# An Analysis of State Evolution for Approximate Message Passing with Side Information

Hangjin Liu
NC State University
Email: hliu25@ncsu.edu

Cynthia Rush
Columbia University
Email: cynthia.rush@columbia.edu

Dror Baron
NC State University
Email: barondror@ncsu.edu

***Abstract*—A common goal in many research areas is to reconstruct an unknown signal x from noisy linear measurements. Approximate message passing (AMP) is a class of low-complexity algorithms for efficiently solving such high-dimensional regression tasks. Often, it is the case that side information (SI) is available during reconstruction. For this reason a novel algorithmic framework that incorporates SI into AMP, referred to as approximate message passing with side information (AMP-SI), has been recently introduced. An attractive feature of AMP is that when the elements of the signal are exchangeable, the entries of the measurement matrix are independent and identically distributed (i.i.d.) Gaussian, and the denoiser applies the same non-linearity at each entry, the performance of AMP can be predicted accurately by a scalar iteration referred to as state evolution (SE). However, the AMP-SI framework uses different entry-wise scalar denoisers, based on the entry-wise level of the SI, and therefore is not supported by the standard AMP theory. In this work, we provide rigorous performance guarantees for AMP-SI when the input signal and SI are drawn i.i.d. according to some joint distribution subject to finite moment constraints. Moreover, we provide numerical examples to support the theory which demonstrate empirically that the SE can predict the AMP-SI mean square error accurately.**

*A full version of this paper is accessible at:* http://www.columbia.edu/~cgr2130/AMPSI_SE.pdf

## I. Introduction

High-dimensional linear regression is a well-studied model being used in many applications including compressed sensing [1], imaging [2], and machine learning and statistics [3]. The unknown signal $\mathbf{x} \in \mathbb{R}^n$ is viewed through the linear model:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^m$ are the measurements, $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a known measurement matrix, and $\mathbf{w} \in \mathbb{R}^m$ is measurement noise. The goal is to estimate the unknown signal $\mathbf{x}$ having knowledge only of the noisy measurements $\mathbf{y}$ and the measurement matrix $\mathbf{A}$. When the problem is under-determined (i.e., $m < n$), in order for reconstruction to be successful, it is necessary to exploit structural or probabilistic characteristics of the input signal $\mathbf{x}$. Often a prior distribution on the input signal $\mathbf{x}$ is assumed, and in this case approximate message passing (AMP) algorithms [1] can be used for the reconstruction task.

AMP [1], [4] is a class of low-complexity algorithms for efficiently solving high-dimensional regression tasks (1). AMP works by iteratively generating estimates of the unknown input vector, $\mathbf{x}$, using a possibly non-linear denoiser function tailored to any prior knowledge about $\mathbf{x}$. One favorable feature of AMP is that under some technical conditions on the measurement matrix $\mathbf{A}$ and $\mathbf{x}$, the observations at each iteration of the algorithm are almost surely equal in distribution to $\mathbf{x}$ plus independent and identically distributed (i.i.d.) Gaussian noise in the large system limit.

**AMP with Side Information (AMP-SI):** In information theory [5], when different communication systems share side information (SI), overall communication can become more efficient. Recently [6], [7], a novel algorithmic framework, referred to as AMP-SI, has been introduced for incorporating SI into AMP for high-dimensional regression tasks (1). AMP-SI has been empirically demonstrated to have good reconstruction quality and is easy to use. For example, we have proposed to use AMP-SI for channel estimation in emerging millimeter wave communication systems [8], where the time dynamics of the channel structure allow previous channel estimates to be used as SI when estimating the current channel structure [7].

We model the observed SI, denoted by $\widetilde{\mathbf{x}} \in \mathbb{R}^n$, as depending statistically on the unknown signal $\mathbf{x}$ through some joint probability density function (pdf), $f(\mathbf{X}, \widetilde{\mathbf{X}})$. AMP-SI uses a conditional denoiser, $g_t : \mathbb{R}^{2n} \to \mathbb{R}^n$, to incorporate SI,

$$g_t(\mathbf{a}, \mathbf{b}) = \mathbb{E}[\mathbf{X} | \mathbf{X} + \lambda_t \mathcal{N}(0, \mathbb{I}_n) = \mathbf{a}, \widetilde{\mathbf{X}} = \mathbf{b}]. \tag{2}$$

The AMP-SI algorithm iteratively updates estimates of the input signal $\mathbf{x}$: let $\mathbf{x}^0 = \mathbf{0}$, the all-zeros vector, then

$$\mathbf{r}^t = \mathbf{y} - \mathbf{A}\mathbf{x}^t + \frac{\mathbf{r}^{t-1}}{\delta}[\operatorname{div} g_{t-1}(\mathbf{x}^{t-1} + \mathbf{A}^T\mathbf{r}^{t-1}, \widetilde{\mathbf{x}})], \tag{3}$$

$$\mathbf{x}^{t+1} = g_t(\mathbf{x}^t + \mathbf{A}^T\mathbf{r}^t, \widetilde{\mathbf{x}}), \tag{4}$$

where $\mathbf{x}^t \in \mathbb{R}^n$ is the estimate of $\mathbf{x}$ at iteration $t$ and $\delta = \frac{m}{n}$ is the measurement rate. For a differential function $g : \mathbb{R}^{2n} \to \mathbb{R}^n$ we use $\operatorname{div} g(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^n \frac{\partial g_i}{\partial a_i}(\mathbf{a}, \mathbf{b})$. Using the denoiser in (2), the AMP-SI algorithm (3)-(4) provides the minimum mean squared error (MMSE) estimate of the signal when SI $\widetilde{\mathbf{x}}$ is available [6].

