Optimal Data Detection and Signal Estimation in Systems with Input Noise

Ramina Ghods, Charles Jeon, Arian Maleki, and Christoph Studer

***Abstract*—Practical systems often suffer from hardware impairments that already appear during signal generation. Despite the limiting effect of such input-noise impairments on signal processing systems, they are routinely ignored in the literature. In this paper, we propose an algorithm for data detection and signal estimation, referred to as Approximate Message Passing with Input noise (AMPI), which takes into account input-noise impairments. To demonstrate the efficacy of AMPI, we investigate two applications: Data detection in massive multiple-input multiple-output (MIMO) wireless systems and sparse signal recovery in compressive sensing. For both applications, we provide precise conditions in the large-system limit for which AMPI achieves optimal performance. We furthermore use simulations to demonstrate that AMPI achieves near-optimal performance at low complexity in realistic, finite-dimensional systems.**

***Index Terms*—Approximate message passing (AMP), compressive sensing, data detection, hardware impairments, input noise, massive MIMO systems, noise folding, sparsity, state evolution.**

I. Introduction

We consider a general class of data detection and signal estimation problems in a noisy linear channel affected by input noise. As illustrated in Fig. 1, we are interested in recovering the N-dimensional *input signal* $\mathbf{s} \in \mathbb{C}^N$ observed from the following model. The input signal $\mathbf{s}$ with prior distribution $p(\mathbf{s}) = \prod_{\ell=1}^{N} p(s_\ell)$ is affected by input-noise characterized by the statistical relation $p(\mathbf{x}|\mathbf{s}) = \prod_{\ell=1}^{N} p(x_\ell|s_\ell)$. The generality of this input-noise model captures a wide range of hardware and system impairments, including hardware non-idealities that exhibit statistical dependence between impairments and the input signal (e.g., phase noise) as well as deterministic effects (e.g., non-linearities). The impaired signal $\mathbf{x} \in \mathbb{C}^N$, which we refer to as the *effective input signal*, is then passed to a noisy linear transform that is modeled as

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

Here, the vector $\mathbf{y} \in \mathbb{C}^M$ is the *measured signal* and M denotes the number of measurements, the *system matrix* $\mathbf{H} \in \mathbb{C}^{M \times N}$ represents the measurement process, and the vector $\mathbf{n} \in \mathbb{C}^M$ models measurement noise. We assume that the entries of the noise vector $\mathbf{n}$ are i.i.d. circularly-symmetric complex Gaussian with variance N_0. In what follows, we make use of the *system ratio* defined as $\beta = N/M$ and the following definitions:

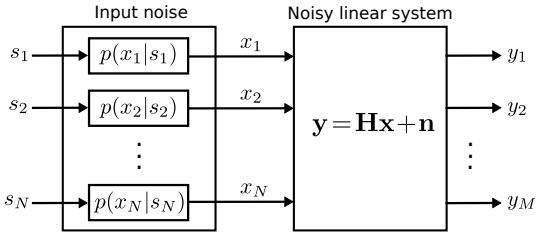


Fig. 1. Illustration of a noisy linear system affected by input noise. The input signal $\mathbf{s}$ is corrupted by input noise, resulting in the effective input signal $\mathbf{x}$ that is observed through a noisy linear system. The goal is to recover the input signal $\mathbf{s}$ from the noisy observations $\mathbf{y}$.

Definition 1. *We define the* large system limit *by fixing the system ratio $\beta = N/M$ and by letting $N \to \infty$.*

Note that N and M can depend on each other as long as their ratio N/M converges to the fixed value β.

Definition 2. *A matrix $\mathbf{H}$ describes* uniform linear measurements *if the entries of $\mathbf{H}$ are i.i.d. circularly-symmetric complex Gaussian with variance $1/M$, i.e., $H_{i,j} \sim \mathcal{CN}(0, 1/M)$.*

Examples of applications that use Definition 2 include Rayleigh fading in wireless communication [2], multi-user communication with randomly-spread code-division multiple access (CDMA) [3], and measurement matrices in compressive sensing that satisfy restricted isometry property with high probability [4]. Definition 2 is frequently used in our theoretical performance analysis of AMPI. We will assume that the system matrix $\mathbf{H}$ is known for our application examples.

A. Two Application Examples

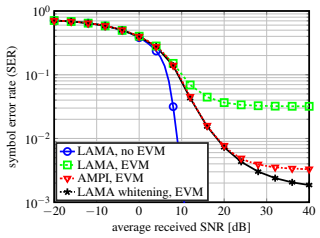
While numerous real-world systems suffer from input noise, we focus on two prominent scenarios.

1) Massive MIMO Data Detection: Massive multiple-input multiple-output (MIMO) is one of the core technologies in fifth-generation (5G) wireless systems [5]. The idea is to equip the infrastructure base-stations with hundreds of antenna elements while simultaneously serving a smaller number of users. One critical challenge in the realization of this technology is the computational complexity of data detection at the base-station [6]. While recent results have shown that the large dimensionality of massive MIMO can be exploited to design near-optimal data-detection algorithms [7]–[9] using approximate message passing (AMP) [10], these

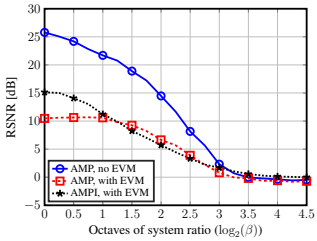
R. Ghods, C. Jeon and C. Studer were with the School of ECE, Cornell University, Ithaca, NY; e-mail: RG is now with Carnegie Mellon University, Pittsburgh, PA; rghods@cs.cmu.edu; CJ is with Apple Inc., San Diego, CA; CS is with ETH Zurich, Zurich, Switzerland; studer@ethz.ch.

A. Maleki is with Department of Statistics at Columbia University, New York City, NY; arian@stat.columbia.edu.

Part of this paper on massive MIMO detection has been presented at the 53rd Annual Allerton Conference on Communication, Control, and Computing [1]. The present paper provides algorithm details that were missing in [1] and includes more theoretical results for AMPI. In addition, we apply the proposed AMPI framework to sparse signal recovery in compressive sensing.



(a) Symbol-error rate (SER) versus average SNR in a 128 user equipment (UE), 128 base-station antenna massive MIMO system with QPSK.



(b) Reconstruction SNR of a sparse signal recovery task for a 5% sparse signal of dimension $N = 1000$ and 20 dB SNR.

Fig. 2. Simulation results of two applications of the proposed AMPI algorithm in massive MIMO and compressive sensing with $EVM = -10$ dB input noise. AMPI yields significant improvements compared to methods that ignore input noise and achieves comparable performance to whitening-based methods that entail prohibitive complexity for the considered system dimensions.

methods ignore the fact that realistic communication systems are affected by impairments that already arise at the transmit side [11], [12]. In this paper, we introduce AMPI (short for AMP with input noise), which mitigates transmit-side RF impairments during data detection. AMPI outperforms existing data-detection methods, e.g., the LAMA algorithm from [8], that ignore the presence of input-noise impairments at virtually no additional computational complexity.

Fig. 2(a) illustrates the symbol error-rate (SER) performance of AMPI in a symmetric massive MIMO system with 128 user equipments (UEs) transmitting QPSK and 128 base-station (BS) antennas. As in [11]–[13], the input noise is modeled as complex Gaussian noise. We consider an additive input noise model $\mathbf{x} = \mathbf{s} + \mathbf{e}$, where $\mathbf{s}$ is the input signal and $\mathbf{e}$ models input noise, with an error vector magnitude (EVM) of $EVM = \mathbb{E}\big[\|\mathbf{e}\|^2\big]/\mathbb{E}\big[\|\mathbf{s}\|^2\big] = -10\,\text{dB}$. The blue curve corresponds to the performance in absence of input noise (no EVM) using LAMA algorithm [7], [8], which achieves—under certain conditions on the MIMO system—the error-rate performance of the individually-optimal (IO) data detector in absence of input noise. After considering input noise, LAMA's performance is drastically reduced (green curve). With input noise, AMPI improves this performance significantly at virtually no complexity increase compared to LAMA. AMPI also achieves comparable SER to a whitening-based approach [11], which is optimal for a Gaussian input noise model; however, noise whitening results in prohibitively high computational complexity in massive MIMO systems.

2) Compressive Sensing Signal Recovery: Compressive sensing (CS) enables sampling and recovery of sparse signals at sub-Nyquist rates [14], [15]. While the CS literature extensively focuses on systems with measurement noise, numerous practical applications already contain noise on the sparse signal to be recovered; see [16]–[18] and the references therein. We will use AMPI to take input-noise into account directly during sparse signal recovery and show substantial improvements compared to that of existing sparse recovery methods for systems with input noise [18] at no additional expense in complexity.

Fig. 2(b) shows the recovery signal-to-noise-ratio (RSNR) defined as $RSNR = \mathbb{E}\big[\|\mathbf{s}\|_2^2/\|\hat{\mathbf{s}} - \mathbf{s}\|_2^2\big]$ where $\mathbf{s}$ and $\hat{\mathbf{s}}$ are the true signal and recovered, respectively, for a compressive sensing scenario in which we recover a $N = 1000$ dimensional sparse signal in the presence of additive input noise with an EVM of -10 dB. The RSNR is plotted for different system ratios β. The signal is assumed to have 5% sparsity and an SNR of $SNR = \mathbb{E}\big[\|\mathbf{Hs}\|^2\big]/\mathbb{E}\big[\|\mathbf{n}\|^2\big] = 20\,\text{dB}$; the non-zero entries are i.i.d. standard normal. The blue curve corresponds to the performance of AMP in the absence of input noise (no EVM). Considering input noise drastically reduces the RSNR of AMP (red curve) whereas AMPI significantly improves the RSNR over AMP for small system ratios β (black curve) at virtually no additional computational complexity.

B. Contributions

In this paper, we consider the recovery of signals in the presence of input noise. We propose a general and computationally-efficient framework called AMPI, which takes into account the effects of input noise on data detection and signal estimation applications. We first provide an asymptotic analysis that characterizes the performance of AMPI in a general AMP-based framework with wide applicability. We then specialize AMPI for two relevant applications: (i) massive MIMO data detection and (ii) sparse signal recovery from compressive sensing measurements. For these applications, we provide a theoretical optimality analysis with the following optimality criteria. For massive MIMO data detection, optimality is achieving the same error-rate performance as the individually-optimal (IO) data detector [8], [19], which solves the minimization problems

$$\hat{s}_\ell^{\text{IO}} = \underset{\tilde{s}_\ell \in \mathcal{O}}{\arg\min}\, \mathbb{P}(\tilde{s}_\ell \neq s_\ell), \quad \ell = 1, \ldots, N. \tag{2}$$

Here, $\mathbb{P}$ stands for probability and $\mathcal{O}$ is a finite set containing possible transmit symbols—in wireless systems this set corresponds to the transmit constellation, e.g., quadrature amplitude modulation (QAM). Part of this analysis generalizes our results from [7], [8] to systems that are affected by input noise and provides precise conditions on the system ratio β for which AMPI is able to achieve IO performance. For signal estimation, optimality is achieved by minimizing the following mean-squared error (MSE):

$$\hat{\mathbf{s}}^{\text{O}} = \underset{\tilde{\mathbf{s}} \in \mathbb{C}^N}{\arg\min}\, \tfrac{1}{N} \|\tilde{\mathbf{s}} - \mathbf{s}\|^2. \tag{3}$$

Here, the superscript O in $\hat{s}_\ell^{\text{O}}$, $\ell = 1, \ldots, N$, stands for optimal. To solve (2) or (3), we need to compute the MAP or minimum MSE (MMSE) estimate of the marginal posterior distribution $p(s_\ell|\mathbf{y}, \mathbf{H})$ for all $\ell = 1, \ldots, N$. Computing the marginal distribution for large-dimensional systems is one of the key challenges in data detection and signal estimation problems as its requires prohibitive complexity [20]. We propose AMPI, which achieves optimal performance in the large-system limit and for uniform linear measurements. Our optimality conditions are derived via the state-evolution (SE) framework [21], [22] of approximate message passing (AMP) [23]–[25]. For both applications, we demonstrate the efficacy and low-complexity of AMPI in more realistic, finite-dimensional systems.

