An Approximate Message Passing Framework for Side Information

Anna Ma[1], You (Joe) Zhou[2], Cynthia Rush[3], Dror Baron[2], and Deanna Needell[4]

[1]Department of Mathematics, University of California San Diego
[2]Department of Electrical and Computer Engineering, NC State University
[3]Department of Statistics at Columbia University
[4]Department of Mathematics, University of California Los Angeles

July 16, 2018

Abstract

Approximate message passing (AMP) methods have gained recent traction in sparse signal recovery. Additional information about the signal, or *side information* (SI), is commonly available and can aid in efficient signal recovery. In this work, we present an AMP-based framework that exploits SI and can be readily implemented in various settings. To illustrate the simplicity and wide applicability of our approach, we apply this framework to a Bernoulli-Gaussian (BG) model and a time-varying birth-death-drift (BDD) signal model, motivated by applications in channel estimation. We develop a suite of algorithms, called AMP-SI, and derive denoisers for the BDD and BG models. We also present numerical evidence demonstrating the advantages of our approach, and empirical evidence of the accuracy of a proposed state evolution.

1 Introduction

The core focus of research in many disciplines, including but not limited to communication [7], compressive imaging [2], matrix completion [8], quantizer design [21], large-scale signal recovery [42], and sparse signal processing [9], is on accurately recovering a high-dimensional, unknown signal from a limited number of noisy linear measurements by exploiting probabilistic characteristics and structure in the signal.

We consider the following model for this task. For an unknown signal $x \in \mathbb{R}^N$,

$$y = Ax + z, \tag{1}$$

where $y \in \mathbb{R}^M$ are noisy measurements, $A \in \mathbb{R}^{M \times N}$ is the measurement matrix, and $z \in \mathbb{R}^M$ is measurement noise. The objective of signal recovery is to recover or estimate x from

knowledge of only y and A, and in some cases statistical knowledge about x and z. A great deal of effort has gone into developing schemes for such signal recovery, for example ℓ_1 minimization based approaches for sparse recovery [12, 38] and computationally efficient iterative algorithms [6, 14, 30], and supporting theory to tackle these challenges as datasets become larger and multidimensional. Approximate message passing (AMP) [14, 20] is an algorithmic framework for recovering sparse signals in high-dimensional regression tasks that is often used when prior information about the signal's distribution is available.

1.1 AMP for Signal Recovery

Approximate message passing or AMP [14, 20, 29] is a low-complexity algorithmic framework for efficiently solving high-dimensional regression tasks (1). AMP algorithms are derived as Gaussian or quadratic approximations of loopy belief propagation algorithms (e.g., min-sum, sum-product) on the dense factor graph corresponding to (1).

AMP has a few features that make it attractive for signal recovery. In certain problem settings, AMP offers convergence in linear time, and its performance can be tracked accurately with a simple scalar iteration known as state evolution (SE), discussed below. In addition, it is well-accepted the performance of AMP will be no worse than the best polynomial-time algorithms available [24].

AMP algorithm: The standard AMP algorithm [14] iteratively updates estimates of the unknown input signal, with $x^t \in \mathbb{R}^N$ being the estimate at iteration t. The algorithm is given by the following set of updates. Assume that x^0 is the all-zero vector and update for $t \geq 0$ with the following iterations:

$$r^t = y - Ax^t + \frac{r^{t-1}}{\delta} \left\langle \eta'_{t-1}(x^{t-1} + A^T r^{t-1}) \right\rangle, \tag{2}$$

$$x^{t+1} = \eta_t(x^t + A^T r^t). \tag{3}$$

Note that $\eta_t \colon \mathbb{R} \to \mathbb{R}$ is an appropriately-chosen sequence of functions and $\delta = \frac{M}{N}$ is the measurement rate. The functions $\{\eta_t(\cdot)\}_{t\geq 0}$ act element-wise on their vector inputs and have derivatives $\eta'_t(w) = \frac{\partial}{\partial w}\eta_t(w)$. Moreover, $\langle w \rangle = \frac{1}{N}\sum_{i=1}^N w_i$ is the empirical mean, where $w \in \mathbb{R}^N$. Here and throughout, we use capital letters to represent *random variables* (RVs) and lower case letters to represent realizations. We also denote a Gaussian RV with mean μ and variance σ^2 by $\mathcal{N}(\mu, \sigma^2)$.

Assuming that the measurement matrix A has independent and identically distributed (i.i.d.) $\mathcal{N}(0, 1/M)$ entries and the entries of the signal x are i.i.d. $\sim f(X)$, where $f(X)$ is the probability density function (pdf) of the signal, one useful feature of AMP is that the input to the denoiser, $x^t + A^T r^t$, which we refer to as the *pseudo-data*, is almost surely equal in distribution to the true signal x plus i.i.d. Gaussian noise with variance λ_t^2, a constant value given by the SE equations, introduced in (4) below, in the large system limit as $N \to \infty$ with fixed δ. These favorable statistical properties of the pseudo-data are due to the presence of the 'Onsager' term, $\frac{r^{t-1}}{\delta}\left\langle \eta'_{t-1}(x^{t-1} + A^T r^{t-1})\right\rangle$, used in the residual step (2) of the AMP updates.

State evolution (SE): One of AMP's attractive features is that under suitable conditions on A and x, its performance can be tracked accurately with a simple scalar iteration

referred to as state evolution (SE) [5, 33]. In particular, performance measures such as the ℓ_2-error or the ℓ_1-error in the algorithm's iterations concentrate to constants predicted by SE. The SE equations follow: let $\lambda_0 = \sigma_z^2 + \mathbb{E}[X^2]/\delta$ and for $t \geq 0$, we have

$$\lambda_t^2 = \sigma_z^2 + \frac{1}{\delta}\mathbb{E}\left[(\eta_{t-1}(X + \lambda_{t-1}Z) - X)^2\right], \tag{4}$$

where $X \sim f(X)$ is independent of $Z \sim \mathcal{N}(0,1)$ and λ_t^2 tracks the variance of the difference between the pseudo-data and signal at iteration t.

The AMP updates (2) - (3) rely on appropriately-chosen *denoisers* $\{\eta_t\}_{t\geq 0}$, which reduce the noise in the optimization task at each iteration. Owing to the favorable properties of the psuedo-data and the fact that one is often interested in evaluating the performance of the algorithm using the mean squared error (MSE), η_t in iteration t is often chosen to be the minimum mean squared error (MMSE) denoiser based on the pdf of x:

$$\eta_t(a) = \mathbb{E}[X|X + \lambda_t Z = a], \tag{5}$$

where $Z \sim \mathcal{N}(0,1)$, and $X \sim f(X)$ is a RV with the same pdf as that of x. See Section 2.3.1 for further insights of how SE behaves in our framework.

1.2 Side information

In information theory [13, 15], it is well known that when different communication systems share *side information* (SI), overall communication can happen more efficiently. As an example, when running a Bayesian signal recovery algorithm on an input x with an unknown probability density, feedback about the estimated density leads to improved signal recovery quality [19].

Signal recovery algorithms often have access to SI, denoted $\widetilde{x}$, that, as we will soon see, offers the potential to markedly improve recovery quality. For the noisy linear model of (1), SI has been shown to aid signal recovery when considering various application settings [10, 17, 22, 25–28, 32, 39–41]. For example, three dimensional (3D) video acquisition could be performed by acquiring each frame of video, which is a 2D image, independently of other frames using a single pixel camera [35]. While recovering the current frame, it is likely that one is simultaneously recovering the previous and next frames, which can be used as SI.

We will demonstrate that our approach is potentially useful in applications by studying a channel estimation problem in wireless communication systems (Fig. 1). In typical channel estimation scenarios, a wireless device transmits a pilot sequence and data payload in *batches*. In batch b, the pilot sequence p is transmitted into the channel, where it is convolved with the channel response x^b, yielding noisy linear measurements (details in Section 4.1). Not only is the channel response x^b in batch b sparse, the slowly time varying nature of the channel ensures that its differences relative to channel responses in previous batches are structured. Therefore, we can use $\widetilde{x} = \widehat{x}^{b-1}$, the channel response estimated in the previous batch, as SI while estimating x^b in the current batch. In Section 5, we demonstrate that SI in the above-mentioned batched manner helps AMP achieve lower MSE for a model motivated by channel estimation.

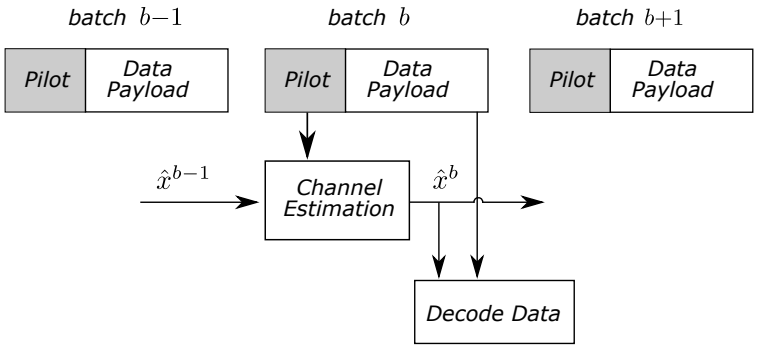

Figure 1: In batch b, the wireless device transmits a pilot and data payload. The channel filter x^b is estimated using the channel's response to the pilot along with SI $\widetilde{x} = \widehat{x}^{b-1}$, the channel filter estimated in the previous batch. The estimated $\widehat{x}^b$ is used to decode the data and as SI in the next batch to estimate x^{b+1}.

1.3 Contributions and Organization

In this work we develop a class of sparse signal recovery algorithms that integrate SI into AMP. Our main contribution is a flexible and general framework that incorporates SI in the denoiser of AMP in a Bayes-optimal manner and can be easily adapted to arbitrary dependencies between the signal and the SI. Moreover, our framework's conceptual simplicity allows us to extend existing SE results to AMP-SI as in (4); these SE results for signal recovery with SI are lacking in prior work [40], [43]. In the case where the SI is a Gaussian-noise corrupted view of the true signal, we rigorously show Bayes-optimality properties for AMP-SI. For more general cases, we demonstrate empirically that our proposed SE formulation tracks the AMP performance.

We demonstrate our framework through its application to two types of signals. First, a Bernoulli-Gaussian (BG) signal and second, motivated by the channel estimation problem discussed in Section 1.2, a time-varying birth-death-drift (BDD) signal. Although we only provide the details for these two models, it is conceptually intuitive to extend our proposed framework to different signal-SI relationships. Our numerical experiments show that our proposed framework achieves a lower MSE than other previously studied SI methods.

