A SIMPLE DERIVATION OF AMP AND ITS STATE EVOLUTION VIA FIRST-ORDER CANCELLATION

Philip Schniter

Dept. of ECE, The Ohio State University, Columbus, OH, 43210. (schniter.1@osu.edu)

ABSTRACT

We consider the linear regression problem, where the goal is to recover the vector $\boldsymbol{x} \in \mathbb{R}^n$ from measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} \in \mathbb{R}^m$ under known matrix $\boldsymbol{A}$ and unknown noise $\boldsymbol{w}$. For large i.i.d. sub-Gaussian $\boldsymbol{A}$, 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. $\boldsymbol{A}$ 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— Approximate message passing, belief propagation, linear regression, compressive sensing, state evolution

1. 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 [3, 2], 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), as outlined in [1] and [7], the LBP perspective is lacking in several respects. First, LBP is generally suboptimal, making it surprising that a simplified approximation of LBP can be optimal. Second, the LBP derivation provides little insight into why large i.i.d. $\boldsymbol{A}$ matrices are important for AMP. Third, the LBP derivation does not suggest a scalar state evolution.

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

2. PROBLEM SETUP

For the linear regression problem (1), we treat $\boldsymbol{y} = [y_1, \dots, y_m]^\top$, $\boldsymbol{x} = [x_1, \dots, x_n]^\top$, and $\boldsymbol{w} = [w_1, \dots, w_m]^\top$ as deterministic vectors and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ as a deterministic matrix. But 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 LSL.

Definition 1. *The "large system limit" (LSL) 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 LSL.

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

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

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

where $\eta^{(t)}(\cdot)$ is component-wise separable (i.e., $[\eta^{(t)}(\boldsymbol{r})]_j = \eta^{(t)}(r_j)\ \forall j$) and $\boldsymbol{\mu}^{(t)}$ is a correction term. 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 [8] uses no correction, i.e., $\boldsymbol{\mu}^{(t)} = \boldsymbol{0}$ for all iterations t, whereas the AMP algorithm [1] uses the "Onsager" correction

$$\boldsymbol{\mu}^{(t)} = \frac{1}{m}\boldsymbol{v}^{(t-1)} \sum_{j=1}^{n} \eta^{(t-1)\prime}(r_j^{(t-1)}), \tag{3}$$

initialized with $\boldsymbol{\mu}^{(0)} = \boldsymbol{0}$. In (3), $\eta^{(t)\prime}$ refers to the derivative of $\eta^{(t)}$. Our goal is to analyze the effect of $\boldsymbol{\mu}^{(t)}$ on algorithm (2) in the LSL and in particular to understand why the Onsager correction (3) is a good choice. To do this, we analyze the errors on $\boldsymbol{r}^{(t)}$ and $\boldsymbol{x}^{(t)}$ in (2) and drop terms that vanish in the LSL.

Supported in part by the National Science Foundation under Grant CCF-1716388.

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

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

3. AMP DERIVATION

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

$$\boldsymbol{e}^{(t)} \triangleq \boldsymbol{r}^{(t)} - \boldsymbol{x}. \tag{4}$$

From (2) and (4) we have that

$$\begin{aligned}\boldsymbol{e}^{(t)} &= \boldsymbol{x}^{(t)} + \boldsymbol{A}^\top(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{(t)} + \boldsymbol{\mu}^{(t)}) - \boldsymbol{x} &(5)\\ &= (\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}^{(t)} + \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} + \boldsymbol{\mu}^{(t)}) - \boldsymbol{x} &(6)\\ &= (\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}^{(t)} - (\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x} + \boldsymbol{A}^\top(\boldsymbol{w} + \boldsymbol{\mu}^{(t)}). &(7)\end{aligned}$$

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

$$\begin{aligned}&[(\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}^{(t)}]_j = x_j^{(t)} - \sum_i a_{ij}\sum_l a_{il}x_l^{(t)}\\ &= \Big(1 - \sum_{i=1}^m a_{ij}^2\Big)x_j^{(t)} - \sum_i a_{ij}\sum_{l\neq j} a_{il}x_l^{(t)} &(8)\\ &= -\sum_i a_{ij}\sum_{l\neq j} a_{il}x_l^{(t)} &(9)\end{aligned}$$