C. Related Results

The effect of input-noise (often called transmit-side impairments) on the performance of communication systems has been studied in [11]–[13], [26]–[34]. Most of these papers use a Gaussian input-noise model, which assumes that the input noise is i.i.d. additive Gaussian noise and independent of the transmit signal $\mathbf{s}$. While the accuracy of this model has been confirmed via real-world measurements [11] for MIMO systems that use orthogonal frequency-division multiplexing (OFDM), it may not be accurate in other scenarios. AMPI is a practical data detection method that allows us to study the fundamental performance of more general input-noise models, which may exhibit statistical dependence with the transmit signal and even include deterministic nonlinearities. For the well-established Gaussian transmit-noise model, we will show in Section III that the SE equations of AMPI coincide to the "coupled fixed point equations" provided in [12], which demonstrates that AMPI is a practical algorithm that achieves the performance predicted by replica-based channel capacity expressions.

In the compressive sensing literature, input noise causes an effect known as "noise folding" [16], [18], [35]. Reference [18] shows that in the large-system limit, the received SNR is increased by a factor of N/M due to input noise. Reference [16] shows that an oracle-based recovery procedure that knows the signal support exhibits a 3 dB loss of reconstructed SNR per octave of sub-sampling. More recently, reference [35] has introduced an ℓ_1-norm based algorithm that reduces the effect of input noise. In contrast to these results, AMPI is a practical signal estimation method that enables one to study the fundamental performance limits of sparse signal recovery in the presence of input noise. We also note that reference [36] investigates signal recovery for a similar model as in (1). The key difference is that AMPI computes an estimate of the original input signal $\mathbf{s}$, whereas the generalized AMP (GAMP) algorithm in [36] recovers the effective input signal $\mathbf{x}$.

D. Notation

Lowercase and uppercase boldface letters represent column vectors and matrices, respectively. For a matrix $\mathbf{H}$, the conjugate transpose is $\mathbf{H}^{\mathsf{H}}$. The entry on the ith row and jth column of the matrix $\mathbf{H}$ is $H_{i,j}$; the kth entry of a vector $\mathbf{x}$ is x_k. The $M \times M$ identity matrix is denoted by $\mathbf{I}_M$ and the $M \times N$ all-zeros matrix by $\mathbf{0}_{M\times N}$. We define $\langle \mathbf{x} \rangle = \frac{1}{N}\sum_{k=1}^{N} x_k$. The quantities $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_2$ represent the ℓ_1 and ℓ_2 norms of the vector $\mathbf{x}$, respectively. Multivariate real-valued and complex-valued Gaussian probability density functions (pdfs) are denoted by $\mathcal{N}(\mathbf{m}, \mathbf{K})$ and $\mathcal{CN}(\mathbf{m}, \mathbf{K})$, respectively, where $\mathbf{m}$ is the mean vector and $\mathbf{K}$ the covariance matrix. The operator $\mathbb{E}_X[\cdot]$ denotes expectation with respect to the probability density function (PDF) of the random variable (RV) X; $p(x)$ represents the probability distribution of RV x, and $\mathbb{P}(A)$ shows probability of event A. The notation $a \xrightarrow{d} b$ represents convergence in distribution of $p(a)$ to $p(b)$. The function $1(\cdot)$ returns 1 if its argument is true and 0 otherwise.

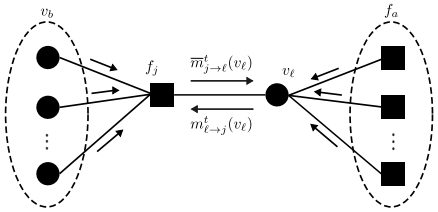


Fig. 3. A factor graph illustrating the sum-product message-passing algorithm.

E. Paper Outline

The rest of the paper is organized as follows. Section II introduces the AMPI algorithm along with its state evolution (SE) framework. Section III analyzes optimality conditions for AMPI. Section IV and Section V investigate AMPI for data detection in massive MIMO systems and for sparse signal recovery, respectively. We conclude in Section VI.

II. AMPI: Approximate Message Passing with Input Noise

We now introduce the message passing algorithm used to derive AMPI. We then develop the complex-valued state-evolution (cSE) framework for AMPI, which will be used in Section III to establish optimality conditions.

A. Sum-Product Message Passing

As discussed in Section I-B, the most challenging step in data detection and signal estimation is calculating the marginal posterior distribution. While the problem of marginalizing a distribution is in general NP-hard [20], there exist heuristics that have been successful in certain applications—one of the most prominent marginalizing schemes is message passing.

Sum-product message passing is a well-established method to compute the marginal distributions [37], [38]. Consider a joint probability distribution $p(v_1, \ldots, v_I)$ with random variables taken from the set $\{v_1, \ldots, v_I\}$. Suppose that $p(v_1, \ldots, v_I)$ factors into a product of J functions with subsets of the variable set $\{v_1, \ldots, v_I\}$ as their argument: $p(v_1, \ldots, v_I) = \prod_{j=1}^{J} f_j(V_j)$, where each V_j is a subset of the variable set $\{v_1, \ldots, v_I\}$ and $\cup_{j=1}^{J} V_j = \{v_1, \ldots, v_n\}$. Such distributions can be expressed as a factor graphs, which are bipartite graphs consisting of variable nodes to represent each random variable v_ℓ, factor nodes for each factor f_j, and edges connecting them if the factor node is an argument of the variable node. Fig. 3 illustrates a factor graph with variable nodes (circles) and factor nodes (squares).

For the sum-product message passing algorithm, we consider messages $m^t_{\ell\to j}(v_i)$ (from every variable node v_i to every factor node f_j) and $\overline{m}^t_{j\to i}(v_i)$ (from every factor node f_j to every variable node v_i) at iteration $t = 1, \ldots, t_{\text{max}}$ with the equations

$$m^t_{i\to j}(v_i) = \prod_{a\neq j} \overline{m}^{t-1}_{a\to i}(v_i) \tag{4}$$

$$\overline{m}^t_{j\to i}(v_i) = \int_{\mathbb{C}} f_j(\partial f_j) \prod_{b\neq i} m^t_{b\to j}(v_b)\, \mathrm{d}(\partial f_j \neq v_i). \tag{5}$$

Here, t_{max} is the maximum number of iterations and ∂f_j is the set of neighbors of node f_j in the graph. After iteratively passing messages between variable nodes and factor nodes, the marginal function of the random variable v_i is approximated by

the product of all messages directed toward that variable node. If a factor graph is cycle-free, then message-passing converges to the exact marginals. If the factor graph has cycles, then general convergence conditions are unknown [39].

B. Deriving Message Passing for our System Model

Consider the system model in Section I. We are interested in recovering the input signal $\mathbf{s}$ by computing the MAP or MMSE estimate of the marginal posterior distributions $p(s_\ell|\mathbf{y},\mathbf{H})$. The marginal distribution $p(s_\ell|\mathbf{y},\mathbf{H})$ can be derived from the joint probability distribution $p(\mathbf{s},\mathbf{x},\mathbf{y}|\mathbf{H})$ as follows:

$$p(s_\ell|\mathbf{y},\mathbf{H}) = \int_{\mathbb{C}^{N-1}} p(\mathbf{s}|\mathbf{y},\mathbf{H})\ \mathrm{d}(s_1,\ldots,s_N \neq s_\ell) \tag{6}$$

$$\propto \int_{\mathbb{C}^{N-1}} \left(\int_{\mathbb{C}^N} p(\mathbf{s},\mathbf{x},\mathbf{y}|\mathbf{H})\mathrm{d}\mathbf{x} \right) \mathrm{d}(s_1,\ldots,s_N \neq s_\ell). \tag{7}$$

Here, the notation $\mathrm{d}(s_1,\ldots,s_N \neq s_\ell)$ indicates integration over all entries $s_1,\ldots,s_N$ except for s_ℓ. Instead of an exhaustive integration for each entry s_ℓ, $\ell = 1,\ldots,N$, we perform sum-product message passing on the factor graph given by the joint PDF $p(\mathbf{s},\mathbf{x},\mathbf{y}|\mathbf{H}) = p(\mathbf{y}|\mathbf{x},\mathbf{H})p(\mathbf{x}|\mathbf{s})p(\mathbf{s})$. The factor graph for this distribution is illustrated in Fig. 4 and consists of the factors $p(\mathbf{y}|\mathbf{x},\mathbf{H})$, $p(\mathbf{x}|\mathbf{s})$, and $p(\mathbf{s})$. For any specific application, we will either know the probability distribution for these factors or make reasonable assumptions about them (see Section IV and Section V for concrete examples). The following observations allow us to simplify message passing: (i) The message from variable node s_ℓ to the factor node $p(x_\ell|s_\ell)$ is equal to $p(s_\ell)$ and remains constant over all iterations. (ii) The message from factor node $p(x_\ell|s_\ell)$ to variable node x_ℓ is $\int_{\mathbb{C}} p(x_\ell|s_\ell)p(s_\ell)\mathrm{d}s_\ell$. From these observations, we see that the factor graph can be divided into two parts. Furthermore, as shown in Fig. 4, let $v^t_{a\to\ell}(x_\ell)$ denote the message from the factor node $p(y_a|\mathbf{x})$ to variable node x_ℓ, and $\overline{v}^t_{\ell\to a}(x_\ell)$ the message from variable node x_ℓ to factor node $p(y_a|\mathbf{x})$. To calculate the messages $v^t_{a\to\ell}(x_\ell)$ and $\overline{v}^t_{\ell\to a}(x_\ell)$ on the right side of the graph (marked with a dashed box in Fig. 4), we can ignore the left part of the graph (on the left of the variable nodes x_ℓ) and identify them as messages $p(x_\ell)$ connected to x_ℓ that are computed by

$$p(x_\ell) = \int_{\mathbb{C}} p(x_\ell|s_\ell)p(s_\ell)\mathrm{d}s_\ell. \tag{8}$$

This implies that we can perform message passing on the effective system model $\mathbf{y} = \mathbf{H}\mathbf{x}+\mathbf{n}$ with effective input signal prior $p(\mathbf{x})$ given in (8). Since we are interested in the estimate of the marginal distribution $p(s_\ell|\mathbf{y},\mathbf{H})$, we can return to the left side of the factor graph and calculate the corresponding messages once $v^t_{a\to\ell}(x_\ell)$ and $\overline{v}^t_{\ell\to a}(x_\ell)$ have been computed. Then, the estimated marginal distribution of $p(s_\ell|\mathbf{y},\mathbf{H})$ is

$$\hat{p}(s_\ell|\mathbf{y},\mathbf{H}) = \int_{x_\ell} \prod_{b=1}^{N} \overline{v}_{b\to\ell}(x_\ell)p(x_\ell|s_\ell)p(s_\ell)dx_\ell, \tag{9}$$

where we use the notation $\hat{p}(s_\ell|\mathbf{y},\mathbf{H})$ to emphasize that this marginalization is an estimate. In the next section, we will provide conditions for which this estimate is exact.

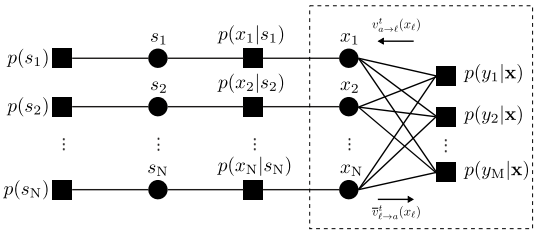


Fig. 4. The factor graph of the joint distribution $p(\mathbf{s},\mathbf{x},\mathbf{y}|\mathbf{H})$. Performing sum-product message passing on this factor graph yields the marginal posterior distributions $p(s_\ell|\mathbf{y},\mathbf{H})$, $\ell = 1,\ldots,N$.