The remainder of the paper is organized as follows. In Section 2, we discuss the AMP algorithm and prior work in AMP approaches that utilize SI. We then present our AMP framework for SI. Next we discuss the BG model in Section 3, which is a simplified version of the BDD model studied in Section 4. In Section 5, we include numerical simulations demonstrating the good performance of AMP-SI. Section 6 concludes.

2 AMP with Side Information

2.1 Prior Work

While integrating SI (or prior information) into signal recovery algorithms is not new [11, 22, 26, 32, 40], our work is a unified framework within AMP that supports arbitrary dependencies between the $(X_n, \widetilde{X}_n)_{n=1}^N$ pairs. Prior work using SI has been either heuristic, limited to specific applications, or outside the AMP framework. For example, Wang and

Liang [40] proposed the Generalized Elastic Net Prior approach, which integrates SI into AMP for a specific signal prior density, but the method lacks Bayes optimality properties and is difficult to apply to other signal models. Our algorithmic framework overcomes these limitations through a generalized, Bayes optimal framework. Ziniel and Schniter [43] developed an AMP-based signal recovery algorithm for a time-varying signal model based on Markov processes for the support and amplitude. The Markov processes and corresponding dependencies between variables are captured by factor graph models. While our BDD model (details in Section 4) is closely related to their time-varying signal model, our emphasis is to introduce the AMP-SI framework and demonstrate how SI can be incorporated in AMP without needing to carefully craft factor graphs for every new signal model.

2.2 Our Approach: AMP-SI

In this paper we introduce a Bayes-optimal algorithmic framework that utilizes available SI. Our SI takes the form of an estimate $\widetilde{x} \in \mathbb{R}^N$, which is statistically dependent on the signal x through some joint pdf $f(X, \widetilde{X})$. We propose a *conditional denoiser*,

$$\eta_t(a, b) = \mathbb{E}[X | X + \lambda_t Z_1 = a, \widetilde{X} = b], \tag{6}$$

which provides an MMSE estimate of the signal while incorporating SI. We refer to our framework using the proposed denoiser (6) within the standard AMP algorithm (2) - (3) as the AMP-SI method. Namely, the AMP-SI algorithm is the following. Assume x^0 is the all-zero vector and update for $t \geq 0$:

$$r^t = y - Ax^t + \frac{r^{t-1}}{\delta} \left\langle \eta'_{t-1}(x^{t-1} + A^T r^{t-1}, \widetilde{x}) \right\rangle, \tag{7}$$

$$x^{t+1} = \eta_t(x^t + A^T r^t, \widetilde{x}). \tag{8}$$

Note that $\eta_t(\cdot, \cdot)$ is the denoising function proposed in (6), its derivative $\eta'_t(w, \cdot) = \frac{\partial}{\partial w}\eta_t(w, \cdot)$ is with respect to the first input, and $\langle w \rangle = \frac{1}{N}\sum_{i=1}^N w_i$ for $w \in \mathbb{R}^N$. The λ_t value in (6) is given by SE equations for AMP-SI: let $\lambda_0 = \sigma_z^2 + \mathbb{E}[X^2]/\delta$ and for $t \geq 0$,

$$\lambda_t^2 = \sigma_z^2 + \frac{1}{\delta}\mathbb{E}\left[\left(\eta_{t-1}(X + \lambda_{t-1}Z_1, \widetilde{X}) - X\right)^2\right], \tag{9}$$

where $(X, \widetilde{X}) \sim f(X, \widetilde{X})$ are independent of Z_1, which is a standard Gaussian RV. In comparison to standard AMP, the conditional denoiser function $\eta_t(\cdot, \cdot)$ uses SI to denoise the pseudo-data in AMP-SI.

We note that while there are rigorous theoretical results [5, 33] proving that for large N the pseudo-data is approximately equal (in distribution) to the true signal x plus i.i.d. Gaussian noise with variance λ_t^2, a constant value given by the SE equations, in the case of standard AMP (2) - (3) with the standard SE (4), we only *conjecture* that such a result is true for AMP-SI (7) - (8) with the corresponding SE (9). However, empirical evidence in Section 5 shows that the SE accurately tracks the MSE of the AMP-SI estimates, and we leave the theoretical study of such properties as future work, some details of which are discussed in Section 2.3.

To show that AMP-SI is conceptually intuitive to apply, while having great potential to improve signal estimation quality in applications where SI is available, we apply AMP-SI to a preliminary channel estimation model (Section 4). While using more realistic channel models is left for future work, our encouraging numerical results show that AMP-SI can be used beyond toy models such as BG (Section 3).

2.3 AMP-SI Theory

2.3.1 State Evolution Analysis

As mentioned previously, the performance of AMP (2)-(3) at each step of the algorithm can be *rigorously* characterized by the SE equations in (4). When the empirical density function of the unknown signal x converges to some pdf $f(X)$ on $\mathbb{R}$ and the denoisers $\{\eta_t(\cdot)\}_{t\geq 0}$ used in the AMP updates are applied element-wise to their input, Bayati and Montanari [5] proved that the SE accurately predicts AMP performance in the large system limit. For example, their result implies that the MSE, $\frac{1}{N}||x^t - x||^2$, equals $\delta(\lambda_t^2 - \sigma_z^2)$ almost surely in the large system limit, but moreover it characterizes the limiting constant values for a fairly general class of loss functions. Rush and Venkataramanan [33] provide a concentration version of the asymptotic result when the prior density of x is i.i.d. sub-Gaussian, showing that the probability of ϵ-deviation between various performance measures and their limiting constant values fall exponentially in N.

Considering AMP-SI, however, we cannot directly apply the theoretical results of Bayati and Montanari [5] or Rush and Venkataramanan [33]. Each entry n of our signal is generated according to the conditional density $f(X_n|\widetilde{X}_n)$, where the conditioning is on the value of the corresponding entry of the SI, meaning the signal x now has independent, but not identically distributed, entries. Owing to x no longer being i.i.d., the conditional denoiser (6) depends on the index n, meaning that different scalar denoisers will be used at different indices, based on different SI at different indices. Both results [5] and [33] require that the same denoiser function be applied to each element of the pseudo-data and our denoiser will change element-wise based on the SI.

Sketch of Future SE Proof. We conjecture that the proofs in [5] and [33] can be extended and sketch the steps of such a proof. Like the proofs in [5] and [33], we would start by analyzing the conditional distribution of the measurement matrix A at any iteration of the algorithm t, conditioned on the algorithm's previous output. These distributional properties of the measurement matrix would not change from the previous work. Using this conditional distribution of A, we would need to show a corresponding result for the conditional distribution of the difference between the pseudo-data and the true signal at each iteration, namely that the conditional distribution (conditional on the past output of the algorithm) has the form $x^t + A^T r^t - x \stackrel{d}{=} \lambda_t Z^t + \Delta^t$ where $Z^t \sim \mathcal{N}(0, \mathbb{I})$ is independent of the conditioning sigma-algebra, Δ^t is a deviation term, and λ_t is given by the AMP-SI SE (9). The final step would be to show that the deviation term is negligible when considering loss functions of interest, like the MSE, in the sense that the normalized ℓ_2 norm of Δ^t concentrates exponentially fast to 0. We note that the proof is inductive on the iteration number t – showing that the norm of Δ^t concentrates to 0 requires showing that the norms of various other quantities from the AMP updates (7) - (8) concentrate on predicted values

– making it technically involved. The full details are left for future work.

2.3.2 Bayes Optimality

When the conditional expectation denoiser (5) is used in AMP (2)-(3), the corresponding SE (4) in its convergent states coincides with Tanaka's fixed point equation [16,37], ensuring that if AMP runs until it converges, in the large system limit the result provides the best possible MSE achieved by any algorithm under certain problem conditions.

In the case that the SI available to the system is a Gaussian-noise corrupted view of the true signal, i.e., $\widetilde{X} = X + \mathcal{N}(0, \sigma_{SI}^2)$, it can be shown [4] that the fixed points of AMP-SI SE (9) coincide with the fixed points of AMP SE (4) with 'effective' measurement rate $\delta_{eff} = \delta/\mu$ and 'effective' measurement noise variance $\sigma_{eff}^2 = \mu\sigma^2$ where $0 \leq \mu \leq 1$ and the μ depends on the prior density of the signal and the SI noise variance σ_{SI}^2. The effective change in δ and σ^2 implies that the incorporation of Gaussian-noise corrupted SI via the AMP-SI algorithm gives us Bayes-optimal signal recovery for a standard (without SI) linear regression problem (1) with *more* measurements and *reduced* measurement noise variance than our own. The details of this argument are provided in Appendix C and first appeared in [4].

We expect that AMP-SI will have similar Bayes-optimality properties to the standard AMP, however proving this rigorously is theoretically difficult. The above analysis relies heavily on the Gaussianity of the SI noise and it is not clear if it can be generalized. Providing rigorous, theoretical guarantees for Bayes optimality is therefore left for future work.

3 Bernoulli-Gaussian Model

The BG model reflects the scenario in which one wants to recover a sparse signal and has access to SI in the form of the signal with additive white Gaussian noise (AWGN). In other words, at every iteration the algorithm has access to SI, $\widetilde{x}$, and pseudo-data, v^t, with

$$\widetilde{x} = x + \mathcal{N}(0, \widehat{\sigma}^2\mathbb{I}), \qquad v_t \approx x + \mathcal{N}(0, \lambda_t^2\mathbb{I}),$$

where the additive noise in the SI and pseudo-data are independent. The entries of x follow a BG pdf:

$$X_n \sim \epsilon \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x_n^2}{2}\right) + (1-\epsilon)\delta_0, \tag{10}$$

so that x is zero with probability $1-\epsilon$ and is standard Gaussian in nonzero entries. Here, δ_0 represents the Dirac delta function at 0.

3.1 The Conditional Denoiser with SI for BG

In this section we will derive the following result:

Result 1. *The AMP-SI denoiser (6) has the following closed form for the BG model:*

$$\eta(a,b) = \left(1 + R_{(a,b)}\right)^{-1} \left[\frac{a\widehat{\sigma}^2 + b\lambda_t^2}{\widehat{\sigma}^2 + \lambda_t^2 + \widehat{\sigma}^2\lambda_t^2}\right], \tag{11}$$

where $R_{(a,b)}$ is a ratio between probabilities (computed in (14)), $\widehat{\sigma}^2$ is the variance of the AWGN of the side information, and λ_t^2 is the variance of the AWGN of the pseudo-data at iteration t.