**State Evolution (SE):** It has been proven that the performance of AMP, as measured, for example, by the normalized squared $\ell_2$-error $\frac{1}{n}||\mathbf{x}^t - \mathbf{x}||_2^2$ between the estimate $\mathbf{x}^t$ and true signal $\mathbf{x}$, can be accurately predicted by a scalar recursion referred as SE [9], [10] when the measurement matrix $\mathbf{A}$ is i.i.d. Gaussian under various assumptions on the elements of the signal. The SE equation for AMP-SI is as follows. Assume the entries of the noise $\mathbf{w}$ are i.i.d. $\sim f(W)$ with $\sigma_w^2 = \mathbb{E}[W^2]$,

978-1-5386-9291-2/19/$31.00 ©2019 IEEE

and let $\lambda_0 = \sigma_w^2 + \mathbb{E}[||\mathbf{X}||^2]/n\delta$. Then for $t \geq 0$,

$$\lambda_t^2 = \sigma_w^2 + \frac{1}{\delta n}\mathbb{E}\left[||g_{t-1}(\mathbf{X} + \lambda_{t-1}\mathbf{Z}, \widetilde{\mathbf{X}}) - \mathbf{X}||^2\right], \quad (5)$$

where $(\mathbf{X}, \widetilde{\mathbf{X}}) \sim f(\mathbf{X}, \widetilde{\mathbf{X}})$ are independent of $\mathbf{Z} \sim \mathcal{N}(0, \mathbb{I}_n)$, where we use $\mathcal{N}(\mu, \sigma^2)$ to denote a Gaussian distribution with mean $\mu$ and variance $\sigma^2$.

Considering AMP-SI (3)-(4), however, we cannot directly apply the existing AMP theoretical results [9], [10], as the conditional denoiser (2) depends on the index $i$ through the SI, meaning that different scalar denoisers will be used at different indices within the AMP-SI iterations. Recent results [11], however, extend the asymptotic SE analysis to a larger class of possible denoisers, allowing, for example, each element of the input to use a different non-linear denoiser as is the case in AMP-SI. We employ these results to rigorously relate the SE presented in (5) to the AMP-SI algorithm in (3)-(4).

**Related Work:** While integrating SI into reconstruction algorithms is not new, AMP-SI introduces a unified framework within AMP supporting arbitrary signal and SI dependencies. Prior work using SI has been either heuristic, limited to specific applications, or outside the AMP framework.

For example, Wang and Liang [12] integrate SI into AMP for a specific signal prior density, but the method is difficult to apply to other signal models. Ziniel and Schniter [13] develop an AMP-based reconstruction algorithm for a time-varying signal model based on Markov processes for the support and amplitude. This signal model is easily incorporated into the AMP-SI framework as discussed in the analysis of the birth-death-drift model of [6], [7]. Manoel et al. implement an AMP-based algorithm in which the input signal is repeatedly reconstructed in a streaming fashion, and information from past reconstruction attempts is aggregated into a prior, thus improving ongoing reconstruction results [14]. This reconstruction scheme resembles that of AMP-SI, in particular when the Bernoulli-Gaussian model is used (see Section II-B).

**Contribution and Outline:** Ma et al. use numerical experiments to show that SE (5) accurately tracks the performance of AMP-SI (3)-(4) [7], as was shown rigorously for standard AMP. Ma et al. conjecture that rigorous theoretical guarantees can be given for AMP-SI as well [7]. In this work, we analyze AMP-SI performance when the input signal and SI are drawn i.i.d. according to a general pdf $f(\mathbf{X}, \widetilde{\mathbf{X}})$ obeying some finite moment conditions, the AMP-SI denoiser (2) is Lipschitz, and the measurement matrix $\mathbf{A}$ is i.i.d. Gaussian.

In Section II, we give the main results, examples for various signal and SI models, and numerical experiments comparing the empirical performance of AMP-SI and the SE predictions. The proof of our main theorem is provided in Section III.

## II. Main Results

### A. Main Theorem

Our main result provides AMP-SI performance guarantees when considering *pseudo-Lipschitz* loss functions.

**Definition II.1.** ***Pseudo-Lipschitz functions*** [11]: For $k \in \mathbb{N}_{>0}$ and any $n \in \mathbb{N}_{>0}$, a function $\phi : \mathbb{R}^n \to \mathbb{R}$ is *pseudo-Lipschitz of order* $k$, or *PL(k)*, if there exists a constant $L$, referred to as the pseudo-Lipschitz constant of $\phi$, such that for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq L\Big(1 + \Big(\frac{||\mathbf{x}||}{\sqrt{n}}\Big)^{k-1} + \Big(\frac{||\mathbf{y}||}{\sqrt{n}}\Big)^{k-1}\Big)\frac{||\mathbf{x} - \mathbf{y}||}{\sqrt{n}}.$$

A sequence (in $n$) of PL(k) functions $\{\phi_n\}_{n \in \mathbb{N}_{>0}}$ is called *uniformly pseudo-Lipschitz* of order $k$ , or *uniformly PL(k)*, if, denoting by $L_n$ the pseudo-Lipschitz constant of $\phi_n$, we have $L_n < \infty$ for each $n$ and $\limsup_{n\to\infty} L_n < \infty$.