Note that (9) is a one-dimensional integral that can either be evaluated in closed form or via numerical integration as long as the distribution of $\prod_{b=1}^{N} \overline{v}_{b\to\ell}(x_\ell)$ is known. To compute the messages $\overline{v}_{b\to\ell}(x_\ell)$, we perform message passing on the right side of the factor graph in Fig. 4. However, an exact message passing algorithm can quickly result in high complexity; this is mainly because the right side of factor graph in Fig. 4 is fully connected and we need to compute $2MN$ messages in each iteration. To reduce complexity, we use AMP introduced in [21], [23], [38]. AMP uses the bipartite structure of the graph and the high dimensionality of the problem to approximate the messages with Gaussian distributions. Passing Gaussian messages only requires the message mean and variance instead of PDFs. Furthermore, the structure and dimensionality also allows AMP to approximate all the messages emerging from or going toward one factor node. Specifically, [23], [38] show that all the messages emerging from one factor node have approximately the same value—similarly, all messages going toward the same factor node share approximately the same value. Hence. $v^t_{a\to\ell} \approx v^t_a, \forall\ell = 1,\ldots,N$ and $\overline{v}^t_{i\to a} \approx \overline{v}^t_i, \forall i = 1,\ldots,M$. This key observation reduces the number of messages that need to be computed in each iteration from $2MN$ to $M+N$. Reference [7] performs approximate message passing on the system model $\mathbf{y} = \mathbf{H}\mathbf{x}+\mathbf{n}$ with complex entries called cB-AMP (short for complex Bayesian AMP), which calculates an estimate for the effective input signal $\hat{x}_\ell$, $\forall\ell$ using the following algorithm.

Algorithm 1 (cB-AMP). *Initialize* $\hat{x}^1_\ell = \mathbb{E}_X[X]$, $\mathbf{r}^1 = \mathbf{y}$, *and* $\gamma_1^2 = N_0 + \beta\,\mathbb{V}\mathrm{ar}_X[X]$ *with* $X \sim p(x_\ell)$ *as defined in (8). For* $t = 1,\ldots,t_{\max}$, *compute*

$$\mathbf{z}^t = \hat{\mathbf{x}}^t + \mathbf{H}^{\mathrm{H}}\mathbf{r}^t \tag{10}$$

$$\hat{\mathbf{x}}^{t+1} = \mathsf{F}(\mathbf{z}^t,\gamma_t^2)$$

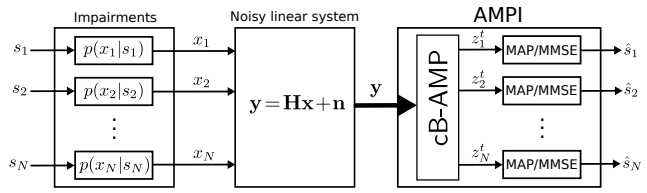
$$\gamma^2_{t+1} = N_0 + \beta\big\langle \mathsf{G}(\mathbf{z}^t,\gamma_t^2)\big\rangle \tag{11}$$

$$\mathbf{r}^{t+1} = \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{t+1} + \beta\tfrac{\mathbf{r}^t}{\gamma_t^2}\left\langle \mathsf{G}(\mathbf{z}^t,\gamma_t^2)\right\rangle.$$

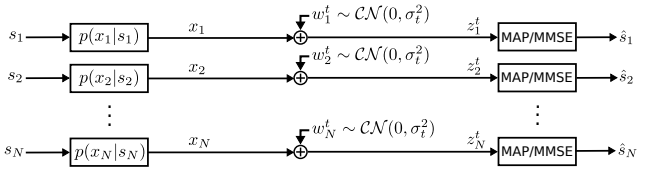
The scalar functions $\mathsf{F}(z^t_\ell,\sigma_t^2)$ *and* $\mathsf{G}(z^t_\ell,\sigma_t^2)$ *operate element-wise on vectors, correspond to the posterior mean and variance, respectively, and are defined as follows:*

$$\mathsf{F}(z^t_\ell,\sigma_t^2) = \int_{\mathbb{C}} x_\ell p(x_\ell|z^t_\ell,\sigma_t^2)\mathrm{d}x_\ell, \tag{12}$$

$$\mathsf{G}(z^t_\ell,\sigma_t^2) = \int_{\mathbb{C}} |x_\ell|^2 p(x_\ell|z^t_\ell,\sigma_t^2)\mathrm{d}x_\ell - \left|\mathsf{F}(z^t_\ell,\sigma_t^2)\right|^2. \tag{13}$$



(a) Impaired linear system with AMPI as the estimator.



(b) Equivalent decoupled system.

Fig. 5. In the large-system limit, AMPI decouples the impaired system into N parallel and independent AWGN channels, which allows us to perform impairment-aware MAP data detection or MMSE estimation.

The message posterior distribution is $p(x_\ell|z_\ell^t,\sigma_t^2) = \frac{1}{Z}p(z_\ell^t|x_\ell,\sigma_t^2)p(x_\ell)$*, where* $p(z_\ell^t|x_\ell,\sigma_t^2) \sim \mathcal{CN}(x_\ell,\sigma_t^2)$ *and* Z *is a normalization constant.*

As detailed in [7, Sec. III.C], by performing Algorithm 1, the so-called Gaussian output $\mathbf{z}^t$ of cB-AMP at iteration t in (10) can be modelled in the large system limit as

$$z_\ell^t = x_\ell + w_\ell^t, \tag{14}$$

with $w_\ell^t \sim \mathcal{CN}(0,\sigma_t^2)$, being independent from x_ℓ (see [22, Sec. 6.4] for details in the real domain). This property is known as the decoupling property as cB-AMP effectively decouples the system into a set of N parallel and independent additive white Gaussian noise (AWGN) channels. Here, σ_t^2 is the effective noise variance that can be computed using state evolution, which we will introduce in Section II-D. While the quantity σ_t^2 cannot be tracked directly within cB-AMP, it can be estimated using the threshold parameter γ_t^2 in (11) as shown in [38]. Algorithm 1 and its Gaussian output $\mathbf{z}^t$ are the results of performing AMP on the right side of the factor graph in Fig. 4. Next, we will use $\mathbf{z}^t$ to perform sum-product message passing on the left side of this factor graph.

C. AMP with Input Noise (AMPI)

We now introduce AMPI, the two-step procedure to recover the input signal $\mathbf{s}$ from the input-output relation illustrated in Fig. 1. First, as illustrated in Fig. 5(a), we use cB-AMP in Algorithm 1 on the effective system model (1) to compute the Gaussian output $\mathbf{z}^t$ and the effective noise variance σ_t^2 at iteration t, where the Gaussian output $\mathbf{z}^t$ is modelled as in (14). Fig. 5(b) shows the equivalent input-output relation of the decoupled system. As detailed in the previous section, this is the result of running AMP on the right side of the factor graph in Fig. 4. Second, we use sum-product message passing on the left side of factor graph in Fig. 4 to compute the estimated marginal distribution of $p(s_\ell|\mathbf{y},\mathbf{H})$. To compute the marginal, we use the Gaussian output $\mathbf{z}^t$ in (14), i.e., $p(z_\ell^t|x_\ell^t) \sim \mathcal{CN}(x_\ell^t,\sigma_t^2)$; this allows us to compute the marginal posterior distribution for each input signal element s_ℓ as follows:

$$\begin{aligned} p(s_\ell|\mathbf{y},\mathbf{H}) &= p(s_\ell|z_\ell^t) \propto p(s_\ell)p(z_\ell^t|s_\ell) \\ &= p(s_\ell)\int_{\mathbb{C}} p(z_\ell^t|x_\ell^t)p(x_\ell^t|s_\ell)\mathrm{d}x_\ell^t. \end{aligned} \tag{15}$$

Finally, we can compute individually optimal MAP data detection or MMSE signal estimation for each entry s_ℓ, $\ell = 1,\ldots,N$, independently using the marginal distribution. Note that (15) can be obtained from (9) by computing $\prod_{b=1}^N \overline{v}_{b\to\ell}$. As shown in [23], we have $\prod_{b=1}^N \overline{v}_{b\to\ell} \sim \mathcal{CN}(z_\ell^t,\sigma_t^2)$. Even though this approach appears to be more straightforward, it lacks the two-step intuition behind AMPI. The resulting AMPI algorithm is summarized as follows.

Algorithm 2 (AMPI). *Initialize* $\hat{x}_\ell^1 = \mathbb{E}_X[X]$, $\mathbf{r}^1 = \mathbf{y}$, *and* $\gamma_1^2 = N_0 + \beta\,\mathbb{V}\mathrm{ar}_X[X]$ *with* $X \sim p(x_\ell)$ *as in (8).*

1) *Run cB-AMP as in Algorithm 1 for* $t_{\max}$ *iterations.*
2) *Compute* $\hat{s}_\ell^{t_{\max}} = D(z_\ell^{t_{\max}},\sigma_{t_{\max}}^2)$*, where the function* D *either computes the MAP or MMSE estimate of* s_ℓ *using the posterior PDF* $p(s_\ell|z_\ell^{t_{\max}})$ *defined in (15). The effective noise variance* $\sigma_{t_{\max}}^2$ *is estimated using the threshold parameter* $\gamma_{t_{\max}}^2$ *from cB-AMP.*

As shown in Sections IV and V, the function D is chosen to satisfy the optimality conditions in (2) or (3).

D. Theoretical Analysis of AMPI via State Evolution

Analyzing message passing methods operating on dense graphs is generally difficult. However, the normality of the messages in our application enables us to study theoretical properties in the large-system limit and for uniform linear measurements. As detailed in [38], the effective noise variance σ_t^2 of AMP can be calculated analytically for every iteration $t = 1,2,\ldots,t_{\max}$, using the state evolution recursion.The following theorem repeats the complex state evolution (cSE) for complex AMP (cB-AMP) [7]. In Section III, we will use the cSE framework to derive optimality conditions for AMPI.

Theorem 1. *Assume the model in (1) with uniform linear measurements. Run cB-AMP using the function* F*, where* F *is a pseudo-Lipschitz function [40, Sec. 1.1, Eq. 1.5]. Then, in the large-system limit the effective noise variance* σ_{t+1}^2 *of cB-AMP at iteration* t *is given by the following cSE recursion:*

$$\sigma_{t+1}^2 = N_0 + \beta\Psi(\sigma_t^2,\sigma_t^2), \tag{16}$$

Here, the MSE function Ψ *is defined by*

$$\Psi(\sigma_t^2,\gamma_t^2) = \mathbb{E}_{X,Z}\Big[\big|\mathsf{F}\big(X+\sigma_t Z,\gamma_t^2\big) - X\big|^2\Big] \tag{17}$$

with $X \sim p(x_\ell)$*,* $Z \sim \mathcal{CN}(0,1)$*, and* F *and* G *are the posterior mean and variance functions from Algorithm 1. The cSE recursion is initialized at* $t = 1$ *by* $\sigma_1^2 = N_0 + \beta\,\mathbb{V}\mathrm{ar}_X[X]$*.*

Remark 1. *The posterior mean function* F *and the MSE function* $\Psi(\sigma_t^2)$ *in (17) depend on the effective input signal prior* $p(\mathbf{x})$*, which, as shown in (8), is a function of the input signal prior* $p(\mathbf{s})$ *and the conditional probability* $p(\mathbf{x}|\mathbf{s})$ *that*

models the transmit-side impairments. Furthermore, Theorem 1 assumes perfect knowledge of the noise variance N_0; [7, Thm. 1] analyzes the case of a mismatch in the noise variance.

III. Optimality of AMPI

We now analyze the optimality of AMPI for the model introduced in Section I using the cSE framework.