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

$$\begin{aligned}&[(\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}^{(t)}]_j\\ &= -\sum_i a_{ij}\sum_{l\neq j} a_{il}\eta^{(t-1)}(r_l^{(t-1)}) &(10)\\ &= -\sum_i a_{ij}\sum_{l\neq j} a_{il}\eta^{(t-1)}\Big(\underbrace{x_l^{(t-1)} + \sum_{k\neq i} a_{kl}v_k^{(t-1)}}_{\triangleq\, r_{il}^{(t-1)}} + a_{il}v_i^{(t-1)}\Big), &(11)\end{aligned}$$

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 one of the main reasons 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 $\{w_k\}_{k=1}^m$.*

Assumption 1 will often be used 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 LSL, 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, $\mathbb{E}[Z_i^2] = \mathbb{E}[(\sum_{j=1}^n A_{ij}u_j)^2] = \sum_{j=1}^n\sum_{l=1}^n \mathbb{E}[A_{ij}A_{il}]u_ju_l = \frac{1}{m}\sum_{j=1}^n u_j^2 = \frac{n}{m}\frac{1}{n}\sum_{j=1}^n u_j^2$, 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 LSL. 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 LSL. □

Later we will make use of the following lemma, whose proof is omitted 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 (3), the elements of $\boldsymbol{v}^{(t)}$, $\boldsymbol{r}^{(t)}$, $\boldsymbol{x}^{(t)}$, and $\boldsymbol{\mu}^{(t)}$ scale as $O(1)$ in the LSL for all iterations t.*

Proof. See [9, App. A]. □

We now perform a Taylor series expansion of the $\eta^{(t-1)}$ term in (11) about $r_{il}^{(t-1)}$:

$$\begin{aligned}&\eta^{(t-1)}(r_{il}^{(t-1)} + a_{il}v_i^{(t-1)}) = \eta^{(t-1)}(r_{il}^{(t-1)}) &(12)\\ &+ a_{il}v_i^{(t-1)}\eta^{(t-1)\prime}(r_{il}^{(t-1)}) + \underbrace{\frac{1}{2}a_{il}^2(v_i^{(t-1)})^2\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)}) + \text{H.O.T.}}_{O(1/m)}\end{aligned}$$

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 2, 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 $O(1)$. Similarly, the 2nd term in (12) scales as $O(1/\sqrt{m})$. We will ignore the $O(1/m)$ term in (12) since it vanishes relative to the $O(1/\sqrt{m})$ term in the LSL. Thus we have

$$\begin{aligned}&[(\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}^{(t)}]_j\\ &\approx -\sum_i a_{ij}\sum_{l\neq j} a_{il}\big[\eta^{(t-1)}(r_{il}^{(t-1)}) + a_{il}v_i^{(t-1)}\eta^{(t-1)\prime}(r_{il}^{(t-1)})\big] &(13)\\ &= -\sum_i a_{ij}\sum_{l\neq j} a_{il}\eta^{(t-1)}(r_{il}^{(t-1)}) - \frac{1}{m}\sum_i a_{ij}v_i^{(t-1)}\sum_{l\neq j}\eta^{(t-1)\prime}(r_{il}^{(t-1)}) &(14)\end{aligned}$$

using $a_{il}^2 = 1/m\ \forall il$. Similar to (9), we have

$$[(\boldsymbol{I} - \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{x}]_j = -\sum_i a_{ij}\sum_{l\neq j} a_{il}x_l, \tag{15}$$

which, combined with (7) and (14), yields

$$\begin{aligned}e_j^{(t)} &\approx \sum_i a_{ij}\sum_{l\neq j} a_{il}\big[x_l - \eta^{(t-1)}(r_{il}^{(t-1)})\big] &(16)\\ &\quad - \frac{1}{m}\sum_i a_{ij}v_i^{(t-1)}\sum_{l\neq j}\eta^{(t-1)\prime}(r_{il}^{(t-1)}) + \sum_i a_{ij}(w_i + \mu_i^{(t)})\\ &= \sum_i a_{ij}\sum_{l\neq j} a_{il}\big[x_l - \eta^{(t-1)}(r_{il}^{(t-1)})\big] &(17)\\ &\quad + \sum_i a_{ij}w_i + \sum_i a_{ij}\Big[\mu_i^{(t)} - v_i^{(t-1)}\frac{1}{m}\sum_{l\neq j}\eta^{(t-1)\prime}(r_{il}^{(t-1)})\Big].\end{aligned}$$

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 (17) behave like realizations of zero-mean Gaussians in the LSL, 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 (17), 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 choices of the correction term $\mu_i^{(t)}$.