Throughout the work, $||\cdot||$ denotes the Euclidean norm, and $\overset{p}{=}$ denotes convergence in probability. In the case of $(\mathbf{X}, \widetilde{\mathbf{X}})$ sampled i.i.d. $f(X, \widetilde{X})$ the AMP-SI denoiser (originally defined in (2)) is separable: define $\eta_t : \mathbb{R}^2 \to \mathbb{R}$, as

$$\eta_t(a, b) = \mathbb{E}[X | X + \lambda_t \mathcal{N}(0, 1) = a, \widetilde{X} = b], \quad (6)$$

and the AMP algorithm in (3)-(4) simplifies to

$$\mathbf{r}^t = \mathbf{y} - \mathbf{A}\mathbf{x}^t + \frac{\mathbf{r}^{t-1}}{\delta}\sum_{i=1}^{n}\eta'_{t-1}([\mathbf{x}^{t-1} + \mathbf{A}^T\mathbf{r}^{t-1}]_i, \widetilde{x}_i), \quad (7)$$

$$x_i^{t+1} = \eta_t([\mathbf{x}^t + \mathbf{A}^T\mathbf{r}^t]_i, \widetilde{x}_i), \qquad \text{for } i = 1, 2, \ldots, n, \quad (8)$$

where the derivative $\eta'_t(s, \cdot) = \frac{\partial}{\partial s}\eta_t(s, \cdot)$. For the denoiser in (6), the SE is as follows: let $\lambda_0 = \sigma_w^2 + \mathbb{E}[X^2]/\delta$ and for $t \geq 0$,

$$\lambda_t^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}\left[(\eta_{t-1}(X + \lambda_{t-1}Z, \widetilde{X}) - X)^2\right], \quad (9)$$

where $(X, \widetilde{X}) \sim f(X, \widetilde{X})$ are independent of $Z \sim \mathcal{N}(0, 1)$.

**Theorem II.1.** *For any PL(2) functions $\phi : \mathbb{R}^2 \to \mathbb{R}$ and $\psi : \mathbb{R}^3 \to \mathbb{R}$, define sequences of functions $\phi_m : \mathbb{R}^{2m} \to \mathbb{R}$ and $\psi_n : \mathbb{R}^{3n} \to \mathbb{R}$ as follows: for vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$ and $\mathbf{x}, \mathbf{y}, \widetilde{\mathbf{x}} \in \mathbb{R}^n$,*

$$\begin{aligned} \phi_m(\mathbf{a}, \mathbf{b}) &:= \frac{1}{m}\sum_{i=1}^{m}\phi(a_i, b_i), \\ \psi_n(\mathbf{x}, \mathbf{y}, \widetilde{\mathbf{x}}) &:= \frac{1}{n}\sum_{i=1}^{n}\psi(x_i, y_i, \widetilde{x}_i). \end{aligned} \quad (10)$$

*Then the functions in* (10) *are uniformly PL(2). Next, assume the following:*

- **(A1)** *The measurement matrix $\mathbf{A}$ has i.i.d. Gaussian entries with mean $0$ and variance $1/m$.*
- **(A2)** *The noise $\mathbf{w}$ is i.i.d. $\sim f(W)$ with finite $\mathbb{E}[|W|^2]$.*
- **(A3)** *The signal and SI $(\mathbf{x}, \widetilde{\mathbf{x}})$ are sampled i.i.d. from $f(X, \widetilde{X})$ with finite $\mathbb{E}[|X|^2]$, finite $\mathbb{E}[|\widetilde{X}|^2]$, and finite $\mathbb{E}[|X\widetilde{X}|]$.*
- **(A4)** *For $t \geq 0$, the denoisers $\eta_t(\cdot, \cdot)$ defined in* (6) *are Lipschitz continuous: for scalars $a_1, a_2, b_1, b_2$, and constant $L > 0$, $|\eta_t(a_1, b_1) - \eta_t(a_2, b_2)| \leq L||(a_1, b_1) - (a_2, b_2)||$.*

*Then, we have the following asymptotic results for the functions defined in* (10),

$$\begin{aligned} \lim_m \phi_m(\mathbf{r}^t, \mathbf{w}) &\overset{p}{=} \lim_m \mathbb{E}[\phi_m(\mathbf{W} + \sqrt{\lambda_t^2 - \sigma_w^2}\,\mathbf{Z}_1, \mathbf{W})]), \\ \lim_n \psi_n(\mathbf{x}^t + \mathbf{A}^T\mathbf{r}^t, \mathbf{x}, \widetilde{\mathbf{x}}) &\overset{p}{=} \lim_n \mathbb{E}[\psi_n(\mathbf{X} + \lambda_t\mathbf{Z}_2, \mathbf{X}, \widetilde{\mathbf{X}})], \end{aligned} \quad (11)$$

*where $\mathbf{Z}_1 \sim \mathcal{N}(0, \mathbb{I}_m)$, $\mathbf{Z}_2 \sim \mathcal{N}(0, \mathbb{I}_n)$, independent of $\mathbf{W} \sim$ i.i.d. $f(W)$ and $(\mathbf{X}, \widetilde{\mathbf{X}}) \sim$ i.i.d. $f(X, \widetilde{X})$. $\mathbf{x}^t$ and $\mathbf{r}^t$ are* *defined in the AMP-SI recursion* (7)-(8)*, and* $\lambda_t$ *in the SE* (9)*.*

Section III contains the proof of Theorem II.1. The proof follows from Berthier et al. [11, Theorem 14] and the strong law of large numbers. The main details involve showing that assumptions **(A1)** $-$ **(A4)** allow us to apply [11, Theorem 14].

As a concrete example of how Theorem II.1 provides performance guarantees for AMP-SI, let us consider a few interesting pseudo-Lipschitz loss functions.

**Corollary II.1.1.** *Under assumptions* ***(A1)*** $-$ ***(A4)****, letting* $\psi^1 : \mathbb{R}^3 \to \mathbb{R}$ *be* $\psi^1(x,y,z) = (x-y)^2$*, then by Theorem II.1,*

$$\lim_{n\to\infty} \frac{1}{n}||\mathbf{x}^t + \mathbf{A}^T\mathbf{r}^t - \mathbf{x}||^2 \overset{p}{=} \lambda_t^2,$$

*where* $\lambda_t^2$ *is defined in* (5)*. Similarly if* $\psi^2 : \mathbb{R}^3 \to \mathbb{R}$ *is defined as* $\psi^2(x,y,z) = (\eta_t(x,z) - y)^2$*, then by Theorem II.1*

$$\lim_{n\to\infty} \frac{1}{n}||\mathbf{x}^{t+1} - \mathbf{x}||^2 \overset{p}{=} \delta(\lambda_{t+1}^2 - \sigma_w^2).$$

*When* $\eta_t$ *is Lipschitz, it is straightforward to show that* $\psi^1$ *and* $\psi^2$ *are both PL(2), and thus Theorem II.1 can be applied.*

### B. Examples

Next, we consider a few signal and SI models to show how one can derive the denoiser in (2), use this to construct the AMP-SI algorithm and the SE, and apply Theorem II.1. Before we get to the examples we state a lemma that demonstrates how functions with bounded derivative are Lipschitz.