A. Optimality of AMPI Within the AMP Framework

In Section II, we have derived AMPI using message-passing. However, there exists a broader class of algorithms for the same task. Specifically, our version of AMPI performs sum-product message passing using the posterior mean function F as defined in (12). One can potentially modify F (or even use different functions at different iterations) to obtain estimates $\hat{x}_\ell$, $\ell = 1, \ldots, N$, and perform MAP data detection or MMSE estimation on these estimates. Such algorithms can still be analyzed through the state evolution framework. The optimality question we ask here is whether it is possible to improve AMPI by choosing functions different to the ones introduced in (2). As we will show in Theorem 2, the functions we used in first and second step of AMPI algorithm are indeed optimal.

Suppose we run AMPI for $t_{\max}$ iterations. Consider a generalization of AMPI, where, in the first step, the posterior mean function F in (12) is replaced with a general pseudo-Lipschitz function F_t that may depend on the iteration index t; the MAP or MMSE function in the second is replaced with another function $\mathsf{F}_{t_{\max}+1}$. More specifically, we consider

$$\hat{\mathbf{x}}^{t+1} = \mathsf{F}_t(\mathbf{z}^t, \gamma_t^2),\ t = 1, \ldots, t_{\max} \tag{18}$$

$$\hat{\mathbf{s}} = \mathsf{F}_{t_{\max}+1}(\mathbf{z}^{t_{\max}+1}, \gamma_{t_{\max}+1}^2),\ \mathbf{z}^{t_{\max}+1} = \hat{\mathbf{x}}^{t_{\max}+1} + \mathbf{H}^{\mathrm{H}}\mathbf{r}^{t_{\max}+1}. \tag{19}$$

We require the sequence of functions $\{\mathsf{F}_1, \mathsf{F}_2, \ldots, \mathsf{F}_{t_{\max}+1}\}$ so that Theorem 1 holds. Now, the question is whether there exists a sequence of functions $\{\mathsf{F}_1, \mathsf{F}_2, \ldots, \mathsf{F}_{t_{\max}+1}\}$, such that given the application, the resulting algorithm achieves lower probability of error or lower MSE than AMPI. Theorem 2 shows that if the solution to the fixed-point equation of (16) is unique, then AMPI is optimal within AMP framework.

The fixed-point equation of (16) is computed by letting the number of iterations $t_{\max} \to \infty$ in (16), which yields

$$\sigma^2 = N_0 + \beta\Psi(\sigma^2, \sigma^2). \tag{20}$$

Theorem 2. *Assume the system model in Section I with uniform linear measurements and the large-system limit. Suppose that we use AMPI with an arbitrary set of pseudo-Lipschitz functions $\mathsf{F}_1, \ldots, \mathsf{F}_{t_{\max}+1}$ as described in (18). If the solution to the fixed-point equation in (20) is unique, then the choice of $\mathsf{F}_1, \ldots, \mathsf{F}_{t_{\max}+1}$ that achieves optimal performance according to (2) and (3) are as introduced in Algorithm 2.*

Theorem 2 shows that it is impossible to improve upon the original choice of AMPI. The proof is given in Appendix A. The fixed-point equation (20) can in general have one or more fixed points. If it has more than one fixed point, then AMPI may converge to different solutions, depending on the initialization [41]. As it is clear from the proof of Theorem 2 in Appendix A, even in cases where AMPI does not have a unique fixed point, one of its fixed points corresponds to the optimal solution in AMP framework. Hence, to provide precise conditions for optimality of AMPI, will analyze the fixed point equation (20) for a unique solution only. To establish conditions under which the fixed point equation (20) has a unique solution, we use the following quantities from [8, Defs. 1–4].

Definition 3. *Fix the input signal prior $p(\mathbf{s})$ and input noise distribution $p(\mathbf{x}|\mathbf{s})$. Then, the exact recovery threshold (ERT) $\beta^{\max}$ and the minimum recovery threshold (MRT) $\beta^{\min}$ are*

$$\beta^{\max} = \min_{\sigma^2>0}\left\{\left(\frac{\Psi(\sigma^2,\sigma^2)}{\sigma^2}\right)^{-1}\right\},\ \beta^{\min} = \min_{\sigma^2>0}\left\{\left(\frac{\mathrm{d}\Psi(\sigma^2,\sigma^2)}{\mathrm{d}\sigma^2}\right)^{-1}\right\}. \tag{21}$$

The minimum critical noise $N_0^{\min}(\beta)$ is defined as

$$N_0^{\min}(\beta) = \min_{\sigma^2>0}\left\{\sigma^2 - \beta\Psi(\sigma^2,\sigma^2) : \beta\tfrac{\mathrm{d}\Psi(\sigma^2,\sigma^2)}{\mathrm{d}\sigma^2} = 1\right\}, \tag{22}$$

and the maximum guaranteed noise $N_0^{\max}(\beta)$ is defined as

$$N_0^{\max}(\beta) = \max_{\sigma^2>0}\left\{\sigma^2 - \beta\Psi(\sigma^2,\sigma^2) : \beta\tfrac{\mathrm{d}\Psi(\sigma^2,\sigma^2)}{\mathrm{d}\sigma^2} = 1\right\}. \tag{23}$$

Using Definition 3, the following theorem establishes three regimes for which fixed-point equation (20) has an unique solution. The proof follows from [7, Sec. IV-D, IV-E].

Lemma 3. *Let the assumptions of Theorem 1 hold and $t_{\max} \to \infty$. Fix $p(\mathbf{s})$ and $p(\mathbf{x}|\mathbf{s})$. If the variance of the receive noise N_0 and system ratio β are in one of the following three regimes:*

1) *$\beta \in \left(0, \beta^{\min}\right]$ and $N_0 \in \mathbb{R}^+$*
2) *$\beta \in \left(\beta^{\min}, \beta^{\max}\right)$ and $N_0 \in \left[0, N_0^{\min}(\beta)\right) \cup (N_0^{\max}(\beta), \infty)$*
3) *$\beta \in [\beta^{\max}, \infty)$ and $N_0 \in (N_0^{\max}(\beta), \infty)$,*

then the fixed point equation (20) has a unique solution.

For AMPI, the quantities in Definition 3 depend on $p(\mathbf{x})$, which is a function of $p(\mathbf{s})$ and $p(\mathbf{x}|\mathbf{s})$ (cf. Remark 1). These quantities can be computed either numerically or in closed-form (see [8, Sec. III]). In many applications, the effective input signal prior $p(\mathbf{x})$ is continuous and bounded which results in certain properties for ERT and MRT as discussed in the following lemma. The proof is given in Appendix B.

Lemma 4. *Suppose the probability density $p(\mathbf{x})$ of the effective input signal $\mathbf{x}$ is continuous and bounded. Furthermore, let the assumptions made in Theorem 1 hold. Then, the ERT and MRT satisfy $\beta^{\max} = 1$ and $\beta^{\min} \leq 1$.*

From Lemma 4, we conclude that for a system with a continuous effective input signal prior $p(\mathbf{x})$ we have $\beta^{\min} \leq 1$. As noted in Lemma 3, $\beta^{\min}$ determines the values of system ratio β under which AMPI can be optimal for any noise variance. In other words, $\beta \leq \beta^{\min} \leq 1$ implies that the system should not be under-determined. As an example, consider a massive MIMO system that uses QPSK constellations. In the absence of input noise, $\beta^{\min} \approx 2.9505$ (see [8, Tbl. I]). However, by adding the slightest amount of input noise with a continuous probability distribution, such as Gaussian input noise, $\beta^{\min}$ abruptly decreases to values no larger than 1.

IV. Application 1: Massive MIMO

As discussed in Section I, impairment-aware data detection is an important part of practical massive MIMO systems. In reference [1], we provided an impairment-aware data detection algorithm called LAMA-I (short for LArge-MIMO Approximate message passing with transmit Impairments). LAMA-I, which is a low-complexity data detection algorithm, is a specialized version of AMPI for massive MIMO systems. In this section, we briefly revisit the signal and system model for massive MIMO and LAMA-I from [1]. We then provide an optimality analysis of LAMA-I which was not shown in [1]. In particular, we prove that besides optimality in the AMP framework, LAMA-I is able to achieve the same error-rate performance as the IO data detector.

A. LAMA-I: AMPI for Massive MIMO Data Detection

Consider an input signal $\mathbf{s} \in \mathbb{C}^N$ sent through an impaired MIMO channel with input-output relation (1) introduced in Section I with the following assumptions. The entries of $\mathbf{s}$ are chosen from a constellation set $\mathcal{O}$, e.g., QAM, and $\mathbf{s}$ is assumed to have i.i.d. prior distribution $p(\mathbf{s}) = \prod_{\ell=1}^{N} p(s_\ell)$ with the following distribution for each transmit symbol [2]:

$$p(s_\ell) = \sum_{a \in \mathcal{O}} p_a \delta(s_\ell - a). \tag{24}$$

The received vector is $\mathbf{y} \in \mathbb{C}^M$ is the received vector, and N and M denote the number of user equipments and base-station antennas, respectively. The MIMO channel matrix $\mathbf{H} \in \mathbb{C}^{M \times N}$ is assumed to be perfectly known at the receiver.

As shown in Algorithm 2, we first use cB-AMP to compute the Gaussian output $\mathbf{z}^{t_{\max}}$ and the effective noise variance $\sigma^2_{t_{\max}} = \gamma^2_{t_{\max}}$ at iteration t. The MIMO system is decoupled into a set of N parallel and independent additive white Gaussian noise (AWGN) channels. Fig. 5(b) shows the equivalent decoupled system. Using (15), we can compute the marginal posterior distribution $p(s_\ell|\mathbf{y}, \mathbf{H}) = p(s_\ell|z_\ell^{t_{\max}})$ from the Gaussian output. The marginal posterior distribution allows us to compute the MAP estimate for each data symbol independently as

$$\hat{s}_\ell^{t_{\max}} = D(z_\ell^{t_{\max}}, \sigma^2_{t_{\max}}) = \arg\max_{s_\ell \in \mathcal{O}} p(s_\ell|z_\ell^{t_{\max}}). \tag{25}$$

We call this procedure the LAMA-I algorithm in [1]. Note that [1, Sec. IV] details the derivation of LAMA-I for a Gaussian input noise model $p(x_\ell|s_\ell) = \mathcal{CN}(s_\ell, N_{\mathrm{T}}), \forall \ell = 1, \ldots, N$.

B. Individually Optimal (IO) Data Detection

We now show that for uniform linear measurements and the large-system limit, LAMA-I is able to achieve the error-rate performance of the IO data detector (2), if the fixed-point equation (20) has a unique solution. There has been some work in the past focusing on optimality of AMP in the absence of input noise such as [42] (see [7, Sec. IV] for a survey). The core of our optimality analysis is the performance of IO data detection in the presence of input noise based on the replica analysis presented in [43]. To prove individual optimality, we first introduce the following definition. The replica analysis for IO data detection makes the following assumption about $\hat{s}_\ell^{\mathrm{IO}}$.

Definition 4. *The IO solution $\hat{s}_\ell^{\mathrm{IO}}$ is said to satisfy hard-soft assumption, if and only if there exist a function $D : \mathbb{R} \to \mathcal{O}$, with the following properties: (i) $\hat{s}_\ell^{\mathrm{IO}} = D(\mathbb{E}(s_\ell|\mathbf{y}, \mathbf{H}))$ and (ii) for every $s \in \mathcal{O}$ the $D^{-1}(s)$ is Borel measurable and its boundary has Lebesgue measure zero.*

We can prove that the hard-soft assumption is in fact true for equiprobable BPSK constellation points, i.e., we have

$$\mathbb{E}(s_\ell|\mathbf{y}, \mathbf{H}) = \mathbb{P}(s_\ell = +1|\mathbf{y}, \mathbf{H}) - \mathbb{P}(s_\ell = -1|\mathbf{y}, \mathbf{H}),$$

and hence, $\hat{s}_\ell^{\mathrm{IO}} = \mathrm{sign}(\mathbb{E}(s_\ell|\mathbf{y}, \mathbf{H}))$. However, it is an interesting open problem whether the assumption in Definition 4 holds for other, more general, constellations as well.