With AMP's choice of $\mu_i^{(t)}$, however, the 3rd term in (17) vanishes in the LSL. In particular, with $\mu_i^{(t)}$ from (3), it becomes

$$\begin{aligned}&\sum_i a_{ij}\Big[\frac{v_i^{(t-1)}}{m}\sum_l \eta^{(t-1)\prime}(r_l^{(t-1)}) - \frac{v_i^{(t-1)}}{m}\sum_{l\neq j}\eta^{(t-1)\prime}(r_{il}^{(t-1)})\Big]\\ &= \frac{1}{m}\sum_i a_{ij}v_i^{(t-1)}\Big[\eta^{(t-1)\prime}(r_j^{(t-1)}) + \sum_{l\neq j}\Big(\eta^{(t-1)\prime}(r_l^{(t-1)}) - \eta^{(t-1)\prime}(r_{il}^{(t-1)})\Big)\Big]\\ &\approx \frac{1}{m}\sum_i a_{ij}v_i^{(t-1)}\Big[\eta^{(t-1)\prime}(r_j^{(t-1)}) + \sum_{l\neq j} a_{il}v_i^{(t-1)}\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)})\Big], &(18)\end{aligned}$$

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

$$\begin{aligned}&\eta^{(t-1)\prime}(r_l^{(t-1)}) = \eta^{(t-1)\prime}(r_{il}^{(t-1)} + a_{il}v_i^{(t-1)}) &(19)\\ &= \eta^{(t-1)\prime}(r_{il}^{(t-1)}) + a_{il}v_i^{(t-1)}\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)}) + O(1/m) &(20)\end{aligned}$$

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 LSL. In (18), the first term is

$$\frac{1}{m}\sum_{i=1}^{m}\underbrace{a_{ij}v_i^{(t-1)}\eta^{(t-1)\prime}(r_j^{(t-1)})}_{O(1/\sqrt{m})} = O(1/\sqrt{m}) \tag{21}$$

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 2. Thus the first term in (18) will vanish in the LSL. The second term in (18) is

$$\underbrace{\frac{1}{m}\sum_{i=1}^{m}a_{ij}(v_i^{(t-1)})^2\underbrace{\sum_{l\neq j}a_{il}\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)})}_{O(1)}}_{O(1/\sqrt{m})} = O(1/\sqrt{m}), \tag{22}$$

which will also vanish in the LSL. The $O(1)$ scaling in (22) follows from Lemma 2 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 $\mu_i^{(t)}$, eq. (17) becomes

$$e_j^{(t)} \approx \sum_i a_{ij}\sum_{l\neq j}a_{il}\big[\underbrace{x_l - \eta^{(t-1)}(r_{il}^{(t-1)})}_{\triangleq\,\epsilon_{il}^{(t)}}\big] + \sum_i a_{ij}w_i. \tag{23}$$

Under Assumption 1, a_{il} is a realization of equiprobable $A_{il} \in \pm\frac{1}{\sqrt{m}}$ that is independent of $\{x_l\}_{l=1}^N$, $r_{il}^{(t-1)}$, and $\{A_{ij}\}_{j\neq l}$. 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

$$\mathbb{E}\Big[\sum_i A_{ij}\sum_{l\neq j}A_{il}\epsilon_{il}^{(t)}\Big] = \sum_i \mathbb{E}[A_{ij}]\sum_{l\neq j}\mathbb{E}[A_{il}]\epsilon_{il}^{(t)} = 0 \tag{24}$$

$$\mathbb{E}\Big[\Big(\sum_i A_{ij}\sum_{l\neq j}A_{il}\epsilon_{il}^{(t)}\Big)^2\Big] = \sum_i \mathbb{E}[A_{ij}^2]\sum_{l\neq j}\mathbb{E}[A_{il}^2](\epsilon_{il}^{(t)})^2 \tag{25}$$

$$= \frac{1}{m^2}\sum_i\sum_{l\neq j}(\epsilon_{il}^{(t)})^2. \tag{26}$$

From the Taylor expansion (12), we have

$$\epsilon_{il}^{(t)} = x_l - \eta^{(t-1)}(r_{il}^{(t-1)}) \tag{27}$$

$$= \underbrace{x_l - \eta^{(t-1)}(r_l^{(t-1)})}_{\triangleq\,\epsilon_l^{(t)}} + \underbrace{a_{il}v_i^{(t-1)}\eta^{(t-1)\prime}(r_{il}^{(t-1)}) + O(1/m)}_{O(1/\sqrt{m})}, \tag{28}$$