**Lemma II.1.1.** *A function* $\phi : \mathbb{R}^2 \to \mathbb{R}$ *having bounded derivatives,* $|\frac{\partial}{\partial x}\phi(x,y)| \leq \mathsf{D}_1$ *and* $|\frac{\partial}{\partial y}\phi(x,y)| \leq \mathsf{D}_2$ *where* $0 < \mathsf{D}_1, \mathsf{D}_2 < \infty$*, is Lipschitz continuous with Lipschitz constant* $\sqrt{\mathsf{D}_1^2 + \mathsf{D}_2^2}$*.*

*1) Gaussian Signal and SI:* In this model, referred to as the GG model henceforth, the signal has i.i.d. Gaussian entries with zero mean and finite variance and we have access to SI in the form of the signal with additive white Gaussian noise (AWGN). The signal, $\mathbf{X}$, and SI, $\widetilde{\mathbf{X}}$, are related by

$$\widetilde{\mathbf{X}} = \mathbf{X} + \mathcal{N}(0, \sigma^2\mathbb{I}). \tag{12}$$

In this case, as shown in [7], because the random variables being estimated are Gaussian, the AMP-SI denoiser (2) has a linear form:

$$\eta_t(a,b) = \frac{\sigma_x^2\sigma^2 a + \sigma_x^2\lambda_t^2 b}{\sigma_x^2(\sigma^2 + \lambda_t^2) + \sigma^2\lambda_t^2}. \tag{13}$$

Then the SE (5) can be computed as

$$\lambda_t^2 = \sigma_w^2 + \frac{1}{\delta}\left[\frac{\sigma_x^2\sigma^2\lambda_{t-1}^2}{\sigma_x^2(\sigma^2 + \lambda_{t-1}^2) + \sigma^2\lambda_{t-1}^2}\right]. \tag{14}$$

We note that the denoiser in (13) is Lipschitz continuous as a result of Lemma II.1.1, because $|\frac{\partial}{\partial a}\eta_t(a,b)| \leq 1$, and $|\frac{\partial}{\partial b}\eta_t(a,b)| \leq 1$. Therefore the assumptions **(A1)** $-$ **(A4)** are satisfied in the GG case and we can apply Thoerem II.1.

*2) Bernoulli-Gaussian Signal and SI:* The Bernoulli-Gaussian (BG) model reflects a scenario in which one wishes to recover a sparse signal and has access to SI in the form of the signal with AWGN as in (12). In this model, each entry of the signal is independently generated according to $x_i \sim \epsilon\mathcal{N}(0,1) + (1-\epsilon)\delta_0$, where $\delta_0$ is the Dirac delta function at 0. In words, the entries of the signal independently take the value 0 with probability $1-\epsilon$ and are $\mathcal{N}(0,1)$ with probability $\epsilon$. In this case, as shown in [7], the AMP-SI denoiser (2) equals

$$\begin{aligned} \eta_t(a,b) &= \Pr(X \neq 0 | a, b)\, \mathbb{E}[X | a, b, X \neq 0] \\ &= (1 + T_{a,b})^{-1} f_{a,b}, \end{aligned} \tag{15}$$

where, letting $\rho_{\tau^2}(x)$ be the zero-mean Gaussian density with variance $\tau^2$ evaluated at $\mathbf{x}$,

$$f_{a,b} := \frac{\sigma^2 a + \lambda_t^2 b}{\sigma^2 + \lambda_t^2 + \sigma^2\lambda_t^2}, \tag{16}$$

$$T_{a,b} := \Big(\frac{1-\epsilon}{\epsilon}\Big)\frac{\nu_t\sqrt{2\pi}}{\lambda_t^2\sigma^2}\rho_{\nu_t}(\sigma^2 a + \lambda_t^2 b), \tag{17}$$

$$\nu_t := \sigma^2\lambda_t^2(\sigma^2 + \lambda_t^2 + \sigma^2\lambda_t^2).$$

Then the SE (5) can be computed as

$$\lambda_t^2 = \sigma_w^2 + \frac{1}{\delta}\Big(\frac{T_{a,b}}{1 + T_{a,b}}\Big)^2\left[\frac{(\sigma^2 + \lambda_{t-1}^2) + \sigma^2\lambda_{t-1}^2}{\sigma^2 + \lambda_{t-1}^2 + \sigma^2\lambda_{t-1}^2}\right]. \tag{18}$$

We again use Lemma II.1.1 to show that the denoiser defined in (15)-(17) is Lipschitz continuous so that the assumptions **(A1)** $-$ **(A4)** are satisfied in the BG case and we can apply Thoerem II.1. We study the partial derivatives. Combining (15)-(17), $\eta_t(a,b) = (1 + T_{a,b})^{-1} f_{a,b}$. Then,

$$\begin{aligned} \Big|\frac{\partial\eta_t(a,b)}{\partial a}\Big| &= \Big|\frac{1}{1 + T_{a,b}}\Big[\frac{\partial f_{a,b}}{\partial a}\Big] - \frac{1}{(1 + T_{a,b})^2}\Big[\frac{\partial T_{a,b}}{\partial a}\Big] f_{a,b}\Big| \\ &\leq \frac{(1 + 2T_{a,b})}{(1 + T_{a,b})^2}\Big|\frac{\partial f_{a,b}}{\partial a}\Big| + \frac{1}{(1 + T_{a,b})^2}\Big|\frac{\partial(T_{a,b}f_{a,b})}{\partial a}\Big|. \end{aligned} \tag{19}$$

Now we show upper bounds for the two terms of (19) separately. For the first term, $\frac{\partial f_{a,b}}{\partial a} \leq 1$, so $\frac{(1+2T_{a,b})}{(1+T_{a,b})^2}\Big|\frac{\partial f_{a,b}}{\partial a}\Big| \leq 1$.

