variance $\tau_{w,n} = \frac{1}{m}\sum_{i=1}^{m} w_i^2$. Each figure plots the empirical mean and standard deviation over $T \in \{100, 1000, 10000\}$ random draws of $\boldsymbol{A}$ for a *single fixed* draw of $\boldsymbol{x}$ and $\boldsymbol{w}$.

Fig. 1 shows the results for the MMSE denoiser at dimension $n = 300$. The figure shows a good, but not great, agreement between the state evolution and average AMP quantities, where the average was computed over $T = 10000$ realizations of $\boldsymbol{A}$. Furthermore, the error bars in Fig. 1, which show the empirical standard deviation over the T realizations, indicate that there was considerable dependence of the trajectories $\{\mathcal{E}_n^{(t)}\}_{t=1}^{30}$ and $\{\tau_{r,n}^{(t)}\}_{t=1}^{30}$ on the realization of $\boldsymbol{A}$ when $n = 300$. Figs. 2 and 3 plot the same quantities for dimensions $n = 3000$ and $n = 30000$, respectively. These figures show that, as the dimension n increases, the agreement between the state-evolution and average AMP trajectories improves and the standard deviation

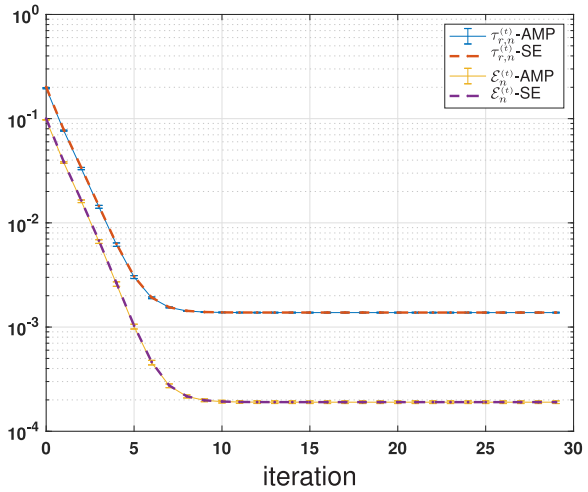

Fig. 3. Denoiser output MSE $\mathcal{E}_n^{(t)}$ and denoiser input-error variance $\tau_{r,n}^{(t)}$ versus iteration for AMP and its state evolution with MMSE denoising and $n = 30000$. Dashed lines show the empirical average over 100 random draws of $\boldsymbol{A}$ and error bars show the empirical standard deviation.

of the AMP trajectories decreases. Table I suggests that the standard deviation decreases proportional to $1/\sqrt{n}$. Fig. 3 shows that, when $n = 30000$, the trajectories $\{\mathcal{E}_n^{(t)}\}_{t=1}^{30}$ and $\{\tau_{r,n}^{(t)}\}_{t=1}^{30}$ are nearly invariant to changes in $\boldsymbol{A}$.

Figs. 4–6 are similar to Figs. 1–3, except that they show the results for the soft-thresholding denoiser. As expected, the use of the soft-thresholding denoiser results in larger MSEs than the MMSE denoiser. But, otherwise, the trends are the same: the agreement between the state-evolution and average AMP trajectories increases with the dimension n, and the standard deviation of the AMP trajectories decreases proportional to $1/\sqrt{n}$.

TABLE I
NUMERICAL EVIDENCE THAT STD$(\mathcal{E}_n^{(t)})\sqrt{n}$ AND STD$(\tau_{r,n}^{(t)})\sqrt{n}$ ARE APPROXIMATELY CONSTANT WITH n, IMPLYING THAT STD$(\mathcal{E}_n^{(t)})$ AND STD$(\tau_{r,n}^{(t)})$ SCALE AS APPROXIMATELY $1/\sqrt{n}$

	n	1000	3000	10000	30000
MMSE denoiser	std$(\mathcal{E}_n^{(29)})\sqrt{n}$	0.0011	0.0008	0.0010	0.0010
	std$(\tau_{r,n}^{(29)})\sqrt{n}$	0.0026	0.0017	0.0022	0.0020
Soft Threshold	std$(\mathcal{E}_n^{(29)})\sqrt{n}$	0.0055	0.0044	0.0037	0.0040
	std$(\tau_{r,n}^{(29)})\sqrt{n}$	0.0118	0.0098	0.0106	0.0093