To simplify our proofs, we make an extra assumption:

Definition 5. *The IO solution $\hat{s}_\ell^{\mathrm{IO}}$ is said to satisfy the fixed-D assumption, if in addition to satisfying the hard-soft assumption, the function D from Definition 4 is only a function of β, $p(s_\ell)$, $p(x_\ell|s_\ell)$, and $p(n_\ell)$. In particular, D does not depend on the dimension N of the input signal.*

Note that for equiprobable BPSK symbols, we have $\hat{s}_\ell^{\mathrm{IO}} = \mathrm{sign}(\mathbb{E}(s_\ell|\mathbf{y}, \mathbf{H}))$ and the fixed-D assumption clearly holds. Intuitively speaking, when the dimensions are large, we do not expect the function D to change with the dimension N.

Before we can establish individual optimality of LAMA-I, we introduce Theorem 5, which analyzes the error probability of the IO solution and provides an equivalent relation, which enables us to compute the error probability of an IO data detector and consequently compare it with other detectors.

Theorem 5. *Suppose that the IO solution satisfies both the hard-soft and fixed-D assumptions in Definitions 4 and 5. Furthermore, assume that the assumptions underlying the replica symmetry in [43] are correct. Then, $\mathbb{P}(\hat{s}_\ell^{\mathrm{IO}} \neq s_\ell)$ converges to $\mathbb{P}(D(Q) \neq S)$ in probability. Here $Q = X + \tilde{\sigma} Z$ with $p(S, X) = p(s_\ell)p(x_\ell|s_\ell)$, $Z \sim \mathcal{CN}(0, 1)$ being independent of (S, X) and $\tilde{\sigma}$ satisfying the following equation:*

$$\tilde{\sigma}^2 = N_0 + \beta \Psi(\tilde{\sigma}^2). \tag{26}$$

We now provide conditions for which LAMA-I algorithm achieves the error-rate performance of IO data detector. The proof is given in Appendix D.

Theorem 6. *Assume the system model in (1) with uniform linear measurements. Suppose that the assumptions made in Theorem 5 hold. Furthermore, assume that the fixed-point equation* (26) *has a unique solution. Let us call the estimate of LAMA-I after t iterations $\hat{s}_\ell^t$. Then in large-system limit, for any $\epsilon \geq 0$ there exists an iteration number t_0 such that*

$$\lim_{N \to \infty} \frac{1}{N} \sum_{\ell=1}^{N} \mathbb{P}(\hat{s}_\ell^{t_0} \neq s_\ell) \leq \lim_{N \to \infty} \frac{1}{N} \sum_{\ell=1}^{N} \mathbb{P}(\hat{s}_\ell^{\mathrm{IO}} \neq s_\ell) + \epsilon,$$

where the limits are taken in probability.[1]

Theorem 6 proves individual optimality of LAMA-I algorithm given certain conditions are met on system size and ratio. The inequality in this theorem shows how LAMA-I with an

[1] If the limit in $\lim_{n \to \infty} X_n = X$ is taken in probability, it means the probability of X_n being far from X should go to zero when n increases.

infinite number of iterations achieves the same error-rate as that of IO data detector. While LAMA-I requires the large-system limit and an infinite number of iterations to achieve the performance of IO data detector, Fig. 2(a) and simulation results in [1] demonstrate that LAMA-I achieves near-IO performance in realistic, finite dimensional large-MIMO systems.

V. Application 2: Compressive Sensing

We now apply AMPI to compressive sensing signal recovery in the presence of input noise. We first introduce the system model and then derive AMPI for this system. We conclude with simulation results that compare AMPI to existing methods.

A. System Model

The noiseless version of CS signal recovery can be solved perfectly under certain conditions on the system dimension and the sparsity level [14], [21]. Recovery under measurement noise has been analyzed extensively; see, e.g., [22], [44]. However, practical CS systems may be affected by input noise [17], [18]. For example, the input signals may not be perfectly sparse or might be affected by noise that appears prior to the measurement process. In what follows, we introduce AMPI to recover the sparse input signal from measurements contaminated with both input and measurement noise. Let $\mathbf{s} \in \mathbb{R}^N$ be the signal of interest we want to reconstruct from the noisy measurements $\mathbf{y} \in \mathbb{R}^M$ with the system model in (1) introduced in Section I. Since the system model for compressive sensing is assumed to be under-determined, we have $M \leq N$ (or equivalently $\beta \geq 1$). Moreover, the input signal $\mathbf{s}$ is a sparse vector with at most K non-zero entries. And, the system matrix $\mathbf{H} \in \mathbb{C}^{M \times N}$ is assumed to be perfectly known.

B. AMPI for Compressive Sensing

In order to apply Algorithm 2 for CS, we first need to derive the effective input signal prior $p(\mathbf{x})$ in (8), which requires the input signal prior $p(\mathbf{s})$. As explained in [22], a practical model to capture sparsity in $\mathbf{s}$ is to assume an i.i.d. Laplace prior. Concretely, we assume

$$p(\mathbf{s}) = \left(\frac{\lambda}{2}\right)^N \exp\left(-\lambda \|\mathbf{s}\|_1\right), \tag{27}$$

with a regularization parameter $\lambda > 0$ that can be tuned to best model the sparsity of the input signal $\mathbf{s}$. For optimal performance, one should tune λ to minimize the MSE of AMPI. Besides the parameter λ, AMPI requires a threshold parameter γ_t^2 that must be tuned in each iteration. To attain optimal performance, both of these parameters should be tuned in each algorithm iteration. To this end, we will use an iteration index subscript for γ_t^2 and for λ_t.

As noted in Section II-B, there exist different methods to tune the threshold parameter γ_t^2. In the paper [45], the authors propose an asymptotically optimal tuning approach using Stein's unbiased risk estimate (SURE) [46] for the threshold parameter γ_t^2. Here, we follow a similar approach that optimally tunes both parameters λ_t and γ_t^2.

First, run Step 1 of Algorithm 2 for $t_{\max}$ iterations. Optimal tuning for AMPI can be achieved if $\lambda_1, \ldots, \lambda_{t_{\max}}$ and $\gamma_1^2, \ldots, \gamma_{t_{\max}}^2$ are tuned in such a way that the value of the asymptotic MSE or $\lim_{N\to\infty} \frac{1}{N}\|\hat{\mathbf{x}}^{t_{\max}+1} - \mathbf{x}\|^2$ is minimized. This requires a joint optimization of the asymptotic MSE over all variables $\{\lambda_1, \ldots, \lambda_{t_{\max}}, \gamma_1^2, \ldots, \gamma_{t_{\max}}^2\}$. However, such a joint optimization is not practical as the iterative nature of AMPI does not allow one to write an explicit expression for MSE. The following theorem shows that one can simplify the joint parameter optimization by tuning each pair (λ_t, γ_t^2) at iteration t starting from $t = 1$ to $t_{\max}$. The proof for this theorem follows from [45, Thm. 3.7] with minor modifications.

Theorem 7. *Suppose that the parameters $\lambda_1, \ldots, \lambda_{t_{\max}}$, $\gamma_1^2, \ldots, \gamma_{t_{\max}}^2$ are optimally tuned for iteration $t_{\max}$ of AMPI. Then, $\forall t < t_{\max}$, the parameters $\lambda_1, \ldots, \lambda_t, \gamma_1^2, \ldots, \gamma_t^2$ are also optimally tuned for iteration t.*

Theorem 7 implies that instead of jointly tuning all parameters $\{\lambda_1, \ldots, \lambda_{t_{\max}}, \gamma_1^2, \ldots, \gamma_{t_{\max}}^2\}$, we can tune (λ_1, γ_1^2) at iteration 1. Given the parameters (λ_1, γ_1^2), we can then tune (λ_2, γ_2^2) at iteration 2, and repeat this process for $t_{\max}$ iterations.

The missing piece is to minimize the MSE at iteration t with appropriate parameters (λ_t, γ_t^2). As cSE in Theorem 1 suggest, the asymptotic MSE at iteration t is given by the function $\Psi(\sigma_t^2, \gamma_t^2, \lambda_t) = \mathbb{E}_{X,Z}\left[\left|\mathsf{F}\big(X + \sigma_t Z, \gamma_t^2, \lambda_t\big) - X\right|^2\right]$, with $X \sim p(x_\ell)$, $Z \sim \mathcal{CN}(0,1)$, and F as the posterior mean function introduced in (12). Notice that all functions Ψ, F, and G will also be a function of λ_t. We use a similar tuning approach as in [45] and since the MSE function $\Psi(\sigma_t^2, \gamma_t^2, \lambda_t)$ depends on the unknown signal prior $p(X)$, we estimate it using SURE in each iteration. For AMPI, SURE is given by

$$\begin{aligned}\hat{\Psi}(\sigma_t^2, \lambda_t, \gamma_t^2) = \tfrac{1}{N}\|\mathsf{F}(\mathbf{z}^t, \gamma_t^2, \lambda_t) - \mathbf{z}^t\|^2 + \sigma_t^2 \\ + \tfrac{2\sigma_t^2}{\gamma_t^2}\left\langle \mathsf{G}(\mathbf{z}^t, \gamma_t^2, \lambda_t) - 1\right\rangle,\end{aligned} \tag{28}$$

where we estimate σ_t^2 by $\|\mathbf{r}^t\|^2/M$ as in [22]. We now modify AMPI as in Algorithm 2 for CS applications as follows.

Algorithm 3 (AMPI-SURE). *Set $\hat{\mathbf{x}}^1 = 0$ and $\mathbf{r}^1 = \mathbf{y}$.*

1) For $t = 1, 2, \ldots, t_{\max}$ compute

$$\begin{aligned}\mathbf{z}^t &= \hat{\mathbf{x}}^t + \mathbf{H}^{\mathrm{H}}\mathbf{r}^t \\ (\lambda_t, \gamma_t^2) &= \underset{\lambda\geq 0, \gamma^2 \geq 0}{\arg\min}\, \hat{\Psi}(\sigma_t^2, \lambda, \gamma^2) \\ \hat{\mathbf{x}}^{t+1} &= \mathsf{F}(\mathbf{z}^t, \gamma_t^2, \lambda_t) \\ \mathbf{r}^{t+1} &= \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{t+1} + \beta\tfrac{\mathbf{r}^t}{\gamma_t^2}\left\langle \mathsf{G}(\mathbf{z}^t, \gamma_t^2, \lambda_t)\right\rangle.\end{aligned}$$

Here, $\hat{\Psi}$ is given by (28), and F and G are the posterior mean and variance given by (12), where $p(x_\ell) = \int_{\mathbb{C}} p(x_\ell|s_\ell)p(s_\ell)\mathrm{d}s_\ell$ and $p(s_\ell) = \frac{\lambda_t}{2}\exp(-\lambda_t|s_\ell|)$.

2) Compute the MMSE estimate for $t = t_{\max}$ with the posterior PDF $p(s_\ell|z_\ell^{t_{\max}})$ as defined in (15) and $p(w_\ell^{t_{\max}}) \sim \mathcal{N}(0, \sigma_{t_{\max}}^2)$. The effective noise variance $\sigma_{t_{\max}}^2$ and signal prior distribution $p(s_\ell)$ are estimated using $\gamma_{t_{\max}}^2$ and $\frac{\lambda_{t_{\max}}^*}{2}\exp\big(-\lambda_{t_{\max}}^*\|\,\mathrm{s}_\ell\|_1\big)$.

Algorithm 3 summarizes AMPI for CS. The only difference to Algorithm 2 is the presence of the extra parameter λ_t that is optimally tuned using SURE depending on the signal sparsity.