Consider the second term of (19). From (17) and (16),

$$T_{a,b}f_{a,b} = \Big(\frac{1-\epsilon}{\epsilon}\Big)\sqrt{2\pi}(\sigma^2 a + \lambda_t^2 b)\rho_{\nu_t}(\sigma^2 a + \lambda_t^2 b),$$

then using $\frac{\partial}{\partial x}\rho_{\tau^2}(x) = -\frac{x}{\tau^2}\rho_{\tau^2}(x)$, we have

$$\Big|\frac{\partial}{\partial a}\Big[T_{a,b}f_{a,b}\Big]\Big| = \Big[\frac{1-\epsilon}{\epsilon}\Big]\frac{\sqrt{2\pi}\sigma^2}{\nu_t}\rho_{\nu_t}(\sigma^2 a + \lambda_t^2 b)\Big|\nu_t - (\sigma^2 a + \lambda_t^2 b)^2\Big|.$$

To upper bound the above, we use $\exp\{-x\} \leq \frac{1}{1+x}$ when $x \geq 0$, and so $\rho_{\tau^2}(x) \leq \sqrt{2\tau^2}(2\tau^2 + x^2)^{-1}/\sqrt{\pi}$. Thus,

$$\begin{aligned} \Big|\frac{\partial}{\partial a}\Big[T_{a,b}f_{a,b}\Big]\Big| &\leq \frac{2\sigma^2}{\sqrt{\nu_t}}\Big(\frac{1-\epsilon}{\epsilon}\Big)\frac{|\nu_t - (\sigma^2 a + \lambda_t^2 b)^2|}{2\nu_t + (\sigma^2 a + \lambda_t^2 b)^2} \\ &\leq \frac{2\sigma^2}{\sqrt{\nu_t}}\Big(\frac{1-\epsilon}{\epsilon}\Big) \leq \frac{2(1-\epsilon)}{\sigma_w\epsilon}, \end{aligned}$$

where in the final inequality we use $\sqrt{\nu_t}/\sigma^2 \geq \lambda_t$ and $\lambda_t \geq \sigma_w$ by (18). Using the above in (19), $|\frac{\partial}{\partial a}\eta_t(a,b)| \leq 1 + 2(1-\epsilon)/(\sigma_w\epsilon)$. A bound for $|\frac{\partial}{\partial b}\eta_t(a,b)|$ can be shown similarly.

### C. Numerical Examples

Finally, we provide numerical results to compare the empirical mean square error (MSE) of AMP-SI and the performance predicted by SE. Fig. 1 shows the MSE achieved by AMP-SI in the GG scenario and the SE prediction of its performance. In this example, the signal variance $\sigma_x^2 = 1$, the measurement noise variance $\sigma_w^2 = 0.01$, the variance of AWGN in SI

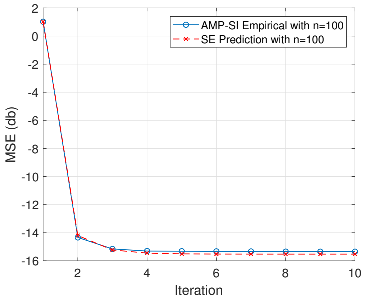

Fig. 1: Empirical MSE performance of AMP-SI and SE prediction. (GG model, $n = 100$, $m = 30$, $\sigma_x = 1$, $\sigma_w = 0.1$, and $\sigma = 0.2$.)

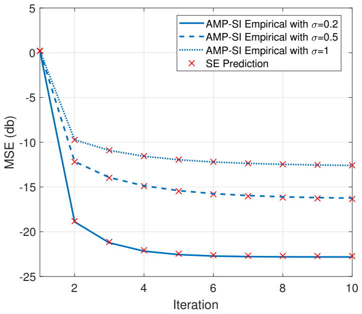

Fig. 2: Empirical MSE performance of AMP-SI and SE prediction. (BG model, $n = 10000$, $m = 3000$, $\epsilon = 0.2$, $\sigma_w = 0.1$)

$\sigma^2 = 0.04$. We averaged over 10 trials of a GG recovery problem for empirical results of AMP-SI. For smaller $n$ there is some gap between the empirical MSE and the SE prediction, as shown in Fig. 1 for $n = 100$, but the gap shrinks as $n$ is increased. (Fig. 2 shows results for the BG scenario with n=10000, and a similar plot for GG would also show the empirical MSE tracking the SE prediction nicely.)

Fig. 2 shows the MSE achieved by AMP-SI in the BG scenario, and the SE prediction of its performance. We again averaged over 10 trials of a BG recovery problem for empirical results of AMP-SI. The signal length $n = 10000$, $m = 3000$, the measurement noise variance $\sigma_w^2 = 0.01$, and $\epsilon = 0.2$, where 20% of the entries in the signal are nonzero. We vary the variance of AWGN in SI from $\sigma^2 = 0.04$, $\sigma^2 = 0.25$, and $\sigma^2 = 1$. The results show that SE can predict the MSE achieved by AMP-SI at every iteration.

## III. Proof of Theorem II.1

The proof of the functions in (10) being uniformly PL(2) when $\phi$ and $\psi$ are PL(2) is a straightforward application of Cauchy-Schwarz.