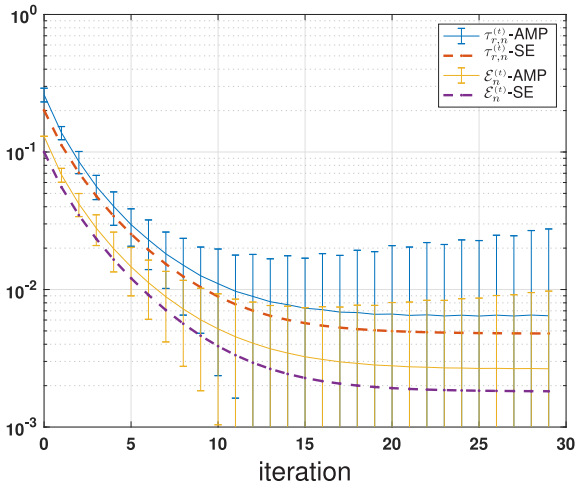

Fig. 4. Denoiser output MSE $\mathcal{E}_n^{(t)}$ and denoiser input-error variance $\tau_{r,n}^{(t)}$ versus iteration for AMP and its state evolution with soft-threshold denoising and $n = 300$. Dashed lines show the empirical average over 10000 random draws of $\boldsymbol{A}$ and error bars show the empirical standard deviation.

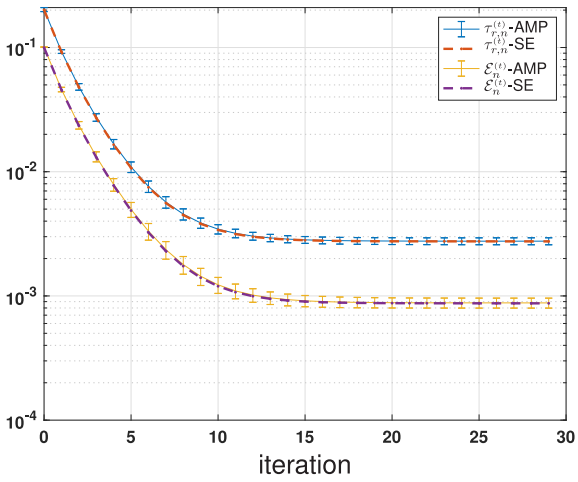

Fig. 5. Denoiser output MSE $\mathcal{E}_n^{(t)}$ and denoiser input-error variance $\tau_{r,n}^{(t)}$ versus iteration for AMP and its state evolution with soft-threshold denoising and $n = 3000$. Dashed lines show the empirical average over 1000 random draws of $\boldsymbol{A}$ and error bars show the empirical standard deviation.

Note that, because the state evolution was computed using the empirical distributions of $\{x_j\}_{j=1}^n$ and $\{w_i\}_{i=1}^m$, which change from one figure to the next (e.g., as n and m change), the state evolution trajectories vary across Figs. 1–6.

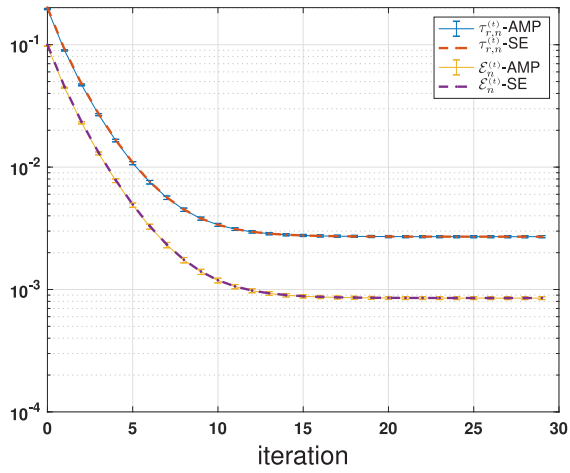

Fig. 6. Denoiser output MSE $\mathcal{E}_n^{(t)}$ and denoiser input-error variance $\tau_{r,n}^{(t)}$ versus iteration for AMP and its state evolution with soft-threshold denoising and $n = 30000$. Dashed lines show the empirical average over 100 random draws of $\boldsymbol{A}$ and error bars show the empirical standard deviation.