C. AMPI Sparse Signal Recovery Under Gaussian Input Noise

AMPI for CS recovery as in Algorithm 3 is defined for a general class of input noise distributions $p(\mathbf{x}|\mathbf{s})$. We now provide a derivation of AMPI for the specific case of Gaussian input noise, i.e., where $p(\mathbf{x}|\mathbf{s}) = \mathcal{CN}(\mathbf{s}, N_{\mathrm{T}}\mathbf{I}_M)$ [17], [18]. The following lemma details the prior $p(\mathbf{x})$ and the functions F and G needed in Steps 1 and 2 of Algorithm 3. The derivations are omitted for brevity and can be found in the supplementary derivations in a slightly longer arXiv version of this paper [47].

Lemma 8. *For a CS system as defined by Section V-A, the prior $p(\mathbf{x})$, the posterior mean F and variance G defined in Algorithm 3 are given by:*

$$p(\mathbf{x}) = \prod_{i=1}^{N} \frac{\lambda}{2} \exp\left(\frac{\lambda^2 N_{\mathrm{T}}}{2}\right)\left(\exp(\lambda x_i) Q\left(\frac{x_i + \lambda N_{\mathrm{T}}}{\sqrt{N_{\mathrm{T}}}}\right) + \exp(-\lambda x_i)\left(1 - Q\left(\frac{x_i - \lambda N_{\mathrm{T}}}{\sqrt{N_{\mathrm{T}}}}\right)\right)\right) \tag{29}$$

$$\mathsf{F}(\hat{x}, \tau, \lambda) = \hat{x} + \lambda\tau\eta(\hat{x}, \tau) \tag{30}$$

$$\mathsf{G}(\hat{x}, \tau, \lambda) = \tau + \lambda^2\tau^2(1 - (\eta(\hat{x}, \tau))^2) - \frac{4}{\gamma(\hat{x}, \tau)} \frac{\lambda\tau^2}{\sqrt{2\pi(N_{\mathrm{T}} + \tau)}}, \tag{31}$$

where we define

$$\eta(\hat{x}, \tau) = \frac{\mathrm{erfcx}(\alpha) - \mathrm{erfcx}(\beta)}{\mathrm{erfcx}(\alpha) + \mathrm{erfcx}(\beta)} \tag{32}$$

$$\gamma(\hat{x}, \tau) = \mathrm{erfcx}(\alpha) + \mathrm{erfcx}(\beta) \tag{33}$$

$$\alpha = \frac{\hat{x} + \lambda(N_{\mathrm{T}} + \tau)}{\sqrt{2(N_{\mathrm{T}} + \tau)}} \tag{34}$$

$$\beta = \frac{-\hat{x} + \lambda(N_{\mathrm{T}} + \tau)}{\sqrt{2(N_{\mathrm{T}} + \tau)}}. \tag{35}$$

The Q-function is $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$, the error function $\mathrm{erfc}(x) = 2Q(\sqrt{2}x)$, and $\mathrm{erfcx}(x) = x^2\mathrm{erfc}(x)$.

Before providing simulation results for AMPI-SURE, we next summarize two baseline algorithms used as a comparison.

D. Two Alternative Algorithms

1) Noise Whitening: Noise whitening has been proposed for data detection and compressive sensing in [11] and [18], respectively. This approach relies on the Gaussian input-noise model, which enables one to "whiten" the impaired system model (1) by multiplying the vector $\mathbf{y}$ with the whitening matrix $\mathbf{W} = N_0\mathbf{Q}^{-\frac{1}{2}}$, where $\mathbf{Q} = N_{\mathrm{T}}\mathbf{H}\mathbf{H}^{\mathrm{H}} + N_0\mathbf{I}_M$ is the covariance matrix of the effective input and receive noise $\mathbf{n} + \mathbf{H}\mathbf{e}$. Whitening results in a statistically equivalent input-output relation $\tilde{\mathbf{y}} = \tilde{\mathbf{H}}\mathbf{s} + \tilde{\mathbf{n}}$, where $\tilde{\mathbf{y}} = \mathbf{W}\mathbf{y}$, $\tilde{\mathbf{H}} = \mathbf{W}\mathbf{H}$ and $\tilde{\mathbf{n}} \sim \mathcal{CN}(0, N_0\mathbf{I}_M)$ is independent of $\mathbf{s}$ [11]. Thus, signal recovery can be performed with conventional algorithms, such as AMP [24], [25]. The drawback of noise whitening is in computing the whitening matrix $\mathbf{W}$, whose dimensions may be extremely large (e.g., in imaging applications). AMPI avoids computation of $\mathbf{W}$, which reduces complexity. Furthermore, AMPI supports more general input noise models—in contrast, noise whitening requires a Gaussian input-noise model.

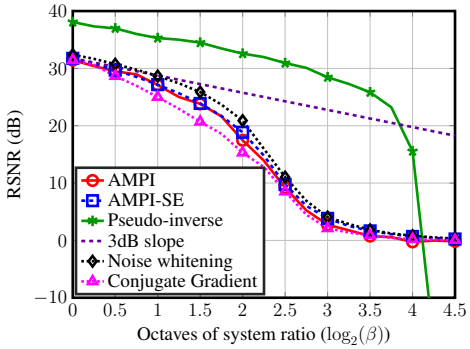


Fig. 6. Reconstruction SNR of AMPI and other algorithms for sparse signal recovery with input noise. The signal has sparsity of 5%, SNR of 30 dB, and is affected by input noise with *EVM* of −30 dB. AMPI achieves the same performance as noise whitening and nonlinear conjugate gradients but at much lower computational complexity.

2) Convex Optimization: Consider the system model in Section V-A. If the input noise is zero, i.e. $\mathbf{x} = \mathbf{s}$, then for a sparse signal $\mathbf{s}$ or equivalently $\mathbf{x}$, sparse signal recovery can be performed by solving [38]

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \tfrac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2 + \lambda\|\mathbf{s}\|_1.$$

If the input noise is non-zero, i.e., $\mathbf{x} \neq \mathbf{s}$, then we can solve the following optimization problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2N_0}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 - \log p(\mathbf{x}) \tag{36}$$

If the term $-\log p(\mathbf{x})$ is convex and differentiable, then we can use efficient algorithms that guarantee convergence to an optimal solution. The following result establishes convexity for the Gaussian input-noise model and provides the gradient, which we use to solve (36). The proof is omitted for brevity and can be found in the supplementary derivations in a slightly longer arXiv version of this paper [47].

Lemma 9. *The objective function $q(\mathbf{x}) = \frac{1}{2N_0}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 - \log p(\mathbf{x})$ is convex and its gradient is given by*

$$\nabla_{\mathbf{x}} q(\mathbf{x}) = \frac{1}{N_0}(\mathbf{H}\mathbf{x} - \mathbf{y})^{\mathrm{T}}\mathbf{H} - \nabla_{\mathbf{x}}[\log p(\mathbf{x})], \tag{37}$$

and, $\nabla_{\mathbf{x}}[\log p(\mathbf{x})] = \lambda[\eta(x_1, 0), \ldots, \eta(x_N, 0)]^{\mathrm{T}}$ with $\eta(\hat{x}, \tau)$ in (32).

Hence, we propose to use the non-linear conjugate gradient method of Polak-Ribiere [48] to solve for (36); see [49, alg. 4.4] for the algorithm details. The downside of such an approach is that there is no known fast approach to set the parameter λ. In contrast, AMPI in Algorithm 3 can be tuned optimally.

E. Results and Comparison

Fig. 6 shows simulation results for sparse signal recovery with a sparsity rate of $\frac{K}{N} = 5\%$ and signal dimension of $N = 1000$. The input signal is generated with a Bernoulli-Gaussian distribution. The indices of the non-zeros entries are selected from an equiprobable Bernoulli distribution and each entry is generated from a standard normal distribution. The reconstruction signal to noise ratio (RSNR) is plotted as a function of the octaves of system ratio β (also referred to as sub-sampling ratio). Furthermore, we consider an average

SNR of $SNR = \mathbb{E}\big[\|\mathbf{Hs}\|^2\big]/\mathbb{E}\big[\|\mathbf{n}\|^2\big] = \beta\frac{E_s}{N_0} = 30\,\text{dB}$ and an error vector magnitude of $EVM = \mathbb{E}\big[\|\mathbf{e}\|^2\big]/\mathbb{E}\big[\|\mathbf{s}\|^2\big] = \frac{N_T}{E_s} = -30\,\text{dB}$. In Fig. 6, the RSNR results of each algorithm is averaged over 20 different randomly-created input signals. The performance of AMPI with 100 iterations is depicted as a solid circle-marked red curve. AMPI's performance almost perfectly matches the cSE predictions in (16) (the dashed square-marked blue curve). As a comparison, we show the performance of noise whitening and convex optimization with non-linear conjugate gradients. The dotted diamond-marked black curve shows the performance of noise whitening followed by AMP. While whitening achieves the same performance as AMPI, it entails significantly higher complexity as one has to first compute the whitening matrix. The dotted triangle-marked magenta curve shows the performance of nonlinear conjugate gradients with 100 iterations. This method requires a computationally expensive grid search to tune the parameter λ. Since we set this tuning parameter using the *optimal* ones from AMPI (solid circle-marked red curve), the conjugate gradients method performs very well. The solid star-marked green curve corresponds to the oracle-based approach of taking the pseudo-inverse assuming the support is known. As [16] suggests, the RSNR of this method decays with a slope of 3 dB per octave. However, none of the proposed algorithms follows the 3dB per octave slope.

We now compare the computational complexity of the three algorithms: (i) AMPI, (ii) non-linear conjugate gradients, and (iii) noise whitening followed by conventional AMP. To this end, we count the number of real-valued multiplications as a proxy for algorithm complexity. For AMPI, the per-iteration complexity is $2MN + 13N + M$. For non-linear conjugate gradients, the complexity is $MN + M^2 + 7N + 21$ plus a per-iteration complexity of $10N$, assuming the algorithm is provided with the optimal sparsity parameter λ. As discussed in the previous paragraph, one needs a grid search to find the optimal parameter λ. For noise whitening, the complexity includes a noise whitening preprocessing stage followed by AMP which is iterative. Specifically, the initial preprocessing complexity of this approach is $M^3+M^2N+2M^2+2$ (in order to apply noise whitening to the system) plus a per-iteration complexity of $2MN + M$ for AMP, assuming the algorithm is provided with the optimal sparsity parameter λ. Similar to non-linear conjugate gradients, one needs a grid search to find the optimal parameter λ. As it is evident from this complexity analysis, the noise whitening approach exhibits the highest complexity followed by non-linear conjugate gradients, which also has higher complexity than our proposed AMPI algorithm. Note that both the noise whitening approach and non-linear conjugate gradients require a grid search in order to determine the optimal parameter λ in every algorithm iteration, which further increases complexity compared to AMPI.

VI. Conclusion

We have introduced AMPI (short for approximate message passing with input noise), a novel data detection and estimation algorithm for systems that are corrupted by input noise. AMPI is computationally efficient and can be used for a broad range of input-noise models. Furthermore, the complex state-evolution (cSE) framework enables a theoretical analysis of AMPI in the large system limit and for i.i.d. Gaussian measurement matrices. Under these conditions, we have investigated optimality conditions of AMPI for data detection and signal estimation. We have shown that AMPI is optimal within the AMP framework and, under additional assumptions, achieves individually-optimal error-rate performance in massive MIMO applications. For the Gaussian input-noise model, we have used numerical results to show that AMPI outperforms methods that ignore input noise and performs on-par with whitening and optimization-based methods, but at (often significantly) lower complexity.