Next we show the asymptotic results given in (11). First we use Berthier *et al.* [11, Theorem 14] and then we make an appeal to the strong law of large numbers (SLLN):

**Theorem III.1. Strong Law of Large Numbers (SLLN)** *[15]: Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean $\mu$. Then the partial averages converge almost surely to $\mu < \infty$.*

We use Berthier *et al.* [11, Theorem 14], restated here for convenience. To apply the result in Berthier *et al.* [11, Theorem 14], one needs to justify the following assumptions:

**(C1)** The measurement matrix $\mathbf{A}$ has Gaussian entries with i.i.d. mean 0 and variance $1/m$.

**(C2)** Define a sequence of denoisers $\widetilde{\eta}_n^t : \mathbb{R}^n \to \mathbb{R}^n$ to be those that apply the denoiser $\eta_t$ defined in (6) elementwise as follows: $\widetilde{\eta}_n^t(\mathbf{x}) := \eta_t(\mathbf{x}, \widetilde{\mathbf{x}})$. For each $t$, $\widetilde{\eta}_n^t(\cdot)$ are uniformly Lipschitz. A function is *uniformly* Lipschitz in $n$ if the Lipschitz constant does not depend on $n$.

**(C3)** $||\mathbf{x}||_2^2/n$ converges to a constant as $n \to \infty$.

**(C4)** The limit $\sigma_w = \lim_{m\to\infty} ||\mathbf{w}||_2/\sqrt{m}$ is finite.

**(C5)** For any iterations $s, t \in \mathbb{N}$ and for any $2 \times 2$ covariance matrix $\mathbf{\Sigma}$, the following limits exist.

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\mathbf{Z}}[\mathbf{x}^T \widetilde{\eta}_n^t(\mathbf{x} + \mathbf{Z})] < \infty,$$

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\mathbf{Z},\mathbf{Z}'} \left[\widetilde{\eta}_n^t(\mathbf{x} + \mathbf{Z})^T \widetilde{\eta}_n^s(\mathbf{x} + \mathbf{Z}')\right] < \infty,$$

where $(\mathbf{Z}, \mathbf{Z}') \sim N(0, \mathbf{\Sigma} \otimes \mathbb{I}_n)$, with $\otimes$ denoting the tensor product and $\mathbb{I}_n$ the identity matrix.

**Theorem III.2.** *Under the assumptions* **(C1)** *–* **(C5)***, for any sequences of uniformly pseudo-Lipschitz functions $\rho_m : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ and $\gamma_n : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,*

$$\lim_m (\rho_m(\mathbf{r}^t, \mathbf{w}) - \mathbb{E}_{\mathbf{Z}_1}[\rho_m(\mathbf{w} + \sqrt{\lambda_t^2 - \sigma_w^2}\, \mathbf{Z}_1, \mathbf{w})]) \overset{p}{=} 0,$$

$$\lim_n \left(\gamma_n(\mathbf{x}^t + \mathbf{A}^T\mathbf{r}^t, \mathbf{x}) - \mathbb{E}_{\mathbf{Z}_2}\left[\gamma_n(\mathbf{x} + \lambda_t \mathbf{Z}_2, \mathbf{x})\right]\right) \overset{p}{=} 0,$$

*where $\mathbf{Z}_1 \sim \mathcal{N}(0, \mathbb{I}_m)$, $\mathbf{Z}_2 \sim \mathcal{N}(0, \mathbb{I}_n)$, $\mathbf{x}^t$ and $\mathbf{r}^t$ are defined in the AMP-SI recursion* (7)-(8)*, and $\lambda_t$ in the SE* (5)*.*

We demonstrate that assumptions **(A1)** – **(A4)** from Section II imply **(C1)** – **(C5)**, so we can apply Theorem III.2.

Assumptions **(A1)** and **(C1)** are identical.

Next consider assumption **(C2)**. The non-separable denoiser $\widetilde{\eta}_n^t(\mathbf{X}) = \eta_t(\mathbf{X}, \widetilde{\mathbf{X}})$ applies the AMP-SI denoiser defined in (6) entrywise to its vector inputs. From **(A4)**, $\{\eta_t(\cdot,\cdot)\}_{t\geq 0}$ are Lipschitz continuous. Thus, for length-$n$ vectors $\mathbf{x}_1, \mathbf{x}_2$, and fixed SI $\widetilde{\mathbf{x}}$, $||\widetilde{\eta}_n^t(\mathbf{x}_1) - \widetilde{\eta}_n^t(\mathbf{x}_2)|| \leq L||\mathbf{x}_1 - \mathbf{x}_2||$. The Lipschitz constant does not depend on $n$, so $\widetilde{\eta}_n^t(\cdot)$ is uniformly Lipschitz.

Assumptions **(C3)** and **(C4)** follow from the SLLN, Definition III.1: by **(A2)** and **(A3)**, $\mathbf{w}$ and $\mathbf{x}$ are entrywise i.i.d. with finite moment conditions $\mathbb{E}[|W|^2] < \infty$ and $\mathbb{E}[|X|^2] < \infty$.

We now show that **(C5)** is met. Define $y_i := x_i \mathbb{E}_Z[\eta_t(x_i + Z_i, \widetilde{x}_i)]$ for $i = 1, 2, \ldots, n$. By **(A3)**, $(\mathbf{X}, \widetilde{\mathbf{X}}) \sim$ i.i.d. $f(X, \widetilde{X})$, meaning $y_1, y_2, \ldots, y_n$ are also i.i.d.. By Def. III.1 if $\mathbb{E}[X\eta_t(X + Z, \widetilde{X})] < \infty$ where $Z \sim \mathcal{N}(0, \sigma_z^2)$ independent of $(\mathbf{X}, \widetilde{\mathbf{X}}) \sim$ i.i.d. $f(X, \widetilde{X})$, then $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n x_i \mathbb{E}_Z\left[\eta_t(x_i + Z_i, \widetilde{x}_i)\right] = \mathbb{E}[X\eta_t(X + Z, \widetilde{X})]$. We now show $\mathbb{E}[X\eta_t(X + Z, \widetilde{X})] < \infty$. By **(A4)**, $\eta_t(\cdot,\cdot)$ is Lipschitz, so it is easy to show that $|\eta_t(a_1, b_1)| \leq L'(1 + |a_1| + |b_1|)$ for some constant $L' > 0$. Using this bound and the triangle inequality,

$$\begin{aligned}\mathbb{E}[X\eta_t(X + Z, \widetilde{X})] &\leq L'\mathbb{E}[|X|(1 + |X + Z| + |\widetilde{X}|)] \\ &\leq L'(\mathbb{E}|X| + \mathbb{E}[X^2] + \mathbb{E}|X|\mathbb{E}|Z| + \mathbb{E}|X\widetilde{X}|).\end{aligned} \tag{20}$$