In what follows, the notation $\psi_{\tau^2}(x)$ refers to the zero-mean Gaussian density with variance τ^2 evaluated at x. We will use $f(\cdot)$ (or $f(\cdot,\cdot)$, $f(\cdot,\cdot,\cdot)$, and so on) to represent a generic pdf (or joint pdf) on the input. Before we begin the derivation of (11), we introduce a few lemmas relating to computations involving two RVs $A = \rho X + \mathcal{N}(0,\sigma_a^2)$ and $B = X + \mathcal{N}(0,\sigma_b^2)$. Deriving the conditional denoiser for BG (and later BDD) requires the joint pdf between A and B (Lemma 1), the product of two Gaussian pdfs (Lemma 2), and the expectation of X conditional on instances of A and B (Lemma 3).

Lemma 1. *Given instances a and b such that $A = \rho X + \mathcal{N}(0,\sigma_a^2)$ for some constant ρ, $B = X + \mathcal{N}(0,\sigma_b^2)$, and $X \sim \mathcal{N}(0,\sigma_x^2)$, the joint pdf between A and B is:*

$$f(a,b) = \frac{1}{\rho}\psi_{\sigma_x^2+\sigma_b^2}(b)\psi_{\frac{\sigma_x^2\sigma_b^2}{\sigma_x^2+\sigma_b^2}+\frac{\sigma_a^2}{\rho^2}}\left(\frac{\sigma_x^2 b}{\sigma_x^2+\sigma_b^2} - \frac{a}{\rho}\right),$$

assuming that the AWGN in A, AWGN in B, and X are independent.

Lemma 1 is proved in Appendix A. Below, we denote the $\mathcal{N}(\mu,\sigma^2)$ density evaluated at x by $\widetilde{\psi}_{\mu,\sigma^2}(x)$.

The next lemma provides a simplified expression for the product of two Gaussian densities.

Lemma 2. *For two Gaussian densities, $\widetilde{\psi}_{\mu_1,\sigma_1^2}(x) \times \widetilde{\psi}_{\mu_2,\sigma_2^2}(x)$ equals*

$$\widetilde{\psi}_{\left(\frac{\mu_1\sigma_2^2+\mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)}(x) \times \widetilde{\psi}_{(\mu_1-\mu_2,\sigma_1^2+\sigma_2^2)}(0).$$

The proof of Lemma 2 involves straightforward algebra and completing the square; the lemma could also be formulated as a convolution of three Gaussian densities.

The final lemma generalizes the conditional expectation of a Gaussian random variable X conditioned on the value of two noisy versions of X, particularly $A \sim \rho X + \mathcal{N}(0,\sigma_a^2)$ and $B \sim X + \mathcal{N}(0,\sigma_b^2)$. We will use the shorthand notation $\mathbb{E}[X \mid a, b]$ to mean

$$\mathbb{E}[X | a = \rho X + \mathcal{N}(0,\sigma_a^2), b = X + \mathcal{N}(0,\sigma_b^2)].$$

Lemma 3. *The conditional expectation of a Gaussian RV $X \sim \mathcal{N}(0,\sigma_x^2)$ given instances a and b such that $A \sim \rho X + \mathcal{N}(0,\sigma_a^2)$ for some constant ρ and $B \sim X + \mathcal{N}(0,\sigma_b^2)$ can be computed as:*

$$\mathbb{E}[X \mid a, b] = \frac{\rho\sigma_x^2\,\sigma_b^2 a + \sigma_x^2\sigma_a^2 b}{\sigma_x^2\,(\sigma_a^2 + \rho^2\sigma_b^2) + \sigma_a^2\sigma_b^2},$$

assuming that the AWGN in A, AWGN in B, and X are independent.

The proof of Lemma 3 can be found in Appendix B.

3.2 Derivation of the Denoiser with SI for BG

Using the aforementioned lemmas, we derive the conditional denoiser for the BG model.

Derivation of Result 1. To derive Result 1, note that

$$\eta(a,b) = \mathbb{E}[X|a = X + \mathcal{N}(0,\lambda_t^2), b = X + \mathcal{N}(0,\widehat{\sigma}^2)],$$

and therefore,

$$\eta(a,b) = \Pr(X \neq 0 \,|\, a,b)\mathbb{E}[X \,|\, a,b,X \neq 0]. \tag{12}$$

Simplifying the expression $\Pr(X \neq 0 \,|\, a,b)$,

$$\begin{aligned}\Pr(X \neq 0 \,|\, a,b) &= \frac{f(X \neq 0,a,b)}{f(X \neq 0,a,b) + f(X = 0,a,b)} \\ &= \left[1 + \frac{\Pr(X=0)f(a,b \,|\, X=0)}{\Pr(X \neq 0)f(a,b \,|\, X \neq 0)}\right]^{-1}.\end{aligned} \tag{13}$$

Note that here we slightly abuse the notation of a pdf with an event (i.e., $X \neq 0$ or $X = 0$) as an input to the density function. Considering the ratio in (13), define

$$R_{(a,b)} = \frac{\Pr(X=0)f(a,b \,|\, X=0)}{\Pr(X \neq 0)f(a,b \,|\, X \neq 0)}.$$

Conditioned on $X \neq 0$, we can compute $f(a,b \,|\, X \neq 0)$ using Lemma 1 with $\rho = 1$, $\sigma_x^2 = 1$, $\sigma_a^2 = \lambda_t^2$, and $\sigma_b^2 = \widehat{\sigma}^2$:

$$f(a,b|X \neq 0) = \psi_{1+\widehat{\sigma}^2}(b)\psi_{\frac{\widehat{\sigma}^2}{1+\widehat{\sigma}^2}+\lambda_t^2}\left(\frac{b}{1+\widehat{\sigma}^2} - a\right).$$

Also, when $X = 0$, A and B are independent so

$$\begin{aligned}f(a,b \,|\, X=0) &= f(a \,|\, X=0)f(b \,|\, X=0) \\ &= \psi_{\lambda_t^2}(a)\psi_{\widehat{\sigma}^2}(b).\end{aligned}$$

With these elements, we can compute $R_{(a,b)}$:

$$R_{(a,b)} = \frac{(1-\epsilon)\psi_{\lambda_t^2}(a)\psi_{\widehat{\sigma}^2}(b)}{\epsilon\psi_{1+\widehat{\sigma}^2}(b)\psi_{\frac{\widehat{\sigma}^2}{1+\widehat{\sigma}^2}+\lambda_t^2}\left(\frac{1}{1+\widehat{\sigma}^2}b - a\right)}. \tag{14}$$

The last term we must compute is the conditional expectation in (12). Using Lemma 3 with $\rho = 1$, $\sigma_x^2 = 1$, $\sigma_a^2 = \lambda_t^2$, and $\sigma_b^2 = \widehat{\sigma}^2$, we have that

$$\mathbb{E}[X|a,b] = \frac{\widehat{\sigma}^2 a + \lambda_t^2 b}{\lambda_t^2 + \widehat{\sigma}^2 + \lambda_t^2\sigma_b^2}. \tag{15}$$

Result 1 is obtained by combining the above computations. In particular, we have that

$$\eta(a,b) = \left(1 + R_{(a,b)}\right)^{-1}\mathbb{E}[X|a,b],$$

where $R_{(a,b)}$ and $\mathbb{E}[X|a,b]$ are computed in (14) and (15), respectively.

3.3 State Evolution for BG

Using the denoiser in (11), we can compute the SE equations (9). Letting $\delta = \frac{M}{N}$, we have $\lambda_0^2 = \frac{1}{\delta}\mathbb{E}[X^2] + \sigma_z^2$ and for $t \geq 0$,

$$\lambda_{t+1}^2 = \sigma_z^2 + \frac{1}{\delta}\mathbb{E}\left[(\eta_t(X + \lambda_t Z_2, X + \widehat{\sigma} Z_1) - X)^2\right],$$

where $\eta_t(\cdot, \cdot)$ is defined in (11), Z_1 and Z_2 are independent, standard Gaussian RVS that are independent of $X \sim f(X)$, and the expectation is with respect to Z_1, Z_2, and X.

4 Birth-Death-Drift Model

In this section, we investigate the application of AMP-SI on a stochastic signal model closely resembling the channel estimation problem in wireless communications.

4.1 Connections to Channel Estimation

BDD Motivation: Our channel estimation scenario is illustrated in Fig. 1. Typical wireless devices transmit a pilot sequence and data payload in *batches*. In batch b, the pilot sequence p is transmitted into the channel, where it is convolved with the channel response x^b, yielding noisy linear measurements,

$$y^b = \text{conv}(p, x^b) + z.$$

This convolution, $\text{conv}(\cdot, \cdot)$, can be expressed as the product of a Toeplitz matrix with a vector,

$$y^b = \text{Toeplitz}(p)x^b + z, \tag{16}$$

where $\text{Toeplitz}(p)$ is the Toeplitz matrix that corresponds to the pilot sequence p. To perform channel estimation using AMP-SI, we will consider (16) as a linear inverse problem (1), where $\text{Toeplitz}(p)$ is the measurement matrix. Our goal will be to estimate the channel response x^b in batch b using the noisy measurements y^b, matrix $\text{Toeplitz}(p)$, and $\widetilde{x} = \widehat{x}^{b-1}$, our estimate of the channel response in the previous batch, $b-1$ (Fig. 1). Our resulting estimate for the channel response, $\widehat{x}^b$, will then help us estimate the channel response in the next batch, x^{b+1}. To develop a conditional denoiser, we need a channel model that describes the channel response x^b, and especially its dependence on x^{b-1}, the channel response in the previous batch. We model the channel as an (unknown) finite impulse response (FIR) filter, whose taps correspond to the amplitude of the channel response at different delays. Many filter taps are close to zero, and this sparsity makes the channel estimation problem a sparse signal recovery task.