Appendix A
Proof of Theorem 2

A. Proof Outline

We provide an optimality proof for an application where AMPI solves the IO problem in (2). This means that in Step 2 of AMPI, we use the MAP estimate. The optimality proof where AMPI is supposed to minimize the MSE follows analogously. Suppose that we use AMPI with an arbitrary set of pseudo-Lipschitz functions $\mathsf{F}_1, ..., \mathsf{F}_{t_{\max}+1}$ as described in (18). In this proof, we will establish that AMPI in Algorithm 2 chooses these functions such that the outputs $\hat{s}_\ell$ for $\ell = 1, ..., N$, correspond to the solution of the IO problem (2) in the large-system limit. We start by writing down the optimality criterion in (2) as

$$\mathsf{F}_{t_{\max}+1}(z_\ell^{t_{\max}+1}, \sigma^2_{t_{\max}+1}) = \arg\min_{\mathsf{F}} \mathbb{P}\big(\mathsf{F}(z_\ell^{t_{\max}+1}, \sigma^2_{t_{\max}+1}) \neq s_\ell\big). \quad (38)$$

The estimate generated by the function $\mathsf{F}_{t_{\max}+1}(z_\ell^{t_{\max}+1}, \sigma^2_{t_{\max}+1})$ during Step 2 at iteration $t_{\max}+1$ minimizes the per-entry symbol-error probability. Note that the functions $\mathsf{F}_1, \ldots, \mathsf{F}_{t_{\max}+1}$ operate element-wise on vectors.

Remark 2. *The criterion (38) appears to only consider the ℓth entry. Since, however, the probability in (38) is taken w.r.t. the randomness in the matrix* $\mathbf{H}$*, the vector* $\mathbf{s}$*, input and receive noise, the criterion is in fact affected by all other entries.*

We now establish the optimality proof by the following two lemmas, with proofs in Appendix A-B and A-C. In what follows, we assume the random variables $S \sim p(s_\ell), X|S \sim p(x_\ell|s_\ell)$, and $Z \sim \mathcal{CN}(0,1)$ to be independent of X and S.

Lemma 10. *Let the assumptions of Thm. 2 hold. For the criterion (38) to hold for $\ell = 1, \ldots, N$, $\mathsf{F}_{t_{\max}+1}(z_\ell^{t_{\max}+1}, \sigma^2_{t_{\max}+1})$ at iteration $t_{\max}+1$ must be the MAP estimator*

$$\mathsf{F}_{t_{\max}+1}(z_\ell^{t_{\max}+1}, \sigma^2_{t_{\max}+1}) = \arg\max_{s_\ell \in \mathcal{O}} p_{S|X+\sigma_{t_{\max}+1}Z}(s_\ell|z_\ell^{t_{\max}+1}) \quad (39)$$

Lemma 11. *Let the assumptions of Thm. 2 hold. For the criterion (38) to hold for $\ell = 1, \ldots, N$, the functions $\mathsf{F}_t(z_\ell^t, \sigma_t^2)$, $t = 1, \ldots, t_{\max}$, must be the unique set of MMSE estimators, i.e., $\mathsf{F}_t(z_\ell^t, \sigma_t^2) = \mathbb{E}_{X|X+\sigma_t Z}[x_\ell|z_\ell^t]$.*

Lemma 10 suggests that for optimality to hold, the function $\mathsf{F}_{t_{\max}+1}(z_\ell^{t_{\max}+1}, \sigma^2_{t_{\max}+1})$ must be the MAP estimator as provided in Step 2 of Algorithm 2. Furthermore, Lemma 11 suggest that if the solution to the fixed-point equation of (20) is unique, then the set of functions $\mathsf{F}_1, \ldots, \mathsf{F}_{t_{\max}}$ that satisfy

the optimality criterion are given by the unique set of MMSE functions $\mathsf{F}_t(z_\ell^t, \sigma_t^2) = \mathbb{E}_{X|X+\sigma_t Z}[x_\ell | z_\ell^t]$ for $t = 1, \ldots, t_{\max}$. These functions match the posterior mean function F (defined by (12)) in Step 1 of Algorithm 2. Thus, from Lemma 10 and Lemma 11, we conclude that Algorithm 2 solves the IO problem (2) given that the fixed point equation (20) is unique.

B. Proof of Lemma 10

We start with the following lemma.

Lemma 12. *Define* $\zeta_N = \frac{1}{N}\sum_{\ell=1}^N \mathbf{1}(\mathsf{F}_{t_{\max}+1}(z_\ell^{t_{\max}+1}, \sigma_{t_{\max}+1}^2) \neq s_\ell)$. *Fix the system ratio* $\beta = N/M$ *and let* $N \to \infty$. *Then, for a given* $\mathbf{H}$, *we have*

$$\zeta_N \xrightarrow{a.s.} \mathbb{E}_{X,S,Z}\left[\mathbf{1}(\mathsf{F}_{t_{\max}+1}(X + \sigma_{t_{\max}+1} Z, \sigma_{t_{\max}+1}^2) \neq S)\right] = \zeta_\infty, \tag{40}$$

Proof. The proof follows from [40, Thm. 1] and we briefly outline the main ideas. Reference [40] established that for any Pseudo-Lipschitz function ψ in the large-system limit we have

$$\frac{1}{N}\sum_{\ell=1}^N \psi(z_\ell^t, x_\ell) \to \mathbb{E}\left[\psi(X + \sigma_t Z, X)\right]. \tag{41}$$

Note that the expression on the left in (41) is the expectation under the empirical distribution of the joint random variables (z_ℓ^t, x_ℓ). Hence, we can say this equation corresponds to one of the forms of convergence in distribution as pointed out in [50, Lem. 2.2]. Based on this lemma, (41) suggests that the empirical distribution of (z_ℓ^t, x_ℓ) also converges weakly to the distribution of $(X + \sigma_t Z, X)$. Furthermore, since $s_\ell \to x_\ell \to \mathbf{y}$ forms a Markov chain, z_ℓ^t (which is a function of $\mathbf{y}$) is independent of s_ℓ given x_ℓ; this implies that the empirical distribution of (z_ℓ^t, s_ℓ) converges weakly to the distribution of $(X + \sigma_t Z, S)$. Hence, based on the same Lemma 2.2 in [50] we can conclude that if $\mathcal{B}$ is a Borel measurable set, whose boundary has Lebesgure measure zero (and if $X + \sigma_t Z$ is absolutely continuous with respect the Lebesgue measure) then $\frac{1}{N}\sum_{\ell=1}^N \mathbf{1}\left((z_\ell^t, s_\ell) \in \mathcal{B}\right) \xrightarrow{d} \mathbb{P}\left((X + \sigma_t Z, S) \in \mathcal{B}\right)$. From this result, we conclude that (40) holds. □

Notice from Lemma 12 that ζ_N is bounded. As a consequence, in the large-system limit, we have

$$\mathbb{E}_{\mathbf{y},\mathbf{H}}[\zeta_N] \xrightarrow{\text{a.s}} \mathbb{E}_{\mathbf{y},\mathbf{H}}[\zeta_\infty]. \tag{42}$$

Let us compute

$$\begin{aligned}\mathbb{E}_{\mathbf{y},\mathbf{H}}\left[\zeta_N\right] &= \mathbb{E}_{\mathbf{y},\mathbf{H}}\left[\frac{1}{N}\sum_{\ell=1}^N \mathbf{1}(\mathsf{F}_{t_{\max}+1}(z_\ell^{t_{\max}+1}, \sigma_{t_{\max}+1}^2) \neq s_\ell)\right] \\ &= \mathbb{P}\left(\mathsf{F}_{t_{\max}+1}(z_\ell^{t_{\max}+1}, \sigma_{t_{\max}+1}^2) \neq s_\ell\right),\ \ell = 1, \ldots, N. \end{aligned} \tag{43}$$

Here, the last equality holds because under the permutations of the entries in $\mathbf{s}$, the distribution does not change and thus, $\mathbb{P}\left(\mathsf{F}(z_\ell^{t_{\max}+1}, \sigma_{t_{\max}+1}^2) \neq s_\ell\right)$ does not depend on the index ℓ. Hence, using (42) and (43), in the large system limit we have

$$\begin{aligned}&\mathbb{P}\left(\mathsf{F}_{t_{\max}+1}(z_\ell^{t_{\max}+1}, \sigma_{t_{\max}+1}^2) \neq s_\ell\right) \\ &\xrightarrow{\text{a.s}} \mathbb{E}_{\mathbf{y},\mathbf{H}}[\zeta_\infty] = \mathbb{P}\left(\mathsf{F}_{t_{\max}+1}(X + \sigma_{t_{\max}+1} Z, \sigma_{t_{\max}+1}^2) \neq S\right). \end{aligned} \tag{44}$$

Hence, instead of minimizing $\mathbb{P}\left(\mathsf{F}(z_\ell^{t_{\max}+1}, \sigma_{t_{\max}+1}^2) \neq s_\ell\right)$ in (38), we can minimize $\mathbb{P}\left(\mathsf{F}(X + \sigma_{t_{\max}+1} Z, \sigma_{t_{\max}+1}^2) \neq S\right)$. Thus, the optimal choice of F at iteration $t_{\max}+1$ is the MAP estimator in (39).

C. Proof of Lemma 11

We next show that satisfying (38), requires the functions $\mathsf{F}_t(z_\ell^t, \sigma_t^2)$, $t = 1, \ldots, t_{\max}$, to be the MMSE estimators. Let us call the MAP estimator $\mathsf{F}_{t_{\max}+1}$ at iteration $t_{\max}+1$ from Lemma 10 as $\mathsf{F}_\sigma^{\text{MAP}}$. Then, the following lemma holds.

Lemma 13. $\mathbb{P}\left(\mathsf{F}_\sigma^{\text{MAP}}(x_\ell + \sigma Z, \sigma^2) \neq s_\ell\right)$ *is a non-decreasing function in* σ.

The proof follows by contradiction. In particular, we try to show that exists two quantities $\sigma_1 < \sigma_2$ such that

$$\begin{aligned}&\mathbb{P}\left(\mathsf{F}_{\sigma_1}^{\text{MAP}}(x_\ell + \sigma_1 Z, \sigma_1^2) \neq s_\ell\right) \\ &\qquad > \mathbb{P}\left(\mathsf{F}_{\sigma_2}^{\text{MAP}}(x_\ell + \sigma_2 Z, \sigma_2^2) \neq s_\ell\right). \end{aligned} \tag{45}$$

Based on $x_\ell + \sigma_1 Z$, we consider the randomized estimator $\mathsf{F}_{\sigma_2}^{\text{MAP}}(x_\ell + \sigma_1 Z + \sqrt{\sigma_2^2 - \sigma_1^2}\tilde{Z}, \sigma_2^2)$, where $\tilde{Z} \sim \mathcal{CN}(0,1)$ independent of Z. It is easy to see that since $\sigma_1 Z + \sqrt{\sigma_2^2 - \sigma_1^2}\tilde{Z}$ is distributed according to $\mathcal{CN}(0, \sigma_2^2)$, we have

$$\begin{aligned}&\mathbb{P}\Big(\mathsf{F}_{\sigma_2}^{\text{MAP}}(x_\ell + \sigma_1 Z + \sqrt{\sigma_2^2 - \sigma_1^2}\tilde{Z}, \sigma_2^2) \neq s_\ell\Big) \\ &\qquad = \mathbb{P}\left(\mathsf{F}_{\sigma_2}^{\text{MAP}}(x_\ell + \sigma_2 Z, \sigma_2^2) \neq s_\ell\right). \end{aligned} \tag{46}$$

Hence,

$$\begin{aligned}&\mathbb{E}_{\tilde{Z}}\left[\mathbb{P}\Big(\mathsf{F}_{\sigma_2}^{\text{MAP}}(x_\ell + \sigma_1 Z + \sqrt{\sigma_2^2 - \sigma_1^2}\tilde{Z}, \sigma_2^2) \neq s_\ell \Big| \tilde{Z}\Big)\right] \\ &\qquad = \mathbb{P}\left(\mathsf{F}_{\sigma_2}^{\text{MAP}}(x_\ell + \sigma_2 Z, \sigma_2^2) \neq s_\ell\right). \end{aligned} \tag{47}$$

Hence, there exists a value of $\tilde{Z}$, call it $\bar{Z}$, for which

$$\begin{aligned}&\mathbb{P}\left(\mathsf{F}_{\sigma_2}^{\text{MAP}}(x_\ell + \sigma_1 Z + \sqrt{\sigma_2^2 - \sigma_1^2}\bar{Z}, \sigma_2^2) \neq s_\ell\right) \\ &\qquad < \mathbb{P}\left(\mathsf{F}_{\sigma_2}^{\text{MAP}}(x_\ell + \sigma_2 Z, \sigma_2^2) \neq s_\ell\right). \end{aligned} \tag{48}$$

Note that this estimator is the non-randomized estimator of $x_\ell + \sigma_1 Z$. Consequently, we have

$$\begin{aligned}&\mathbb{P}\left(\mathsf{F}_{\sigma_2}^{\text{MAP}}(x_\ell + \sigma_1 Z + \sqrt{\sigma_2^2 - \sigma_1^2}\tilde{\theta}, \sigma_2^2) \neq s_\ell\right) \\ &\qquad \leq \mathbb{P}\left(\mathsf{F}_{\sigma_2}^{\text{MAP}}(x_\ell + \sigma_2 \theta, \sigma_2^2) \neq s_\ell\right), \end{aligned} \tag{49}$$

which is in contradiction with (46).