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, which is also $O(1)$. Because the $O(1/\sqrt{m})$ term in (28) vanishes in the LSL, we see that (26) becomes

$$\frac{1}{m^2}\sum_i\sum_{l\neq j}(\epsilon_{il}^{(t)})^2 \approx \frac{1}{m^2}\sum_{i=1}^m\sum_{l\neq j}(\epsilon_l^{(t)})^2 = \frac{1}{m}\sum_{l\neq j}(\epsilon_l^{(t)})^2 \tag{29}$$

$$= \underbrace{\frac{n}{m}\frac{1}{n}\sum_{l=1}^n(\epsilon_l^{(t)})^2}_{O(1)} - \underbrace{\frac{1}{m}(\epsilon_j^{(t)})^2}_{O(1/m)} \approx \delta^{-1}\mathcal{E}^{(t)}, \tag{30}$$

where $\mathcal{E}^{(t)}$ is the average squared error on the denoiser output $\boldsymbol{x}^{(t)}$:

$$\mathcal{E}^{(t)} \triangleq \lim_{n\to\infty}\frac{1}{n}\sum_{l=1}^n(\epsilon_l^{(t)})^2. \tag{31}$$

We have thus deduced that, in the LSL, the first term in (23) behaves like a zero-mean Gaussian with variance $\delta^{-1}\mathcal{E}^{(t)}$. For the second term in (23), 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

$$\mathbb{E}\Big[\sum_i A_{ij}w_i\Big] = \sum_i \mathbb{E}[A_{ij}]w_i = 0 \tag{32}$$

$$\mathbb{E}\Big[\Big(\sum_i A_{ij}w_i\Big)^2\Big] = \sum_i \mathbb{E}[A_{ij}^2]w_i^2 = \frac{1}{m}\sum_{i=1}^m w_i^2 \approx \tau_w, \tag{33}$$

where τ_w denotes the empirical second moment of the noise:

$$\tau_w \triangleq \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^m w_i^2. \tag{34}$$

To summarize, with AMP's choice of $\boldsymbol{\mu}^{(t)}$ from (3), the jth component of the denoiser input-error behaves like

$$e_j^{(t)} \sim \mathcal{N}\big(0, \underbrace{\delta^{-1}\mathcal{E}^{(t)} + \tau_w}_{\triangleq\,\tau_r^{(t)}}\big) \tag{35}$$

in the LSL, 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., ISTA'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.

4. AMP STATE EVOLUTION

In Section 3, 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 LSL.

Recalling the definition of $\mathcal{E}^{(t)}$ from (31), we can write

$$\frac{1}{n}\sum_{l=1}^n(\epsilon_l^{(t)})^2 \approx \frac{1}{n}\sum_{l=1}^n\big[\eta^{(t-1)}(x_l + \mathcal{N}(0, \tau_r^{(t-1)})) - x_l\big]^2 \tag{36}$$

$$= \mathbb{E}\big[\eta^{(t-1)}(X + \mathcal{N}(0, \tau_r^{(t-1)})) - X\big]^2 \tag{37}$$

where X is a scalar random variable with the empirical distribution

$$X \sim p(x) = \frac{1}{n}\sum_{l=1}^n\delta(x - x_l), \tag{38}$$

where $\delta(\cdot)$ denotes the Dirac delta. Thus, in the LSL, we can argue

$$\mathcal{E}^{(t)} = \mathbb{E}\big[\eta^{(t-1)}\big(X + \mathcal{N}(0, \tau_r^{(t-1)})\big) - X\big]^2, \tag{39}$$

where X now is distributed according to the $n \to \infty$ limit of the empirical distribution. Combining (39) with the update equation for $\tau_r^{(t)}$ gives the following recursion for $t = 0, 1, 2, \ldots$:

$$\tau_r^{(t)} = \delta^{-1}\mathcal{E}^{(t)} + \tau_w \tag{40a}$$

$$\mathcal{E}^{(t+1)} = \mathbb{E}\big[\eta^{(t)}\big(X + \mathcal{N}(0, \tau_r^{(t)})\big) - X\big]^2, \tag{40b}$$

initialized with $\mathcal{E}^{(0)} = \mathbb{E}[X^2]$. The recursion (40) is known as AMP's "state evolution" for the mean-squared error [1, 3, 2].