Due to the slowly varying time dynamics of the channel, x^b is not only sparse, but has strong dependencies with the channel response in adjacent batches. A possible model for changes from x^b to x^{b+1} involves (*i*) *birth* of new nonzeros in x^{b+1} (corresponding to new wireless paths); (*ii*) *death* of nonzeros in x^b that become zero in x^{b+1} (existing paths are obscured as the user moves); and (*iii*) slow *drift* of existing nonzeros. We call these time-varying channel dynamics a *birth-death-drift* (BDD) model. To demonstrate the efficacy of our BDD model, we looked at ray tracing simulations for a mobile user moving in an urban

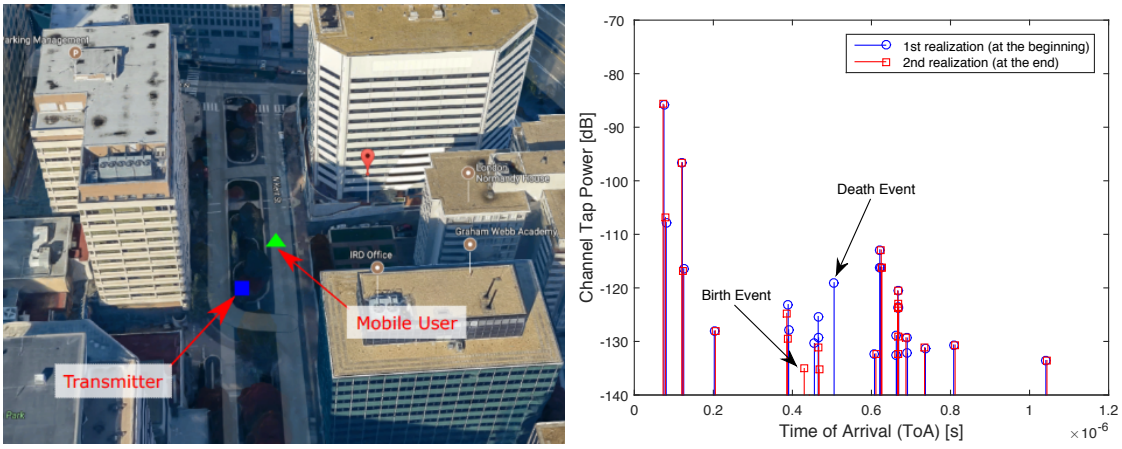

Figure 2: Ray tracing simulation results for a mobile user moving in an urban environment (top) show that the channel realizations at the beginning and end help (bottom) lend credence to our BDD model.

environment. A photo of the urban environment (a suburb of Washington, DC) is shown in the top panel of Fig. 2. The bottom panel shows two realizations of the channel filter. The realization corresponding to the beginning of the mobile user's motion is depicted by circles, and the realization corresponding to the end of the user's motion is marked by squares. It can be seen that most nonzero taps of the channel filter drift slowly; birth and death events are highlighted for the reader's convenience. Not only is the channel filter in each batch sparse, but its differences relative to filters in previous batches are highly structured.

The proposed BDD model resembles that of Saleh and Velanzuela [34], though their model does not involve time variation, in contrast to the one studied here. Their model also supports dependencies between filter taps within each batch; zeros and nonzeros tend to be clustered together within the communication channel. To keep things simple, our paper uses the BDD model for filter taps that are independent within each batch; possible inter-batch dependencies and corresponding non-separable denoisers [23, 36] are left for future work. For communication-minded readers, it should also be highlighted that AMP-SI is a flexible framework for incorporating SI. By demonstrating the efficacy of AMP-SI on this channel model, we believe the adaptation of AMP-SI to mmWave channel estimation is an exciting and promising direction for future research.

Formal definition of BDD model: To formally introduce the BDD model, we start by considering a single time batch. Between the previous and current batch, the signal elements independently change according to a BDD process which defines the joint pdf $f(X_p, X_c)$, where 'p' denotes the previous signal, a noisy version that serves as side information, and 'c' the current signal.

The elements of the signal evolve following four cases in the BDD model: for any entry $n \in 1, 2, \ldots, N$, **Case 1:** Zero entry remains zero, i.e., $[x_p]_n = 0$ and $[x_c]_n = 0$.
Case 2: *Death* – nonzero entry becomes zero, i.e., $[x_p]_n \sim \mathcal{N}(0, \sigma_s^2)$ and $[x_c]_n = 0$.
Case 3: *Drift* – nonzero entry remains nonzero, i.e., $[x_p]_n \sim \mathcal{N}(0, \sigma_s^2)$ and $[x_c]_n = \rho[x_p]_n + \mathcal{N}(0, \sigma^2)$.
Case 4: *Birth* – zero entry becomes nonzero, i.e., $[x_p]_n = 0$ and $[x_c]_n \sim \mathcal{N}(0, \sigma_s^2)$.

We define $\sigma^2 > 0$ to be the variance in the zero-mean Gaussian drift and $\sigma_s^2 > 0$ to be the steady-state variance, or the variance of the nonzero entries in the signal at every batch. Indeed, an entry of the current signal is nonzero in Cases 3 and 4, and by choosing the constant $\rho > 0$ such that $\rho^2\sigma_s^2 + \sigma^2 = \sigma_s^2$, we ensure $var(X_c) = \sigma_s^2$ for both these cases. Finally, Case j occurs with probability ϵ_j and $\sum_{j=1}^4 \epsilon_j = 1$.

Remark 1. The BG model is a simplified version of the BDD model. One can confirm that setting $\epsilon_2 = \epsilon_4 = 0$, $\epsilon_1 = 1 - \epsilon$, $\epsilon_3 = \epsilon$, $\sigma = 0$, and $\sigma_s^2 = 1$ obtains the model discussed in Section 3.

In the BDD model, the SI takes the form of the previous batch's signal x_p with AWGN. The pseudo-data, which we label v_t, is approximately the current batch's signal x_c with AWGN. That is, at every iteration the algorithm has access to:

$$\widetilde{x} = x_p + \mathcal{N}(0, \widehat{\sigma}^2\mathbb{I}), \qquad v_t \approx x_c + \mathcal{N}(0, \lambda_t^2\mathbb{I}),$$

where the additive noise in the SI and pseudo-data are independent. In the multiple batch setting, the pseudo-data in the final iteration of AMP-SI for approximating the b^{th} signal, which is a noisy version of x_p, becomes the SI for the approximation of the $(b+1)^{th}$ signal and the variance of this SI is available through λ_t^2 given by the SE equations 9.

4.2 The Conditional Denoiser with SI for BDD

We now derive the conditional denoiser for the BDD model presented in Section 4.1. Recall that the inputs a and b of the conditional denoiser $\eta(a, b)$ are instances of the pseudo-data v_t and SI $\widetilde{x}$, respectively.

Result 2. *The AMP-SI denoiser* (6) *has the following closed form for the BDD model,*

$$\begin{aligned}\eta(a,b) = &\frac{\epsilon_4\,\mu^4_{(a,b)}}{S_{(a,b)}}\left[\frac{\sigma_s^2\,a}{\sigma_s^2+\lambda_t^2}\right] \\ &+ \frac{\epsilon_3\,\mu^3_{(a,b)}}{S_{(a,b)}}\left[\frac{\sigma_s^2\,(\sigma^2+\widehat{\sigma}^2)\,a + \rho\,\sigma_s^2\,\lambda_t^2\,b}{\sigma_s^2\,(\sigma^2+\lambda_t^2+\widehat{\sigma}^2) + \lambda_t^2\,\widehat{\sigma}^2}\right],\end{aligned} \tag{17}$$

where $\epsilon_i\mu^i_{(a,b)}$ is the the joint pdf evaluated for Case i and instances a and b. Additionally, $S_{(a,b)}$ is the marginal pdf evaluated at instances a and b. The variables $\mu^3_{(a,b)}$, $\mu^4_{(a,b)}$, and $S_{(a,b)}$ are defined in (18) *below.*

In what follows, the notation $\psi_{\tau^2}(x)$ refers to the zero-mean Gaussian density with variance τ^2 evaluated at x.

$$\begin{aligned}\mu^3_{(a,b)} &= \psi_{\frac{\sigma_s^2\,(\widehat{\sigma}^2+\sigma^2)}{\widehat{\sigma}^2+\sigma_s^2}+\lambda_t^2}\left(\frac{\rho\,\sigma_s^2\,b}{\widehat{\sigma}^2+\sigma_s^2} - a\right)\psi_{\widehat{\sigma}^2+\sigma_s^2}(b), \\ \mu^4_{(a,b)} &= \psi_{\sigma_s^2+\lambda_t^2}(a)\,\psi_{\widehat{\sigma}^2}(b), \\ S_{(a,b)} &= \epsilon_1\,\psi_{\lambda_t^2}(a)\,\psi_{\widehat{\sigma}^2}(b) + \epsilon_2\,\psi_{\lambda_t^2}(a)\,\psi_{\widehat{\sigma}^2+\sigma_s^2}(b) \\ &\quad + \epsilon_3\,\mu^3_{(a,b)} + \epsilon_4\,\mu^4_{(a,b)}.\end{aligned} \tag{18}$$

4.3 Derivation of the Denoiser for BDD

Using the lemmas presented in Section 3, we derive the conditional denoiser for the BDD model.

Derivation of Result 2. To derive Result 2, note that

$$\eta(a,b) = \mathbb{E}[X_c | a = X_c + \mathcal{N}(0,\lambda_t^2), b = X_p + \mathcal{N}(0,\widehat{\sigma}^2)],$$

which we represent with shorthand $\mathbb{E}[X_c|a,b]$. Then,

$$\eta(a,b) = \sum_{j=3}^{4} \Pr(\text{Case j} \,|\, a,b)\mathbb{E}[X_c|a,b,\text{Case j}], \tag{19}$$

where we use the fact that $x_c = 0$ in Cases 1 and 2, and so $\mathbb{E}[X_c \,|\, a,b,\text{Case 1}] = \mathbb{E}[X_c|a,b,\text{Case 2}] = 0$.

Considering (19), let us simplify the expression $\Pr(\text{Case j} \,|\, a,b)$. In the following we use $f(\cdot)$ (or $f(\cdot,\cdot)$, $f(\cdot,\cdot,\cdot)$, and so on) to represent a generic pdf (or joint pdf) on the input. By Bayes' Rule,

$$\Pr(\text{Case j} \,|\, a,b) = \frac{f(\text{Case j},a,b)}{\sum_{i=1}^{4} f(\text{Case i},a,b)}. \tag{20}$$

To derive the denoiser (17) from (19) and (20), we must compute, for $j = \{1,2,3,4\}$:

$$f(\text{Case j},a,b) = \Pr(\text{Case j}) f(b|\text{Case j}) f(a|\text{Case j},b), \tag{21}$$

along with $\mathbb{E}[X_c \,|\, a,b,\text{Case 3}]$ and $\mathbb{E}[X_c \,|\, a,b,\text{Case 4}]$.

We first address Cases 1, 2, and 4 since $a = X_c + \mathcal{N}(0,\lambda_t^2)$ and $b = X_p + \mathcal{N}(0,\widehat{\sigma}^2)$ are independent in these cases. In Case 3, these values are *dependent* and therefore that case is handled carefully at the end.