By **(A3)**, $\mathbb{E}[|X|^2], \mathbb{E}[|\widetilde{X}|^2]$, and $\mathbb{E}|X\widetilde{X}|$ are all finite. Then noting that for any random variable, $Y$, $|Y|^r \leq 1 + |Y|^k$ for $1 \leq r \leq k$, meaning $\mathbb{E}[|Y|]^r < 1 + \mathbb{E}[|Y|^k]$ the boundednes of $\mathbb{E}[X\eta_t(X + Z, \widetilde{X})]$ follows from (20) and **(A3)**. The proof of the second equation in **(C5)** follows similarly to the proof of the first equation in **(C5)**.

Now that we've justified **(C1)** – **(C5)**, we make an appeal to Theorem III.2 and the SLLN in order to finally prove (11).

The first result in (11), namely the asymptotic result for $\phi_m$ uniformly PL(2), follows *almost* immediately by applying Theorem III.2 using $\rho_m = \phi_m$. Namely, by Theorem III.2,

$$\lim_m (\phi_m(\mathbf{r}^t, \mathbf{w}) - \mathbb{E}_{\mathbf{Z}_1}[\phi_m(\mathbf{w} + \sqrt{\lambda_t^2 - \sigma_w^2}\, \mathbf{Z}_1, \mathbf{w})]) \overset{p}{=} 0$$

since $\phi_m$ is assumed to be uniformly PL(2). To complete the proof, we will finally prove that

$$\begin{aligned} &\lim_m \mathbb{E}_{\mathbf{Z}_1}[\phi_m(\mathbf{w} + \sqrt{\lambda_t^2 - \sigma_w^2}\, \mathbf{Z}_1, \mathbf{w})] \\ &\quad = \lim_m \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathbf{Z}_1}[\phi(w_i + \sqrt{\lambda_t^2 - \sigma_w^2}\, [\mathbf{Z}_1]_i, w_i)] \qquad (21) \\ &\quad \overset{a.s.}{=} \mathbb{E}[\phi(W + \sqrt{\lambda_t^2 - \sigma_w^2}\, Z_1, W)], \end{aligned}$$

where $W \sim f(W)$ independent of $Z_1$ standard Gaussian. Then the desired result follows since

$$\mathbb{E}[\phi_m(\mathbf{W} + \sqrt{\lambda_t^2 - \sigma_w^2}\, \mathbf{Z}_1, \mathbf{W})] = \mathbb{E}[\phi(W + \sqrt{\lambda_t^2 - \sigma_w^2}\, Z_1, W)].$$

Result (21) follows by the SLLN (Definition III.1) if $\mathbb{E}[\phi(W + \sqrt{\lambda_t^2 - \sigma_w^2}\, Z_1, W)]$ is finite. By Definition II.1 it is easy to see that if $\phi : \mathbb{R}^2 \to \mathbb{R}$ is PL(2), then there is a constant $L' > 0$ such that for all $\mathbf{x} \in \mathbb{R}^2 : |\phi(\mathbf{x})| \leq L'(1 + ||\mathbf{x}||^2)$. Using this,

$$\begin{aligned} &|\phi(W + \sqrt{\lambda_t^2 - \sigma_w^2}\, Z_1, W)| \leq L_1'(1 + ||(W + \sqrt{\lambda_t^2 - \sigma_w^2}\, Z_1, W)||^2) \\ &\leq L_1'(1 + 3|W|^2 + 2(\lambda_t^2 - \sigma_w^2)|Z_1|^2), \end{aligned} \qquad (22)$$

where we have used: for any $r > 0$ and any $a_1, a_2$ scalars, $||(a_1, a_2)||^2 = a_1^2 + a_2^2$ and $(|a_1| + |a_2|)^r \leq 2^{r-1}(|a_1|^r + |a_2|^r)$. Then, using (22), and the boundedness of $\mathbb{E}[|W|^2]$ by **(A2)**, $\mathbb{E}|\phi(W + \sqrt{\lambda_t^2 - \sigma_w^2}\, Z_1, W)| < \infty$.

The second result of (11) requires a bit more care as it is not immediate that the function $\gamma_n : \mathbb{R}^{2n} \to \mathbb{R}$ defined as $\gamma_n(\mathbf{a}, \mathbf{b}) := \psi_n(\mathbf{a}, \mathbf{b}, \widetilde{\mathbf{x}})$ for a sequence of side informations $\{\widetilde{\mathbf{x}}\}_n$ is uniformly PL(2) as needed to apply Theorem III.2. The next step of the proof deals with carefully handling this issue. We note that once we have shown that

$$\lim_n (\psi_n(\mathbf{x}^t + \mathbf{A}^T\mathbf{r}^t, \mathbf{x}, \widetilde{\mathbf{x}}) - \mathbb{E}_{\mathbf{Z}_2}[\psi_n(\mathbf{x} + \lambda_t \mathbf{Z}_2, \mathbf{x}, \widetilde{\mathbf{x}}]) \overset{p}{=} 0 \quad (23)$$