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.

5. 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., $\tau_r^{(t)} \approx \|\boldsymbol{v}^{(t)}\|^2/m$ [7]. We now explain this approach using arguments similar to those used above.

To begin, it is straightforward to show (see [9, eq.(44)]) that

$$v_i^{(t)} = y_i - \sum_{l=1}^{n} a_{il}\eta^{(t-1)}(r_{il}^{(t-1)}) - v_i^{(t-1)}\frac{1}{m}\sum_{l=1}^{n}\eta^{(t-1)\prime}(r_{il}^{(t-1)}) + \mu_i^{(t)} + O(1/m). \quad (41)$$

Ignoring the $O(1/m)$ term and use AMP's $\mu_i^{(t)}$ from (3), we get

$$v_i^{(t)} \approx y_i - \sum_{l=1}^{n} a_{il}\eta^{(t-1)}(r_{il}^{(t-1)}) + v_i^{(t-1)}\frac{1}{m}\sum_{l=1}^{n}\big[\eta^{(t-1)\prime}(r_l^{(t-1)}) - \eta^{(t-1)\prime}(r_{il}^{(t-1)})\big] \quad (42)$$

$$= y_i - \sum_{l=1}^{n} a_{il}\eta^{(t-1)}(r_{il}^{(t-1)}) + v_i^{(t-1)}\frac{n}{m}\frac{1}{n}\sum_{l=1}^{n}\underbrace{\big[a_{il}v_i^{(t-1)}\eta^{(t-1)\prime\prime}(r_{il}^{(t-1)}) + O(1/m)\big]}_{O(1/\sqrt{m})}, \quad (43)$$

where we used the Taylor series (20) 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 (43) 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 LSL. With this and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$, we get

$$v_i^{(t)} \approx w_i + \sum_{l=1}^{n} a_{il}\underbrace{\big[x_l - \eta^{(t-1)}(r_{il}^{(t-1)})\big]}_{=\epsilon_{il}^{(t)}}, \quad (44)$$

recalling $\epsilon_{il}^{(t)}$ defined in (23). Squaring and averaging over i yields

$$\frac{1}{m}\sum_{i=1}^{m}(v_i^{(t)})^2 \approx \frac{1}{m}\sum_{i=1}^{m}w_i^2 + \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{l=1}^{n}a_{il}\epsilon_{il}^{(t)}\right)^2 + \frac{2}{m}\sum_{i=1}^{m}\left(w_i\sum_{l=1}^{n}a_{il}\big[x_l - \eta^{(t-1)}(r_{il}^{(t-1)})\big]\right). \quad (45)$$

We now examine the components of (45) in the LSL. By definition, the first term in (45) converges to τ_w. By the law of large numbers, the second term converges to

$$\lim_{n\to\infty}\mathbb{E}\left[\left(\sum_{l=1}^{n}A_{il}\epsilon_{il}^{(t)}\right)^2\right] = \lim_{n\to\infty}\sum_{l=1}^{n}\sum_{j=1}^{n}\mathbb{E}[A_{il}A_{ij}]\epsilon_{il}^{(t)}\epsilon_{ij}^{(t)} \quad (46)$$

$$= \lim_{n\to\infty}\frac{1}{m}\sum_{l=1}^{n}(\epsilon_{il}^{(t)})^2, \quad (47)$$

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 (28), we have

$$\lim_{n\to\infty}\frac{1}{m}\sum_{l=1}^{n}(\epsilon_{il}^{(t)})^2 = \lim_{n\to\infty}\frac{n}{m}\frac{1}{n}\sum_{l=1}^{n}(\epsilon_l^{(t)})^2 = \delta^{-1}\mathcal{E}^{(t)}, \quad (48)$$

where m depends on n because $m/n = O(1)$. In summary,

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}(v_i^{(t)})^2 = \tau_w + \delta^{-1}\mathcal{E}^{(t)} = \tau_r^{(t)}, \quad (49)$$