Cases 1, 2, and 4: Here, we can simplify (21) by noting that $f(a \,|\, \text{Case j},b) = f(a \,|\, \text{Case j})$ due to the independence of a and b in these cases. For $j \in \{1,2,4\}$,

$$\begin{aligned} f(\text{Case j},a,b) &= \Pr(\text{Case j}) f(b \,|\, \text{Case j}) f(a \,|\, \text{Case j}) \\ &= \epsilon_j \, \psi_{\sigma^2_{b,j}}(b) \, \psi_{\sigma^2_{a,j}}(a), \end{aligned} \tag{22}$$

where $\sigma^2_{a,j} = \mathbb{E}[a^2 \,|\, \text{Case j}]$, and $\sigma^2_{b,j} = \mathbb{E}[b^2 \,|\, \text{Case j}]$. We also compute $\mathbb{E}[X_c \,|\, a,b,\text{Case 4}]$. This equals $\mathbb{E}[X_c \,|\, a,\text{Case 4}]$ since $b = \mathcal{N}(0,\widehat{\sigma}^2)$ is independent of X_c. Since $a = X_c + \mathcal{N}(0,\lambda_t^2)$, the conditional expectation is computed using a Wiener filter,

$$\mathbb{E}[X_c \,|\, a,\text{Case 4}] = \mathbb{E}[X_c \,|\, X_c + \mathcal{N}(0,\lambda_t^2)] = \frac{\sigma_s^2 a}{\sigma_s^2 + \lambda_t^2}. \tag{23}$$

Case 3: Here, $a = \rho X_p + \mathcal{N}(0,\sigma^2) + \mathcal{N}(0,\lambda_t^2)$ and $b = X_p + \mathcal{N}(0,\widehat{\sigma}^2)$ which, in contrast to the above cases, are now dependent through $X_p \sim \mathcal{N}(0,\sigma_s^2)$. To compute $f(\text{Case 3},a,b) = P(\text{Case 3}) f(a,b|\text{Case 3})$ note that conditional on Case 3, we may apply Lemma 1 to $f(a,b|\text{Case 3})$ with $X = X_p$, $\sigma_a^2 = \sigma^2 + \lambda_t^2$, and $\sigma_b^2 = \widehat{\sigma}^2$ to obtain:

$$f(\text{Case 3},a,b) = \Pr(\text{Case 3}) \, f(a,b|\text{Case 3})$$

$$= \frac{\epsilon_3}{\rho} \psi_{\sigma_s^2+\widehat{\sigma}^2}(b)\, \psi_{\frac{\sigma_s^2(\widehat{\sigma}^2+\sigma^2)}{\sigma_s^2+\widehat{\sigma}^2}+\frac{\sigma^2+\lambda_t^2}{\rho^2}} \left(\frac{\sigma_s^2 b}{\sigma_s^2+\widehat{\sigma}^2} - \frac{a}{\rho} \right). \tag{24}$$

We also need to compute $\mathbb{E}[X_c|a, b, \text{Case 3}]$. By linearity of expectation we have

$$\begin{aligned} &\mathbb{E}[X_c \,|\, a, b, \text{Case 3}] = \mathbb{E}[\rho X_p + \mathcal{N}(0, \sigma^2) \,|\, a, b, \text{Case 3}] \\ &= \rho \mathbb{E}[X_p \,|\, a, b, \text{Case 3}] + \mathbb{E}[\mathcal{N}(0, \sigma^2) \,|\, a, b, \text{Case 3}]. \end{aligned} \tag{25}$$

Conditional on Case 3, we can compute the first expectation in (25) using Lemma 3 with $X = X_p$, $\sigma_a^2 = \sigma^2 + \lambda_t^2$ since $a = \rho X_p + \mathcal{N}(0, \sigma^2 + \lambda_t^2)$, and $\sigma_b^2 = \widehat{\sigma}^2$ since $b = X_p + \mathcal{N}(0, \widehat{\sigma}^2)$:

$$\begin{aligned} \mathbb{E}[X_p|a, b, \text{Case 3}] &= \frac{\sigma_s^2[\rho\,\widehat{\sigma}^2 a + (\sigma^2 + \lambda_t^2) b]}{\sigma_s^2(\sigma^2 + \lambda_t^2 + \rho^2\widehat{\sigma}^2) + (\sigma^2 + \lambda_t^2)\widehat{\sigma}^2} \\ &= \frac{\sigma_s^2[\rho\,\widehat{\sigma}^2 a + (\sigma^2 + \lambda_t^2) b]}{\sigma_s^2(\sigma^2 + \lambda_t^2 + \widehat{\sigma}^2) + \lambda_t^2\widehat{\sigma}^2}, \end{aligned} \tag{26}$$

where we use the fact that $\rho^2\sigma_s^2 + \sigma^2 = \sigma_s^2$ to simplify. Letting $Z_c \sim \mathcal{N}(0, \sigma^2)$ be such that $X_c = \rho X_p + Z_c$, one can use the same approach as in Lemma 3 to obtain:

$$\begin{aligned} &\mathbb{E}[Z_c \,|\, a = Z_c + \rho X_p + \mathcal{N}(0, \lambda^2), \qquad\qquad (27) \\ &\qquad\qquad b = X_p + \mathcal{N}(0, \widehat{\sigma}^2), \text{Case 3}] \\ &= [\sigma^2 \;\; 0] \begin{bmatrix} \rho^2\sigma_s^2 + \lambda^2 & \rho\sigma_s^2 \\ \rho\sigma_s^2 & \sigma_s^2 + \widehat{\sigma}^2 \end{bmatrix}^{-1} \begin{bmatrix} a \\ b \end{bmatrix} \\ &= \frac{\sigma^2[(\sigma_s^2 + \widehat{\sigma}^2)a - \rho\sigma_s^2 b]}{\sigma_s^2(\sigma^2 + \lambda_t^2 + \widehat{\sigma}^2) + \lambda_t^2\widehat{\sigma}^2}. \qquad\qquad (28) \end{aligned}$$

Combining (26) and (28):

$$\begin{aligned} &\mathbb{E}[X_c \,|\, a, b, \text{Case 3}] \\ &= \rho \mathbb{E}[X_p \,|\, a, b, \text{Case 3}] + \mathbb{E}[\mathcal{N}(0, \sigma^2) \,|\, a, b, \text{Case 3}] \\ &= \frac{\rho\sigma_s^2[\rho\,\widehat{\sigma}^2 a + (\sigma^2 + \lambda_t^2) b] + \sigma^2[(\sigma_s^2 + \widehat{\sigma}^2)a - \rho\sigma_s^2 b]}{\sigma_s^2(\sigma^2 + \lambda_t^2 + \widehat{\sigma}^2) + \lambda_t^2\widehat{\sigma}^2} \\ &= \frac{\sigma_s^2(\sigma^2 + \widehat{\sigma}^2)a + \rho\sigma_s^2\lambda_t^2 b}{\sigma_s^2(\sigma^2 + \lambda_t^2 + \widehat{\sigma}^2) + \lambda_t^2\widehat{\sigma}^2}. \end{aligned} \tag{29}$$

Result 2 is obtained by combining the above calculations. Considering (19) and (20),

$$\eta(a, b) = \frac{\sum_{j=3}^4 f(\text{Case j}, a, b)\mathbb{E}[X_c|a, b, \text{Case j}]}{\sum_{i=1}^4 f(\text{Case i}, a, b)}, \tag{30}$$

which results in the denoiser presented in (17) - (18) with $S_{(a,b)} = \sum_{i=1}^4 f(\text{Case i}, a, b)$, where the probabilities are calculated in (22) and (24), $\epsilon_3\, \mu_{(a,b)}^3 = f(\text{Case 3}, a, b)$ and $\epsilon_4\, \mu_{(a,b)}^4 = f(\text{Case 4}, a, b)$, and finally with $\mathbb{E}[X_c \,|\, a, b, \text{Case 3}]$ and $\mathbb{E}[X_c \,|\, a, b, \text{Case 4}]$ calculated in (32) and (23), respectively.

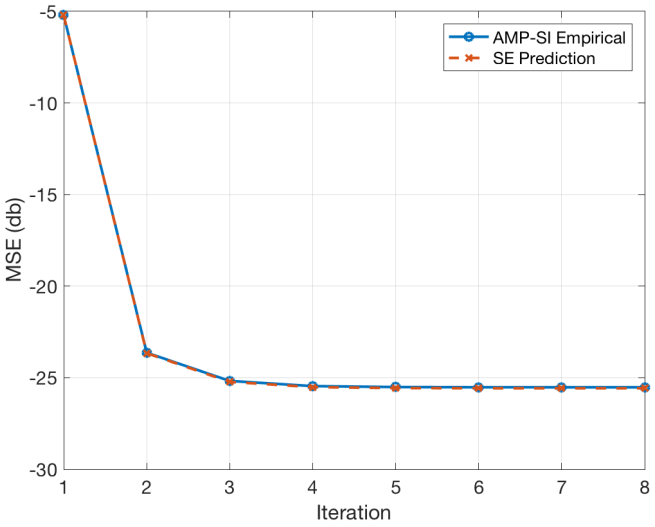

Figure 3: Empirical performance of AMP-SI and performance predicted by SE across iterations. (BG signal, $N = 10000$, $M = 3000$, $\sigma_z = 0.1$, $\epsilon = 0.3$, $\widehat{\sigma} = 0.1$.)

4.4 State Evolution for BDD

Using the results from the previous section, specifically the form of the denoiser in (17), we can calculate the SE equations (9). Letting $\delta = \frac{M}{N}$, we have $\lambda_0^2 = \frac{1}{\delta}\mathbb{E}[X_c^2] + \sigma_z^2$ and for $t \geq 0$,

$$\lambda_{t+1}^2 = \sigma_z^2 + \frac{1}{\delta}\mathbb{E}\left[(\eta_t(X_c + \lambda_t Z_2, X_p + \widehat{\sigma} Z_1) - X_c)^2\right],$$

where $\eta_t(\cdot,\cdot)$ is defined in (17), and the RVs Z_1 and Z_2 are both zero mean unit norm Gaussian, and are independent of the RVs X_p and X_c, which are distributed according to the prior distributions of x_p and x_c. The expectation is with respect to Z_1, Z_2, X_p, and X_c, where X_p and X_c are dependent. Because the form of the denoiser given in (17) is complicated, it seems infeasible to find a closed-form value for the expectation in the SE equations, so we estimate these values numerically.