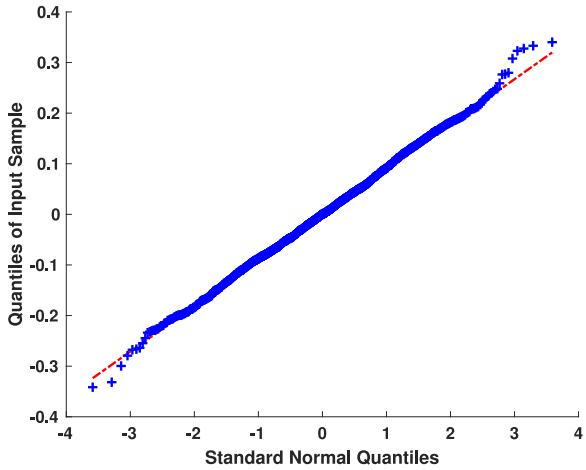

Fig. 7. Quantiles of standard normal versus quantiles of AMP denoiser input-error $\{e_j^{(t)}\}$ at iteration $t = 5$ with MMSE denoising and $n = 3000$.

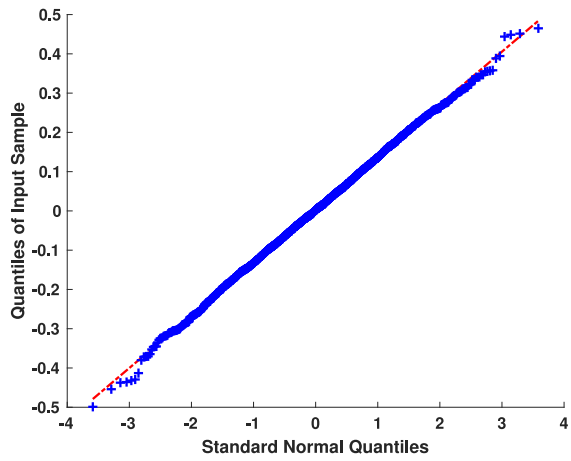

Fig. 8. Quantiles of standard normal versus quantiles of AMP denoiser input-error $\{e_j^{(t)}\}$ at iteration $t = 5$ with soft-threshold denoising and $n = 3000$.

Finally, to give evidence that the denoiser input error $\{e_j^{(t)}\}$ is approximately Gaussian, we show quantile-quantile (QQ) plots in Figs. 7 and 8 at iteration $t = 5$ and dimension $n = 3000$ for the MMSE and soft-thresholding denoisers, respectively. The figures show that the quantiles of $\{e_j^{(t)}\}$ are very close to those of a zero-mean Gaussian random variable. Although not shown here, QQ plots at other iterations t look similar, and the QQ plots become more linear (i.e., $\{e_j^{(t)}\}$ looks more Gaussian) as n grows larger.

VII. Conclusion

For the linear regression problem, we presented a simple derivation of AMP and its MSE state-evolution based on the idea of "first-order cancellation." In particular, our derivation writes the linear transform of the denoiser output, $\boldsymbol{A}\boldsymbol{x}^{(t)}$, as the sum of a term that is weakly dependent on the previous iteration and another term that is strongly dependent but canceled by AMP's Onsager correction term in the large-system limit. Our derivation provides insights that are missing from the usual loopy belief-propagation derivation of AMP, while being much more accessible than Bayati et al.'s rigorous analysis of AMP.

Appendix

Here we establish Lemma 2, which says that the elements of $\boldsymbol{v}^{(t)}, \boldsymbol{r}^{(t)}, \boldsymbol{x}^{(t)}$, and $\boldsymbol{\mu}^{(t)}$ scale as $O(1)$ in the large-system limit under Assumption 1. We do this by induction.

From the initialization $\boldsymbol{x}^{(0)} = \boldsymbol{0} = \boldsymbol{\mu}^{(0)}$, we have that $\boldsymbol{v}^{(0)} = \boldsymbol{y}$, whose elements are $O(1)$, and we have $\boldsymbol{r}^{(0)} = \boldsymbol{A}^\top \boldsymbol{y} = \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^\top \boldsymbol{w}$. Examining the jth entry, we see that

$$r_j^{(0)} = \sum_i a_{ij} \sum_l a_{il} x_l + \sum_i a_{ij} w_i \tag{55}$$

$$= x_j \underbrace{\sum_{i=1}^m a_{ij}^2}_{=1} + \underbrace{\sum_i a_{ij} \underbrace{\sum_{l \neq j} a_{il} x_l}_{O(1)}}_{O(1)} + \underbrace{\sum_i a_{ij} w_i}_{O(1)} \tag{56}$$