then the last step showing that

$$\lim_n \mathbb{E}_{\mathbf{Z}_2}[\psi_n(\mathbf{x} + \lambda_t \mathbf{Z}_2, \mathbf{x}, \widetilde{\mathbf{x}})] \overset{p}{=} \lim_n \mathbb{E}[\psi_n(\mathbf{X} + \lambda_t \mathbf{Z}_2, \mathbf{X}, \widetilde{\mathbf{X}})],$$

follows by the SLLN as in (21) - (22). However, the function $\gamma_n$ is not obviously uniformly PL(2) since an upper bound on $|\psi_n(\mathbf{a}, \widetilde{\mathbf{a}}, \widetilde{\mathbf{x}}) - \psi_n(\mathbf{b}, \widetilde{\mathbf{b}}, \widetilde{\mathbf{x}})|$ necessarily has an $||\widetilde{\mathbf{x}}||/\sqrt{n}$ factor. This is mainly a technicality as $||\widetilde{\mathbf{x}}||/\sqrt{n}$ is bounded by a constant (independent of $n$) with high probability.

To show (23) we would like to show that for any $\epsilon > 0$,

$$P(|\psi_n(\mathbf{x}^t + \mathbf{A}^T\mathbf{r}^t, \mathbf{x}, \widetilde{\mathbf{x}}) - \mathbb{E}_{\mathbf{Z}_2}[\psi_n(\mathbf{x} + \lambda_t \mathbf{Z}_2, \mathbf{x}, \widetilde{\mathbf{x}})]| > \epsilon) \to 0 \qquad (24)$$

as $n \to \infty$. Define a pair of events $\mathcal{T}_n(\epsilon)$ and $\mathcal{B}_n(C)$ as $\mathcal{T}_n(\epsilon) := \{|\psi_n(\mathbf{x}^t + \mathbf{A}^T\mathbf{r}^t, \mathbf{x}, \widetilde{\mathbf{x}}) - \mathbb{E}_{\mathbf{Z}_2}[\psi_n(\mathbf{x} + \lambda_t \mathbf{Z}_2, \mathbf{x}, \widetilde{\mathbf{x}})]| > \epsilon\}$ and for constant $C > 0$ independent of $n$, $\mathcal{B}_n(C) := \{\widetilde{\mathbf{x}} \in \mathbb{R}^n : ||\widetilde{\mathbf{x}}||/\sqrt{n} < C\}$. Then demonstrating (24) means showing, for any $\epsilon > 0$, that $\lim_n P(\mathcal{T}_n(\epsilon)) = 0$. Note that,

$$\begin{aligned} P(\mathcal{T}_n(\epsilon)) &= P(\mathcal{T}_n(\epsilon) \text{ and } \mathcal{B}_n(C)) + P(\mathcal{T}_n(\epsilon) \text{ and not } \mathcal{B}_n(C)) \\ &\leq P(\mathcal{T}_n(\epsilon)|\mathcal{B}_n(C)) + P(\text{not } \mathcal{B}_n(C)). \end{aligned}$$

Considering the above, the first term approaches 0 as $n$ gets large due to Theorem III.2, since one can argue $P(\mathcal{T}_n(\epsilon)|\mathcal{B}_n(C)) = P(\mathcal{T}_n(\epsilon)|\mathcal{B}_p(C)$ for all $p > p_0$) and conditional on the event $\mathcal{B}_p(C)$ being true for all integers $p > p_0$ (constant $p_0 > 0$), the function $\gamma_n$ defined in (III) is uniformly PL(2) in $n$. This uses that $\widetilde{\mathbf{x}}(n)$ is independent of the other random elements, namely $\mathbf{A}(n)$ and $\mathbf{w}(n)$. Next, by choosing $C$ large enough, the second probability $P(\text{not } \mathcal{B}_n(C))$ goes to zero almost surely by the SLLN as $||\widetilde{\mathbf{x}}||/\sqrt{n}$ concentrates to the elementwise expectation of $\widetilde{\mathbf{x}}$.

## Acknowledgment

We thank You (Joe) Zhou for insightful conversations and valuable advice. Liu and Baron acknowledge support from NSF EECS #1611112 and Rush from NSF #1217023.

## References

[1] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. Nat. Academy Sci.*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[2] H. Arguello and G. Arce, "Code aperture optimization for spectrally agile compressive imaging," *J. Opt. Soc. Am.*, vol. 28, no. 11, pp. 2400–2413, Nov. 2011.

[3] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Springer, Aug. 2001.

[4] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," *Arxiv preprint arXiv:1010.5141*, Oct. 2010.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 2006.

[6] D. Baron, A. Ma, D. Needell, C. Rush, and T. Woolf, "Conditional approximate message passing with side information," in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, 2017.

[7] A. Ma, Y. Zhou, C. Rush, D. Baron, and D. Needell, "An approximate message passing framework for side information," *arXiv:1807.04839*, July 2018.

[8] A. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE J. Select. Areas Commun.*, vol. 5, no. 2, pp. 128–137, Feb. 1987.

[9] 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.

[10] 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.

[11] R. Berthier, A. Montanari, and P. M. Nguyen, "State evolution for approximate message passing with non-separable functions," *arXiv:1708.03950*, Aug. 2017.

[12] X. Wang and J. Liang, "Approximate message passing-based compressed sensing reconstruction with generalized elastic net prior," *Signal Process. Image*, vol. 37, pp. 19–33, Sept. 2015.

[13] J. Ziniel and P. Schniter, "Dynamic compressive sensing of time-varying signals via approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5270–5284, Nov. 2013.

[14] A. Manoel, F. Krzakala, E. W. Tramel, and L. Zdeborová, "Streaming Bayesian inference: theoretical limits and mini-batch approximate message-passing," in *Proc. Annual Allerton Conf. on Commun., Control, and Comput.*, Oct. 2017, pp. 1048–1055.

[15] J. S. Rosenthal, *A First Look at Rigorous Probability Theory*, 2nd ed. World Scientific Publishing Co. Pte. Ltd., 2006.