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

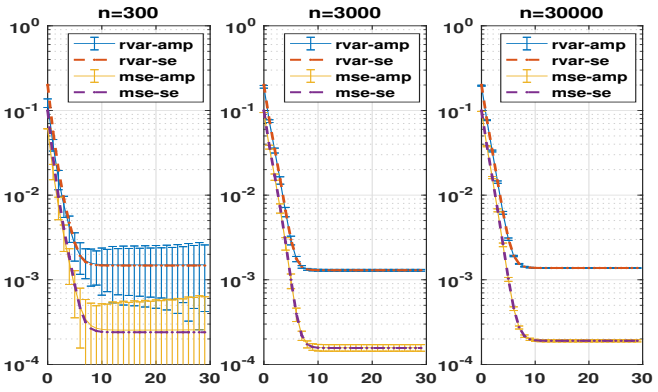

Fig. 1. Denoiser output MSE $\mathcal{E}_n^{(t)}$ and denoiser input-error variance $\tau_{r,n}^{(t)}$ versus iteration for AMP and its state evolution. Dashed lines show the empirical average over 1000 random draws of $\boldsymbol{A}$ and error bars show the empirical standard deviation.

Table 1. Numerical evidence that $\text{std}(\mathcal{E}_n^{(t)})\sqrt{n}$ and $\text{std}(\tau_{r,n}^{(t)})\sqrt{n}$ are approximately constant with n, implying that $\text{std}(\mathcal{E}_n^{(t)})$ and $\text{std}(\tau_{r,n}^{(t)})$ scale as $1/\sqrt{n}$ for sufficiently large n.

n	1000	3000	10 000	30 000
$\text{std}(\mathcal{E}_n^{(29)})\sqrt{n}$	0.0011	0.0008	0.0010	0.0010
$\text{std}(\tau_{r,n}^{(29)})\sqrt{n}$	0.0026	0.0017	0.0022	0.0020

6. 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 $\beta = 0.1$ and $\{w_i\}$ drawn i.i.d. zero-mean Gaussian with variance such that $\mathbb{E}[\|\boldsymbol{A}\boldsymbol{x}\|^2]/\mathbb{E}[\|\boldsymbol{w}\|^2] \approx 20$ dB. We used MMSE denoising: $\eta^{(t)}(r_j) = \mathbb{E}[X \,|\, r_j = X + \mathcal{N}(0, \tau_r^{(t)})]$.

Below we 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 the AMP algorithm (2) and the AMP state evolution (40). For the 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 state evolution, the denoiser output MSE was computed as

$$\mathcal{E}_n^{(t)} = \begin{cases}\mathbb{E}\left[\eta^{(t-1)}(X + \mathcal{N}(0, \tau_{r,n}^{(t-1)})) - X\right]^2 & t > 0\\ \mathbb{E}\left[X^2\right] & t = 0,\end{cases} \quad (50)$$

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 random draws of $\boldsymbol{A}$ for a *single fixed* draw of $\boldsymbol{x}$ and $\boldsymbol{w}$.

Figure 1 shows the results at dimension $n \in \{300, 3000, 30000\}$. The figures show an excellent agreement between the state evolution and average AMP quantities when $n \geq 3000$, where the average was computed over $T = 1000$ realizations of $\boldsymbol{A}$. The error bars, which show the empirical standard deviation over the T realizations, decrease as the dimension n increases. Table 1 suggests that the standard deviation scales proportional to $1/\sqrt{n}$ at large n.

7. REFERENCES

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

[2] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and message passing algorithms," *Annals of Applied 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. on Information Theory*, vol. 57, 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 Internat. Symp. on Information Theory*, 2016.

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

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

[7] A. Montanari, "Graphical models concepts in compressed sensing," in *Compressed Sensing: Theory and Applications* (Y. C. Eldar and G. Kutyniok, eds.), Cambridge Univ. Press, 2012.

[8] 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. on Image Processing*, vol. 7, pp. 319–335, Mar. 1998.

[9] P. Schniter, "A simple derivation of AMP and its state evolution via first-order cancellation," *arXiv:1907.04235*, 2019.