since $a_{ij}^2 = 1/m \, \forall ij$ and where the $O(1)$ scalings follow from Lemma 1 under Assumption 1. Thus the elements of $\boldsymbol{r}^{(0)}$ are $O(1)$. Because $x_j^{(1)} = \eta^{(0)}(r_j^{(0)})$, the elements of $\boldsymbol{x}^{(1)}$ are also $O(1)$. And from (4),

$$\mu_i^{(1)} = \frac{n}{m} v_i^{(0)} \frac{1}{n} \sum_{j=1}^n \eta^{(0)\prime}(r_j^{(0)}), \tag{57}$$

where $\frac{n}{m}$, $v_i^{(0)}$, and $\frac{1}{n}\sum_{j=1}^n \eta^{(0)\prime}(r_j^{(0)})$ are all $O(1)$, implying that the elements of $\boldsymbol{\mu}^{(1)}$ are $O(1)$.

Now, suppose that the elements of $\boldsymbol{r}^{(t-1)}, \boldsymbol{v}^{(t-1)}, \boldsymbol{x}^{(t)}, \boldsymbol{\mu}^{(t)}$ are all $O(1)$, which we know occurs when $t = 1$. Then from (2) we have that

$$v_i^{(t)} = y_i - \sum_{l=1}^n a_{il} x_l^{(t)} + \mu_i^{(t)} \tag{58}$$

$$= y_i - \sum_{l=1}^n a_{il} \eta^{(t-1)}(r_l^{(t-1)}) + \mu_i^{(t)} \tag{59}$$

$$= y_i - \sum_{l=1}^n a_{il} \eta^{(t-1)} \left(\underbrace{x_l^{(t-1)} + \sum_{k \neq i} a_{kl} v_k^{(t-1)}}_{= r_{il}^{(t-1)}} + a_{il} v_i^{(t-1)} \right) + \mu_i^{(t)}, \tag{60}$$

where $r_{il}^{(t-1)}$ is only weakly dependent on $\{a_{ij}\}_{j=1}^n$, leading us to invoke Assumption 1. The Taylor expansion (14) (which is admissible since, under the induction hypothesis, the elements of $\boldsymbol{v}^{(t-1)}$ and $\boldsymbol{r}^{(t-1)}$ scale as $O(1)$) then yields

$$v_i^{(t)} = y_i - \sum_{l=1}^n a_{il} \left[\eta^{(t-1)}(r_{il}^{(t-1)}) + a_{il} v_i^{(t-1)} \eta^{(t-1)\prime}(r_{il}^{(t-1)}) + O(1/m) \right] + \mu_i^{(t)} \tag{61}$$

$$= y_i - \underbrace{\sum_{l=1}^n a_{il} \eta^{(t-1)}(r_{il}^{(t-1)})}_{O(1)} - v_i^{(t-1)} \underbrace{\frac{1}{m} \sum_{l=1}^n \eta^{(t-1)\prime}(r_{il}^{(t-1)})}_{O(1)} + O(1/m) \underbrace{\sum_{l=1}^n a_{il}}_{O(1)} + \mu_i^{(t)}, \tag{62}$$

where we used the fact that $a_{il}^2 = 1/m \, \forall il$. In (62), the $O(1)$ scaling of the first and last sums follows from Lemma 1 under Assumption 1. Thus the elements of $\boldsymbol{v}^{(t)}$ are $O(1)$.

Next, we establish that the elements of $\boldsymbol{r}^{(t)}$ are $O(1)$ when the elements of $\boldsymbol{v}^{(t-1)}, \boldsymbol{x}^{(t)}$ and $\boldsymbol{\mu}^{(t)}$ are. From (5)–(8) we have

$$\boldsymbol{r}^{(t)} = (\boldsymbol{I} - \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{x}^{(t)} - (\boldsymbol{I} - \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{x} + \boldsymbol{A}^\top(\boldsymbol{w} + \boldsymbol{\mu}^{(t)}) + \boldsymbol{x} \tag{63}$$

$$= (\boldsymbol{I} - \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{x}^{(t)} + \boldsymbol{A}^\top \boldsymbol{\mu}^{(t)} + \underbrace{\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^\top \boldsymbol{w}}_{= \boldsymbol{r}^{(0)}}, \tag{64}$$

where we previously established that the elements in $\boldsymbol{r}^{(0)}$ are $O(1)$. As for the remaining term in (64), we have from (13) and (4) that