5 Numerical Results

Here, we present a comparison between the empirical performance of AMP-SI and AMP for the BG and BDD signal models. All numerical results were generated using MATLAB.

BG signal: Fig. 3 presents the empirical performance of AMP-SI on a BG signal and the SE prediction of its performance. For this experiment, the signal has dimension $N = 10000$, the SI has standard deviation $\widehat{\sigma} = 0.10$, the number of measurements is $M = 3000$, and the measurement noise standard deviation is $\sigma_z = 0.10$. We set $\epsilon = 0.30$ so that approximately 30% of the entries in the signal are nonzero. The measurement matrix $A \in \mathbb{R}^{M \times N}$ has i.i.d. standard Gaussian entries. The empirical normalized MSE for AMP-SI is averaged over 20 trials of a BG recovery problem. We are also plotting MSE results predicted by SE, and it can be seen that the SE prediction accurately tracks the empiricial performance of AMP-SI.

BDD signal: Fig. 4 presents experimental results for recovering a signal x_c over 10 time batches following the BDD model of Section 4. In each time batch, the SI is the pseudo-data output from AMP-SI in the previous batch, except for the first batch where no SI is available and we default to standard AMP. The signal is of dimension $N = 10000$, the steady-state standard deviation is

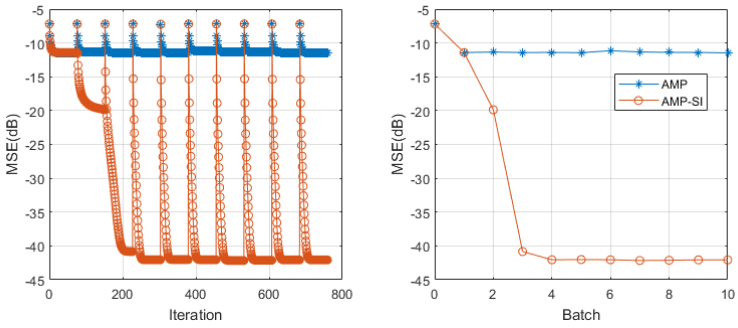

Figure 4: Empirical performance of AMP-SI and AMP across (left) iteration and (right) time batches. (BDD signal, $N = 10000$, $M = 3000$, $\sigma_s = 1$, $\rho = 0.95$, $\sigma_z = 0.01$, $\epsilon_1 = 0.80$, $\epsilon_2 = \epsilon_4 = 0.01$, $\epsilon_3 = 0.18$.)

$\sigma_s = 1$, the decay rate of nonzeros is $\rho = 0.95$, and the measurement noise has standard deviation $\sigma_z = 0.01$. The empirical MSE is averaged over 20 trials. For each batch, AMP-SI (or AMP in the first iteration) runs for 10 iterations. We set $\epsilon_1 = 0.80$, $\epsilon_2 = \epsilon_4 = 0.01$, and $\epsilon_3 = 0.18$ so that there are approximately $K = N(\epsilon_3 + \epsilon_4) = 1900$ nonzero entries per signal. The measurement matrix has i.i.d. standard Gaussian entries in each batch, and the number of measurements is $M = 3000$. It can be seen that AMP-SI outperforms AMP in every batch (expect Batch 1 where they are both AMP since no SI is available).

SE for BDD: To highlight the advantages of SI, Fig. 5 shows the recovery quality predicted by SE. Here, all parameters are set as in the experiments used for Fig. 4, except for the number of measurements M (to show different $\delta = M/N$) and ϵ_1 and ϵ_3 (to show different percentages of nonzeros, $\gamma = K/N$). To vary γ, we keep $\epsilon_2 = \epsilon_4 = 0.01$ while modifying the probability of the drift case, ϵ_3, accordingly. In each panel, the horizontal axis corresponds to δ, the vertical axis to γ, and shades of gray to the SE prediction of mean squared error (MSE). Batch 1 corresponds to the first time the signal is recovered without SI, Batch 3 uses recovered signals from the second batch as SI, and Batch 10 uses the recovered signal from Batch 9 as SI. The high-quality dark gray region in the upper right portion of each panel is expanding, while the low-quality light gray region is shrinking, showing improved signal recovery due to the SI. It can be seen that the same MSE quality is obtained from a measurement rate δ lower than without SI.

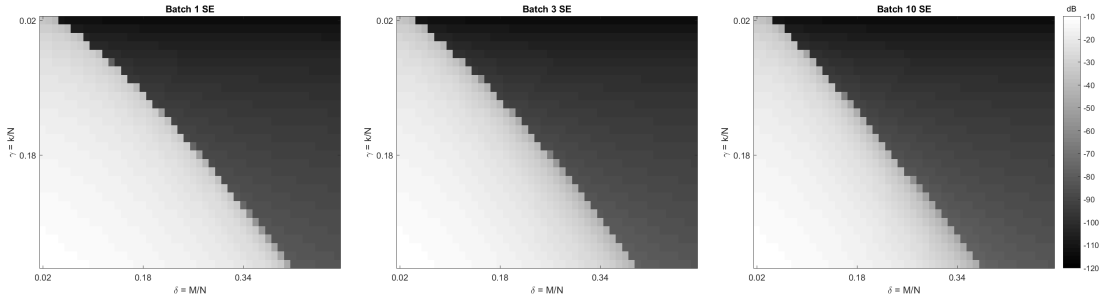

Figure 5: AMP-SI for BDD signals. The MSEs predicted by SE are plotted as shades of gray; they are functions of measurement rate $\delta = \frac{M}{N}$ and sparsity rate $\gamma = \frac{k}{N}$. From left to right: Batch 1 (without SI), Batch 3 (SI=Batch 2), and Batch 10 (SI=Batch 9). The 'good' dark gray region (upper right corners) expands with more SI.

Channel estimation with Topelitz matrices: So far we used i.i.d. Gaussian matrices, and we now transition to Toeplitz matrices in order to demonstrate that AMP-SI is suitable for channel

estimation (details in Section 4). Based on (16), the channel estimation problem deviates from the BDD model in two aspects. First, as mentioned, A is Toeplitz rather than i.i.d. Gaussian. It is well known that for non-i.i.d. sensing matrices, the standard AMP prescribed by (2) and (3) often suffers from divergence over iterations. A common approach to improve convergence of iterative algorithms is damping; in AMP, the standard iteration (3) is replaced by $x^{t+1} = \lambda x^t + (1-\lambda)\eta_t(x^t + A^T r^t)$. Rangan et al. [31] demonstrate that damping is effective in aiding the convergence of AMP for some non-i.i.d. sensing matrices. Second, for a pilot sequence p, the number of rows of the measurement matrix, M, equals length$(p) + N - 1$, which typically exceeds N, the number of columns. This inverse problem is expansive ($M > N$) instead of compressive ($M < N$), where we remind the reader that AMP and SE theory support arbitrary $\delta > 0$ where $\delta = \frac{M}{N}$.

Our experiment had 5 time batches. We set the length of the channel response N to 4000, the length of the pilot sequence length$(p) = 1001$, the standard deviation of the steady signal $\sigma_s = 1$, the decay rate of nonzeros $\rho = 0.95$, and the measurement noise standard deviation $\sigma_z \in \{0.01, 0.1, 1\}$. This setting corresponds to SNR= 0dB, 20dB and 40dB, and $\delta = 1.25$. For BDD model parameters, we set $\epsilon_1 = 0.78$, $\epsilon_2 = \epsilon_4 = 0.01$. Thus at each time batch, 21% of the entries of the channel response are nonzero. The individual entries of the pilot p are $\pm 1/\sqrt{\text{length}(p)} = \pm 0.0316$, each with probability 0.5. We performed damping using parameter $\lambda = 0.9$. Table 1 demonstrates the empirical channel estimation performance of AMP-SI averaged over 50 realizations. Compared to standard AMP (batch 1 in Table 1), AMP-SI consistently achieves lower MSE levels starting from batch 2. One striking observation from Fig. 6 is the similar performance of AMP-SI for Toeplitz (channel estimation) and i.i.d. matrices. This similarity leads us to conjecture that for the given BDD signal model, SE prediction tracks the performance of AMP/AMP-SI with Toeplitz matrices as well as the i.i.d. Gaussian case. The conjecture is further evident from Table 2. Observations from other BDD time batches resemble batch 5 (Table 2) and not included.

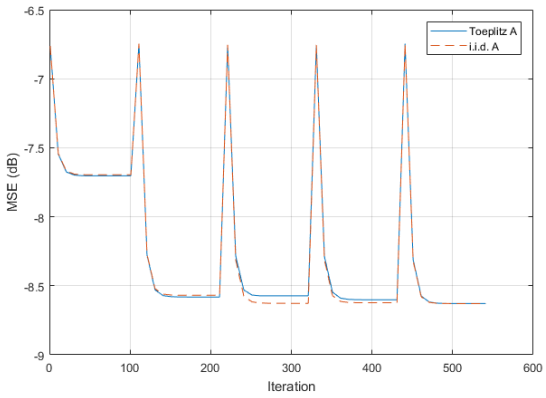

Figure 6: Empirical AMP-SI performance with i.i.d. and Toeplitz sensing matrices. (BDD signal, SNR=0dB, averaged over 100 realizations.)

6 Challenges and Future Work

In this work, we presented AMP-SI, a suite of Approximate Message Passing (AMP) based algorithms that utilize *side information* (SI) to aid in signal recovery using *conditional* denoisers. We derive conditional denoisers for a Bernoulli-Gaussian (BG) signal model and a more complicated time-varying birth-death-drift (BDD) signal model, motivated by channel estimation, to show the wide-applicability of our work. We also conjectured state evolution (SE) properties. Numerical

	Channel Estimation MSE(dB)		
SNR	**Batch 1**	**Batch 2**	**Batch 5**
0dB	-7.70	-8.58	-8.63
20dB	-23.48	-24.78	-24.79
40dB	-45.34	-45.76	-45.80

Table 1: Empirical AMP-SI performance for channel estimation. (BDD signal, averaged over 50 realizations.)

	AMP-SI Performance in MSE(dB) at Time Batch 5		
SNR	**i.i.d. A**	**SE Prediction**	**Toeplitz A**
0dB	-8.63	-8.63	-8.63
20dB	-24.79	-24.86	-24.79
40dB	-45.72	-45.91	-45.80

Table 2: Empirical AMP-SI performance for i.i.d. matrices, SE predictions, and empirical performance for Toeplitz matrices. (BDD signal, Batch 5, averaged over 50 realizations.)

experiments show that the proposed SE accurately tracks the performance of AMP-SI, and that AMP-SI achieves the same MSE as AMP using a lower measurement rate.