$$\left[(\boldsymbol{I} - \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{x}^{(t)} + \boldsymbol{A}^\top \boldsymbol{\mu}^{(t)} \right]_j = -\sum_i a_{ij} \sum_{l \neq j} a_{il} \eta^{(t-1)} \left(r_{il}^{(t-1)} + a_{il} v_i^{(t-1)} \right) + \sum_i a_{ij} \frac{1}{m} v_i^{(t-1)} \sum_l \eta^{(t-1)\prime} \left(r_l^{(t-1)} \right) \tag{65}$$

$$= \sum_i a_{ij} \left[\frac{1}{m} v_i^{(t-1)} \sum_l \eta^{(t-1)\prime} \left(r_l^{(t-1)} \right) - \sum_{l \neq j} a_{il} \eta^{(t-1)} \left(r_{il}^{(t-1)} + a_{il} v_i^{(t-1)} \right) \right] \tag{66}$$

$$= \sum_i a_{ij} \Bigg[\frac{1}{m} v_i^{(t-1)} \sum_l \eta^{(t-1)\prime}\left(r_l^{(t-1)}\right) - \sum_{l\neq j} a_{il}\eta^{(t-1)}\left(r_{il}^{(t-1)}\right) - \sum_{l\neq j} a_{il}^2 v_i^{(t-1)} \eta^{(t-1)\prime}\left(r_{il}^{(t-1)}\right) + \underbrace{O(1/m)\sum_{l\neq j} a_{il}}_{O(1/\sqrt{m})} \Bigg], \tag{67}$$

where, for the last step, we applied the Taylor expansion (14). Applying $a_{il}^2 = 1/m \,\forall il$ and rearranging, we get

$$\left[(\boldsymbol{I} - \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{x}^{(t)} + \boldsymbol{A}^\top \boldsymbol{\mu}^{(t)}\right]_j$$

$$= \sum_i a_{ij} \Bigg[\frac{v_i^{(t-1)}}{m} \sum_l \eta^{(t-1)\prime}\left(r_l^{(t-1)}\right) - \frac{v_i^{(t-1)}}{m} \sum_{l\neq j} \eta^{(t-1)\prime}\left(r_{il}^{(t-1)}\right) - \sum_{l\neq j} a_{il}\eta^{(t-1)}\left(r_{il}^{(t-1)}\right) + O(1/\sqrt{m}) \Bigg] \tag{68}$$

$$= \frac{1}{m}\sum_i a_{ij} v_i^{(t-1)} \left[\sum_l \eta^{(t-1)\prime}\left(r_l^{(t-1)}\right) - \sum_{l\neq j} \eta^{(t-1)\prime}\left(r_{il}^{(t-1)}\right) \right] - \sum_i a_{ij} \sum_{l\neq j} a_{il}\eta^{(t-1)}\left(r_{il}^{(t-1)}\right) + O(1) \tag{69}$$

$$= \frac{1}{m}\sum_i a_{ij} v_i^{(t-1)} \Bigg[\eta^{(t-1)\prime}\left(r_j^{(t-1)}\right) + \sum_{l\neq j} \left(\eta^{(t-1)\prime}\left(r_l^{(t-1)}\right) - \eta^{(t-1)\prime}\left(r_{il}^{(t-1)}\right)\right) \Bigg] - \sum_i a_{ij} \sum_{l\neq j} a_{il}\eta^{(t-1)}\left(r_{il}^{(t-1)}\right) + O(1), \tag{70}$$

where the $O(1)$ term follows from the fact that $a_{ij} \in \pm 1/\sqrt{m}$. For the 2nd-to-last term in (70), we have

$$\sum_i a_{ij} \underbrace{\sum_{l\neq j} a_{il}\eta^{(t-1)}\left(r_{il}^{(t-1)}\right)}_{O(1)} = O(1), \tag{71}$$

which follows from Lemma 1 under Assumption 1. For the remaining term in (70), we apply the Taylor expansion (23) and rearrange terms as follows:

$$\frac{1}{m}\sum_i a_{ij} v_i^{(t-1)} \left[\eta^{(t-1)'}\left(r_j^{(t-1)}\right) + \sum_{l\neq j} \left(\eta^{(t-1)'}\left(r_l^{(t-1)}\right) - \eta^{(t-1)'}\left(r_{il}^{(t-1)}\right)\right) \right]$$

$$= \frac{1}{m}\sum_i a_{ij} v_i^{(t-1)} \left[\eta^{(t-1)\prime}\left(r_j^{(t-1)}\right) + \sum_{l\neq j} \left(a_{il} v_i^{(t-1)} \eta^{(t-1)\prime\prime}(r_{il}^{(t-1)}) + O(1/m)\right) \right] \tag{72}$$

$$= \frac{1}{m}\sum_{i=1}^m a_{ij} v_i^{(t-1)} \eta^{(t-1)\prime}\left(r_j^{(t-1)}\right) + \frac{1}{m}\sum_{i=1}^m a_{ij}(v_i^{(t-1)})^2 \underbrace{\sum_{l\neq j} a_{il}\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)})}_{O(1)} + \frac{1}{m}\sum_{i=1}^m a_{ij}(v_i^{(t-1)})^2 \sum_{l\neq j} O(1/m) \tag{73}$$

$$= O(1/\sqrt{m}) \tag{74}$$

where the $O(1)$ scaling of the sum follows from Lemma 1 under Assumption 1. The final $O(1/\sqrt{m})$ scaling in (74) follows since each of the three terms in (73) is the average of $O(1/\sqrt{m})$ terms, due to $a_{ij} \in \pm 1/\sqrt{m}$. In conclusion, we have established that the elements of $\boldsymbol{r}^{(t)}$ are $O(1)$ when the elements of $\boldsymbol{v}^{(t-1)}, \boldsymbol{x}^{(t)}$ and $\boldsymbol{\mu}^{(t)}$ are.

To complete the induction proof, we need to establish that the elements of $\boldsymbol{x}^{(t+1)}$ and $\boldsymbol{\mu}^{(t+1)}$ are $O(1)$ when those of $\boldsymbol{v}^{(t)}$ and $\boldsymbol{r}^{(t)}$ are. But this follows straightforwardly from (2) and (4), i.e.,

$$x_j^{(t+1)} = \eta^{(t)}(r_j^{(t)}) \tag{75}$$

$$\mu_i^{(t+1)} = \frac{n}{m} v_i^{(t)} \frac{1}{n}\sum_{j=1}^n \eta^{(t)\prime}(r_j^{(t)}), \tag{76}$$

since n/m is $O(1)$ in the large-system limit.

In summary, we have established Lemma 2, which says the elements of $\boldsymbol{v}^{(t)}, \boldsymbol{r}^{(t)}, \boldsymbol{x}^{(t)}$, and $\boldsymbol{\mu}^{(t)}$ scale as $O(1)$ in the large-system limit under Assumption 1.

ACKNOWLEDGMENT

The author thanks Galen Reeves for inspiring discussions and feedback on early drafts of this manuscript.

REFERENCES

[1] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, pp. 18914–18919, Nov. 2009.

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

[3] 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, Feb. 2011.

[4] G. Reeves and H. D. Pfister, "The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact," in *Proc. IEEE Int. Symp. Inf. Theory*, 2016, pp. 665–669.

[5] J. Barbier, M. Dia, N. Macris, and F. Krzakala, "The mutual information in random linear estimation," in *Proc. Allerton Conf. Commun., Control, Comput.*, 2016, pp. 625–632.

[6] C. Rush and R. Venkataramanan, "Finite-sample analysis of approximate message passing algorithms," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7264–7286, Nov. 2018.

[7] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.

[8] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, pp. 1–305, May 2008.

[9] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. Inf. Theory Workshop*, Jan. 2010, pp. 1–5.
[10] A. Montanari, "Graphical models concepts in compressed sensing," in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok, eds., Cambridge, U.K.: Cambridge Univ. Press, 2012.
[11] T. Minka, *A Family of Approx Algorithms for Bayesian Inference*. Ph.D. thesis, Dept. Comp. Sci. Eng., MIT, Cambridge, MA, Jan. 2001.
[12] T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter, "Approximate inference techniques with expectation constraints," *J. Statistical Mech.: Theory Experiment*, vol. P11015, 2005.
[13] 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, Aug. 2015.
[14] A. Chambolle, R. A. DeVore, N. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 319–335, Mar. 1998.

Philip Schniter (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Illinois at Urbana-Champaign, USA, in 1992 and 1993, respectively, and the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, USA, in 2000. He is currently a Professor in the Department of Electrical and Computer Engineering at The Ohio State University, Columbus, OH, USA. He received the IEEE Signal Processing Society Best Paper Award in 2016, and has served on several IEEE Technical Committees: Signal Processing for Communications and Networking, Sensor Array and Multichannel, and Computational Imaging.