To simulate the channel estimation task, we additionally consider a Toeplitz measurement matrix as opposed to the standard Gaussian i.i.d. matrix. Our results show that AMP-SI is able to obtain a lower MSE than AMP for such a setting. A challenge and future direction with this line of work is that the current theoretical guarantees for AMP assume that A is an i.i.d. matrix. Although AMP often diverges when non-i.i.d. matrices are used, there is empirical evidence that AMP can successfully perform deconvolution and utilize other structures in various settings [3, 18]. We leave these challenges and the rigorous proofs of our conjectures for future work.

Acknowledgments

We are grateful to Yavuz Yapici and Ismail Guvenc who helped us formulate the BDD model, and to Chethan Anjinappa whose numerical ray tracing simulation (Fig. 2) helped confirm the model. Our work originated from earlier work on AMP with side information, which was joint with Tina Woolf. Finally, we thank Junan Zhu and Yanting Ma for helping us formulate the problem and master some of the deeper technical details.

Appendix

A Proof of Lemma 1

Recall from the Lemma statement that $A = \rho X + \mathcal{N}(0, \sigma_a^2)$ and $B = X + \mathcal{N}(0, \sigma_b^2)$ where $X \sim \mathcal{N}(0, \sigma_x^2)$.

Then from Bayes' rule, $f(a, b) = f(b)f(a \,|\, b)$ and computing $f(a \,|\, b)$ we have:

$$f(a \,|\, b) = \int_x f(a, x \,|\, b)\, dx = \int_x f(x \,|\, b) f(a \,|\, b, x)\, dx$$

$$\stackrel{(1)}{=} \int_x \frac{f(x)f(b\,|\,x)}{f(b)} \psi_{\sigma_a^2}(a-\rho x)\,dx$$
$$= \int_x \frac{\psi_{\sigma_x^2}(x)\psi_{\sigma_b^2}(b-x)}{f(b)} \psi_{\sigma_a^2}(a-\rho x)\,dx,$$

where equality (1) relies on Bayes' rule applied to $f(x\,|\,b)$. Therefore,

$$f(a,b) = f(b)f(a\,|\,b)$$
$$= \int_x \psi_{\sigma_x^2}(x)\psi_{\sigma_b^2}(b-x)\psi_{\sigma_a^2}(a-\rho x)\,dx$$
$$\stackrel{(2)}{=} \frac{1}{\rho}\psi_{\sigma_x^2+\sigma_b^2}(b)\psi_{\frac{\sigma_x^2\sigma_b^2}{\sigma_x^2+\sigma_b^2}+\frac{\sigma_a^2}{\rho^2}}\left(\frac{\sigma_x^2}{\sigma_x^2+\sigma_b^2}b - \frac{1}{\rho}a\right),$$

where equality (2) uses Lemma 2.

B Proof of Lemma 3

Recall from the Lemma statement that $A = \rho X + \mathcal{N}(0,\sigma_a^2)$ and $B = X + \mathcal{N}(0,\sigma_b^2)$ where $X \sim \mathcal{N}(0,\sigma_x^2)$.

Because X, A, and B are jointly Gaussian RVs, the MMSE-optimal estimator for X conditioned on a and b is linear,

$$\widehat{x} = \mathbb{E}[X\,|\,a,b] = \alpha a + \beta b + \gamma, \tag{31}$$

where α, β, and γ are constants. A well known result (see, e.g., Theorem 9.1 of [1]) states that $\widehat{x} = W\begin{bmatrix}a\\b\end{bmatrix} + U$, where

$$W = C_1^T(C_2)^{-1}, \quad U = \mathbb{E}[X] - W\mathbb{E}\begin{bmatrix}A\\B\end{bmatrix},$$
$$C_1 = \mathrm{Cov}\left(X, \begin{bmatrix}A\\B\end{bmatrix}\right), \quad C_2 = \mathrm{Cov}\left(\begin{bmatrix}A\\B\end{bmatrix}, \begin{bmatrix}A\\B\end{bmatrix}\right).$$

We compute these terms one by one. First, X, A, and B all have zero mean, and so $U = 0$, which implies that the constant γ in the linear form (31) is zero. Second,

$$C_1 = \mathrm{Cov}\left(X, \begin{bmatrix}A\\B\end{bmatrix}\right) = \begin{bmatrix}\mathbb{E}[XA]\\\mathbb{E}[XB]\end{bmatrix},$$

because the zero means ensure that only the cross terms $\mathbb{E}[XA]$ and $\mathbb{E}[XB]$ appear in the expression for C_1. The cross terms are computed as

$$\mathbb{E}[XA] = \mathbb{E}[X(\rho X + \mathcal{N}(0,\sigma_a^2))] = \rho\sigma_x^2,$$
$$\mathbb{E}[XB] = \mathbb{E}[X(X + \mathcal{N}(0,\sigma_b^2))] = \sigma_x^2.$$

Therefore, $C_1 = \sigma_x^2\begin{bmatrix}\rho\\1\end{bmatrix}$. Third,

$$C_2 = \mathrm{Cov}\left(\begin{bmatrix}A\\B\end{bmatrix}, \begin{bmatrix}A\\B\end{bmatrix}\right) = \mathbb{E}\left[\begin{bmatrix}A\\B\end{bmatrix}[A\ B]\right],$$

where once again only the cross terms need be computed. These cross terms are (i) $\mathbb{E}[A^2] = \rho^2\sigma_x^2 + \sigma_a^2$; (ii) $\mathbb{E}[B^2] = \sigma_x^2 + \sigma_b^2$; and (iii)

$$\begin{aligned}\mathbb{E}[AB] = \mathbb{E}[BA] &= \mathbb{E}[(\rho X + \mathcal{N}(0,\sigma_a^2))(X + \mathcal{N}(0,\sigma_b^2))] \\ &= \rho\sigma_x^2.\end{aligned}$$

The MMSE-optimal estimator is

$$\begin{aligned}\mathbb{E}[X \mid a,b] &= W\begin{bmatrix}a\\b\end{bmatrix} = C_1^T(C_2)^{-1}\begin{bmatrix}a\\b\end{bmatrix} \\ &= [\rho\sigma_x^2 \ \sigma_x^2]\begin{bmatrix}\rho^2\sigma_x^2+\sigma_a^2 & \rho\sigma_x^2 \\ \rho\sigma_x^2 & \sigma_x^2+\sigma_b^2\end{bmatrix}^{-1}\begin{bmatrix}a\\b\end{bmatrix} \\ &= \frac{\rho\sigma_x^2\,\sigma_b^2 a + \sigma_x^2\sigma_a^2 b}{\sigma_x^2\,(\sigma_a^2+\rho^2\sigma_b^2)+\sigma_a^2\sigma_b^2}.\end{aligned} \tag{32}$$

C Fixed points of AMP-SI SE with Gaussian SI

This appendix will show that when the SI is a Gaussian-noise corrupted observation of the true signal, i.e., $\widetilde{X} = X + \mathcal{N}(0,\sigma_{SI}^2)$, the fixed points of AMP-SI SE (9) coincide with the fixed points of AMP SE (4) with 'effective' measurement rate $\delta_{eff} = \delta/\mu$ and 'effective' measurement noise variance $\sigma_{eff}^2 = \mu\sigma_z^2$ where $0 \le \mu \le 1$ and μ depends on the pdf of the signal and the SI noise variance σ_{SI}^2.

Before demonstrating the aforementioned Bayes-optimality property of AMP-SI, we use matched filter arguments to provide a simplified representation of the conditional denoiser of (6) when the SI is the signal viewed with AWGN. In calculating the AMP-SI denoiser (6), we want to calculate the expectation of X conditioned on the pseudo data, $X+\lambda_t Z_1 = a$, and SI, $X+\sigma_{SI}Z_2 = b$, where Z_1 and Z_2 are independent, standard Gaussian RVs. We define signal and noise vectors as $s = [1\ 1]^T$ and $v = [\lambda_t Z_1\ \sigma_{SI}Z_2]^T$, respectively, where $[\cdot]^T$ is the transpose operator. The matched filter estimates the unknown X by computing the inner product between

$$\begin{bmatrix}a\\b\end{bmatrix} = \begin{bmatrix}X+\lambda_t Z_1\\ X+\sigma_{SI}Z_2\end{bmatrix} = sX + v,$$

and a matched filter $h \in \mathbb{R}^2$. An optimal h^* that maximizes the signal to noise ratio while having unit norm is computed by inverting $R_v = E[vv^T]$, the autocovariance matrix of v,

$$h^* = (R_v)^{-1}s/\|(R_v)^{-1}s\|.$$

It can be shown that $h^* = [\sigma_{SI}^2\ \lambda_t^2]^T/(\sigma_{SI}^2+\lambda_t^2)$, and the inner product is defined as $\mu^t(a,b)$:

$$\mu^t(a,b) = \langle [a\ b]^T, h^*\rangle = \frac{a\sigma_{SI}^2 + b\lambda_t^2}{\sigma_{SI}^2+\lambda_t^2}. \tag{33}$$

Note that $\mu^t(X+\lambda_t Z_1, X+\sigma Z_2)$ equals

$$\frac{(X+\lambda_t Z_1)\sigma_{SI}^2 + (X+\sigma_{SI}Z_2)\lambda_t^2}{\sigma_{SI}^2+\lambda_t^2} \overset{d}{=} X + \sigma_t Z,$$

where Z is standard Gaussian, $\overset{d}{=}$ denotes equality in distribution, and the variance term, $(\sigma_t)^2$, is

$$(\sigma_t)^2 = \frac{(\lambda_t\sigma_{SI}^2)^2 + (\sigma_{SI}\lambda_t^2)^2}{(\sigma_{SI}^2+\lambda_t^2)^2} = \frac{\lambda_t^2\sigma_{SI}^2}{\sigma_{SI}^2+\lambda_t^2}. \tag{34}$$

The above provides us with the following simplification of the AMP-SI denoiser (6) for SI with AWGN,

$$\eta_t(a,b) = \mathbb{E}[X|X+\sigma^t Z = \mu^t(a,b)], \tag{35}$$

where $\mu^t(a,b)$ and σ^t are defined in (33) and (34). We note that μ^t is a function of (a,b), but for brevity we drop this dependence in the following. Considering (9) and (35),

$$\eta_t(X+\lambda_t Z_1, X+\sigma_{SI}Z_2) = \mathbb{E}[X|X+\sigma^t Z]. \tag{36}$$

We simplify the SE equations (9) using (36) and the definition of σ^t in (34). Let $\lambda_0 = \sigma_z^2 + \mathbb{E}[X^2]/\delta$ and for $t \geq 0$,

$$\lambda_t^2 = \sigma_z^2 + \frac{1}{\delta}\mathbb{E}\left[\left(\mathbb{E}\left[X\middle|X+\sqrt{\frac{\lambda_{t-1}^2\sigma_{SI}^2}{\sigma_{SI}^2+\lambda_{t-1}^2}}\,Z\right]-X\right)^2\right]. \tag{37}$$

The results in (35) and (37) provide a simplified way to calculate the conditional denoiser of (6) and the SE *when the signal and the SI are related through Gaussian noise*. Moreover, at the stationary point of (37) we have

$$\lambda^2 = \sigma_z^2 + \frac{1}{\delta}\mathbb{E}\left[\left(\mathbb{E}\left[X\middle|X+\sqrt{\frac{\lambda^2\sigma_{SI}^2}{\sigma_{SI}^2+\lambda^2}}\,Z\right]-X\right)^2\right], \tag{38}$$

where λ^2 is the scalar channel variance. Comparing (4) (SE without SI) and (38), we denote the variance in the conditional expectation by $\widetilde{\lambda}^2 = \frac{\lambda^2\sigma_{SI}^2}{\sigma_{SI}^2+\lambda^2}$. Note that $\lambda^2 = \frac{\widetilde{\lambda}^2\sigma_{SI}^2}{\sigma_{SI}^2-\widetilde{\lambda}^2} \geq 0$, because $\widetilde{\lambda}^2 \leq \sigma_{SI}^2$, and we can rewrite the above as

$$\widetilde{\lambda}^2 = \frac{(\sigma_{SI}^2-\widetilde{\lambda}^2)\sigma_z^2}{\sigma_{SI}^2} + \frac{1}{\frac{\delta\sigma_{SI}^2}{\sigma^2-\widetilde{\lambda}^2}}\mathbb{E}\left[\left(\mathbb{E}[X|X+\widetilde{\lambda}Z]-X\right)^2\right]. \tag{39}$$

We see that AMP-SI SE (9) has fixed points coinciding with the fixed points of standard AMP SE (4) with 'effective' measurement rate $\delta_{eff} = \delta\left(\frac{\sigma_{SI}^2+\lambda^2}{\sigma_{SI}^2}\right)$ and 'effective' measurement noise variance $\sigma_{eff}^2 = \left(\frac{\sigma_{SI}^2}{\sigma_{SI}^2+\lambda^2}\right)\sigma_z^2$ where σ_{SI}^2 is the noise in the SI and λ^2 is the stationary point of (37). This effective change in δ and σ^2 implies that the incorporation of SI with AWGN via the AMP-SI algorithm gives us signal recovery for a standard (without SI) linear regression problem (1) with *more* measurements and/or *reduced* measurement noise variance than our own, and the effect becomes more pronounced, as the noise variance in the SI, σ_{SI}^2, gets small.

The above analysis relies on the fact that for the conditional expectation denoiser in standard (without SI) AMP (2)-(3), the corresponding SE equation (4) in its convergent states coincides with Tanaka's fixed point equation [37], ensuring that if AMP runs until it converges, the result provides the best possible MSE achieved by any algorithm under certain conditions. (These conditions on δ and ϵ, while outside the scope of this paper, ensure that there is a single solution to Tanaka's fixed point equation, since multiple solutions may create a disparity between the MSE of AMP and the MMSE [20], implying that AMP-SI might be sub-optimal in such cases.) However, the above analysis relies heavily on the Gaussianity of the SI noise and its generalization is left for future work.

References

[1] Minimum mean square error. https://en.wikipedia.org/wiki/Minimum_mean_square_error#cite_note-1, Retrieved July 7, 2017.

[2] H. Arguello and G. Arce. Code aperture optimization for spectrally agile compressive imaging. *J. Opt. Soc. Am.*, 28(11):2400–2413, Nov. 2011.

[3] J. Barbier, C. Schülke, and F. Krzakala. Approximate message-passing with spatially coupled structured operators, with applications to compressed sensing and sparse superposition codes. *J. Stat. Mech-Theory E.*, 2015(5):P05013, May 2015.

[4] D. Baron, A. Ma, D. Needell, C. Rush, and T. Woolf. Conditional approximate message passing with side information. In *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2017.

[5] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory*, 57(2):764–785, Feb. 2011.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, Jan. 2011.

[7] G. Caire, R. Muller, and T. Tanaka. Iterative multiuser joint decoding: Optimal power allocation and low-complexity implementation. *IEEE Trans. Inf. Theory*, 50(9):1950–1973, Sept. 2004.

[8] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9:717–772, Dec. 2009.

[9] E. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, Dec. 2006.

[10] G.-H. Chen, J. Tang, and S. Leng. Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets. *Medical Physics*, 35(2):600–663, Feb. 2008.

[11] M. Chen, F. Renna, and M. Rodrigues. On the design of linear projections for compressive sensing with side information. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 670–674, Barcelona, Spain, July 2016.

[12] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comp.*, 20(1):33–61, Aug. 1998.

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

[14] D. Donoho, A. Maleki, and A. Montanari. Message passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.*, 106(45):18914–18919, Nov. 2009.

[15] R. Gallager. *Information Theory and Reliable Communications*. Wiley, Jan. 1968.

[16] D. Guo and S. Verdú. Randomly spread CDMA: Asymptotics via statistical physics. *IEEE Trans. Inf. Theory*, 51(6):1983–2010, June 2005.

[17] C. Herzet, C. Soussen, J. Idier, and R. Gribonval. Exact recovery conditions for sparse representation with partial support information. *IEEE Trans. Inf. Theory*, 59(11):7509–7524, Aug. 2013.

[18] U. Kamilov, A. Bourquard, and M. Unser. Sparse image deconvolution with message passing. In *Proc. 5th Workshop on Signal Process. with Adaptive Sparse Structured Representations (SPARS)*, Feb. 2013.

[19] U. Kamilov, S. Rangan, A. Fletcher, and M. Unser. Approximate message passing with consistent parameter estimation and applications to sparse learning. In *Workshop Neural Info. Proc. Sys. (NIPS)*, pages 2447–2455, Dec. 2012.

[20] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices. *J. Stat. Mech. - Theory E.*, 2012(08):P08009, Aug. 2012.

[21] S. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–137, Mar. 1982.

[22] H. V. Luong, J. Seiler, A. Kaup, S. Forchhammer, and N. Deligiannis. Measurement bounds for sparse signal reconstruction with multiple side information. *Arxiv preprint arXiv:1605.03234*, Jan. 2017.

[23] Y. Ma, J. Zhu, and D. Baron. Approximate message passing algorithm with universal denoising and Gaussian mixture learning. *IEEE Trans. Signal Process.*, 65(21):5611–5622, Nov. 2016.

[24] A. Maleki. *Approximate message passing algorithms for compressed sensing*. Stanford University, Nov. 2010.

[25] H. Mansour and R. Saab. Recovery analysis for weighted ℓ_1-minimization using the null space property. *Appl. Comput. Harmon. Anal.*, 43(1):23–38, July 2017.

[26] J. Mota, N. Deligiannis, and M. Rodrigues. Compressed sensing with prior information: Strategies, geometry, and bounds. *IEEE Trans. Inf. Theory*, 63(7):4472–4496, July 2017.

[27] J. Mota, N. Deligiannis, A. Sankaranaraynan, V. Cevher, and M. Rodrigues. Adaptive-rate reconstruction of time-varying signals with application in compressive foreground extraction. *IEEE Trans. Signal Process.*, 64(14):3651–3666, Mar. 2016.

[28] D. Needell, R. Saab, and T. Woolf. Weighted-minimization for sparse recovery under arbitrary prior information. *Inst. Math. Inf. Infer.*, 6(3):284–309, Jan. 2017.

[29] S. Rangan. Generalized approximate message passing for estimation with random linear mixing. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 2168–2172, July 2011.

[30] S. Rangan, A. Fletcher, P. Schniter, and U. Kamilov. Inference for generalized linear models via alternating directions and Bethe free energy minimization. In *Proc. Int. Symp. Inf. Theory (ISIT)*, pages 1640–1644, June 2015.

[31] S. Rangan, P. Schniter, and A. Fletcher. On the convergence of approximate message passing with arbitrary matrices. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pages 236–240, Feb. 2014.

[32] F. Renna, L. Wang, X. Yuan, J. Yang, G. Reeves, A. Calderbank, L.Carin, and M. Rodrigues. Classification and reconstruction of high-dimensional signals from low-dimensional features in the presence of side information. *IEEE Trans. Inf. Theory*, 62(11):6459–6492, Sept. 2016.

[33] C. Rush and R. Venkataramanan. Finite sample analysis of approximate message passing. *IEEE Trans. Inf. Theory*, (forthcoming). Available: https://ieeexplore.ieee.org/document/8318695/.

[34] A. Saleh and R. Valenzuela. A statistical model for indoor multipath propagation. *IEEE J. Select. Areas Commun.*, 5(2):128–137, Feb. 1987.

[35] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R. Baraniuk. A new compressive imaging camera architecture using optical-domain compression. Feb. 2006.

[36] J. Tan, Y. Ma, and D. Baron. Compressive imaging via approximate message passing with image denoising. *IEEE Trans. Signal Process.*, 63(8):2085–2092, Apr. 2015.

[37] T. Tanaka. A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Trans. Inf. Theory*, 48(11):2888–2910, Nov. 2002.

[38] R. Tibshirani. Regression shrinkage and selection via the LASSO. *J. Royal Stat. Soc. Series B (Methodological)*, 58(1):267–288, Jan. 1996.

[39] N. Vaswani and W. Lu. Modified-CS: Modifying compressive sensing problems for partially known support. *IEEE Trans. Signal Process.*, 58(9):4595–4607, May 2010.

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

[41] L. Weizman, Y. Eldar, and D. Bashat. Compressed sensing for longitudinal MRI: An adaptive-weighted approach. *Medical Physics*, 42(9):5195–5208, Nov. 2015.

[42] J. Zhu, D. Baron, and A. Beirami. Optimal trade-offs in multi-processor approximate message passing. *Arxiv preprint arXiv:1601.03790*, Nov. 2016.

[43] J. Ziniel and P. Schniter. Dynamic compressive sensing of time-varying signals via approximate message passing. *IEEE Trans. Signal Process.*, 61(21):5270–5284, Nov. 2013.