Lemma 13 shows that in order for $\mathsf{F}_\sigma^{\text{MAP}}$ to provide the smallest probability of error in (38), the function sequence $\{\mathsf{F}_1, \ldots, \mathsf{F}_{t_{\max}}\}$ should lead to the minimum possible $\sigma_{t_{\max}+1}^2$. In Lemma 15, we prove that $\sigma_{t_{\max}+1}^2$ is minimal only if $\{\mathsf{F}_1, \ldots, \mathsf{F}_{t_{\max}}\}$ are the MMSE estimators. We first provide Lemma 14, which is required in the proof for Lemma 15.

Lemma 14. $\inf_{\mathsf{F}} \mathbb{E}_{X,Z}\left[\left|\mathsf{F}(X + \sigma Z, \sigma^2) - X\right|^2\right]$ *is a nondecreasing function in* σ.

The proof follows by contradiction. Suppose that the statement of Lemma 14 is not true. Then, there exists two quantities $\hat{\sigma}_1 < \hat{\sigma}_2$ such that

$$\begin{aligned}&\inf_{\mathsf{F}} \mathbb{E}_{X,Z}\left[\left|\mathsf{F}(X + \hat{\sigma}_1 Z, \hat{\sigma}_1^2) - X\right|^2\right] \\ &\qquad > \inf_{\mathsf{F}} \mathbb{E}_{X,Z}\left[\left|\mathsf{F}(X + \hat{\sigma}_2 Z, \hat{\sigma}_2^2) - X\right|^2\right]. \end{aligned} \tag{50}$$

Now suppose that both infima in (50) are achieved with $\mathsf{F}_{\hat{\sigma}_1}$ and $\mathsf{F}_{\hat{\sigma}_2}$, respectively. Then, we can construct a new estimator $\tilde{\mathsf{F}}_{\hat{\sigma}_1}$ for the variance $\hat{\sigma}_1$ as

$$\tilde{\mathsf{F}}_{\hat{\sigma}_1}(X+\hat{\sigma}_1 Z,\hat{\sigma}_1^2) = \mathbb{E}_{\tilde{Z}}\left[\mathsf{F}_{\hat{\sigma}_2}(X+\hat{\sigma}_1 Z+\sqrt{\hat{\sigma}_2^2-\hat{\sigma}_1^2}\tilde{Z},\hat{\sigma}_2^2)\Big| Z\right], \quad (51)$$

where $\tilde{Z} \sim \mathcal{CN}(0,1)$. Hence, $\hat{\sigma}_1 Z + \sqrt{\hat{\sigma}_2^2-\hat{\sigma}_1^2}\tilde{Z} \sim \mathcal{CN}(0,\hat{\sigma}_2^2)$. We now prove that $\tilde{\mathsf{F}}_{\hat{\sigma}_1}$ has a lower risk than $\mathsf{F}_{\hat{\sigma}_1}$, which is in contradiction with $\mathsf{F}_{\hat{\sigma}_1}$ achieving infimum of the function $\mathbb{E}_{X,Z}\left[\left|\mathsf{F}(X+\hat{\sigma}_1 Z,\hat{\sigma}_1^2)-X\right|^2\right]$ for $\hat{\sigma}_1$, i.e.,

$$\mathbb{E}^\star\left[\left|\tilde{\mathsf{F}}_{\hat{\sigma}_1}(X+\hat{\sigma}_1 Z,\hat{\sigma}_1^2)-X\right|^2\right] \quad (52)$$

$$= \mathbb{E}^\star\left[\left|\mathbb{E}_{\tilde{Z}}\left[\mathsf{F}_{\hat{\sigma}_2}(X+\hat{\sigma}_1 Z+\sqrt{\hat{\sigma}_2^2-\hat{\sigma}_1^2}\tilde{Z},\hat{\sigma}_2^2)\Big| Z\right]-X\right|^2\right]$$

$$\overset{(a)}{\leq} \mathbb{E}^\star\left[\mathbb{E}_{\tilde{Z}}\left[\left|\mathsf{F}_{\hat{\sigma}_2}(X+\hat{\sigma}_1 Z+\sqrt{\hat{\sigma}_2^2-\hat{\sigma}_1^2}\tilde{Z},\hat{\sigma}_2^2)-X\right|^2\Big| Z\right]\right] \quad (53)$$

$$= \mathbb{E}^\star\left[\left|\mathsf{F}_{\hat{\sigma}_2}(X+\hat{\sigma}_2 Z,\hat{\sigma}_2^2)-X\right|^2\right] \quad (54)$$

$$\overset{(b)}{<} \mathbb{E}^\star\left[\left|\mathsf{F}_{\hat{\sigma}_1}(X+\hat{\sigma}_1 Z,\hat{\sigma}_1^2)-X\right|^2\right], \quad (55)$$

where $\mathbb{E}^\star[\cdot]$ is the expectation over the random variables X and Z. Here, the two inequalities (a) and (b) come from Jensen's inequality and assumption (50), respectively.

Lemma 15. *The sequence of functions $\{\mathsf{F}_1,\ldots,\mathsf{F}_{t_{\max}}\}$ must be the MMSE estimators to lead to the minimum $\sigma^2_{t_{\max}+1}$.*

The proof follows by induction. Suppose that the functions $\mathsf{F}_1,\ldots,\mathsf{F}_{t-1}$ are MMSE estimators to minimize σ_t^2. Then, we prove by contradiction that to minimize σ_{t+1}^2, all functions $\mathsf{F}_1,\ldots,\mathsf{F}_t$ must be MMSE estimators. Now, suppose that $\mathsf{F}_1^*,\ldots,\mathsf{F}_t^*$ are the optimal functions that lead to the minimum effective noise variance that we call σ_{t+1}^{*2}. And assume that at least one of these functions is not an MMSE estimator. Then, we prove that if $\bar{\mathsf{F}}_1,\ldots,\bar{\mathsf{F}}_t$ are all MMSE estimators, they generate a lower variance $\bar{\sigma}_{t+1}^2$. Let us compute $\bar{\sigma}_{t+1}^2$ from the c-SE framework in Theorem 1:

$$\bar{\sigma}_{t+1}^2(\bar{\mathsf{F}}_1,\ldots,\bar{\mathsf{F}}_t) \overset{(a)}{=} N_0+\beta\,\mathbb{E}_{X,Z}\left[\left|\bar{\mathsf{F}}_t(X+\bar{\sigma}_t Z,\bar{\sigma}_t^2)-X\right|^2\right] \quad (56)$$

$$\overset{(b)}{=} N_0+\beta\inf_{\mathsf{F}_t}\mathbb{E}_{X,Z}\left[\left|\mathsf{F}_t(X+\bar{\sigma}_t Z,\bar{\sigma}_t^2)-X\right|^2\right] \quad (57)$$

$$\overset{(c)}{\leq} N_0+\beta\inf_{\mathsf{F}_t}\mathbb{E}_{X,Z}\left[\left|\mathsf{F}_t(X+\sigma_t^* Z,\sigma_t^{*2})-X\right|^2\right] \quad (58)$$

$$\leq N_0+\beta\,\mathbb{E}_{X,Z}\left[\left|\mathsf{F}_t^*(X+\sigma_t^* Z,\sigma_t^{*2})-X\right|^2\right] \quad (59)$$

$$\overset{(d)}{=} \sigma_{t+1}^{*2}(\mathsf{F}_1^*,...,\mathsf{F}_t^*). \quad (60)$$

Here, (a) and (d) follow from cSE, (b) follows from $\bar{\mathsf{F}}_t$ being an MMSE estimator. Lastly, (c) follows from Lemma 14 and the base case of induction, i.e., $\bar{\sigma}_t^2(\bar{\mathsf{F}}_1,...,\bar{\mathsf{F}}_{t-1}) < \sigma_t^{*2}(\mathsf{F}_1^*,...,\mathsf{F}_{t-1}^*)$. By inspecting inequality (60), we see that it is in contradiction with the optimality assumption of $\mathsf{F}_1^*,...,\mathsf{F}_t^*$ unless $\mathsf{F}_i^*=\bar{\mathsf{F}}_i$ for $i=1,...,t$. Note that here we have assumed that at every stage, the MMSE estimator is unique. Because otherwise, there can be another set of MMSE estimators $\tilde{\mathsf{F}}_1,...,\tilde{\mathsf{F}}_t$ which generates a lower $\tilde{\sigma}_{t+1}^2$. Thus, this lemma proves that if $\mathsf{F}_1,...,\mathsf{F}_{t_{\max}}$ are a unique set of MMSE estimators, then they generate the minimum $\sigma^2_{t_{\max}+1}$ by letting $t_{\max}\to\infty$.

From Lemmas 13 and 15, we conclude that to satisfy (38), the functions $\mathsf{F}_t(z_\ell^t,\sigma_t^2)$, $t=1,\ldots,t_{\max}$, must be the set of MMSE estimators, i.e., $\mathsf{F}_t(z_\ell^t,\sigma_t^2)=\mathbb{E}_{X|X+\sigma_t Z}[x_\ell|z_\ell^t]$, which is equivalent to the message mean (12) for $t=1,\ldots,t_{\max}$ in Step 1 of AMPI. Note that for optimality to hold, we need the MMSE estimators to be unique. If the solution to the fixed-point equation of (20) is unique, then this guarantees uniqueness of the MMSE estimators.

Appendix B
Proof of Lemma 4

Starting with Definition 3 of $\beta^{\max}$, assume that the minimum in this equation is achieved by $\sigma^2=\bar{\sigma}^2$. Thus,

$$\beta^{\max}=\left(\frac{\Psi(\bar{\sigma}^2,\bar{\sigma}^2)}{\bar{\sigma}^2}\right)^{-1}\geq 1. \quad (61)$$

Here, the inequality comes from [51, Prop. 4] which provides an upper bound for the MSE function $\Psi(\sigma^2,\sigma^2)$ as follows: $\Psi(\sigma^2,\sigma^2)\leq\sigma^2, \quad \forall\sigma^2\geq 0$. Recall from Section II-B that AMPI decouples the system into a set of N parallel and independent AWGN channels $z_\ell=x_\ell+\sigma Z$ with $Z\sim\mathcal{N}(0,1)$ and σ_t^2 being the effective noise variance computed using state evolution equations in Section II-D. Hence, using [52, Thm. 12] for each AWGN channels, we conclude that if the prior signal distribution $p(x_\ell)$ is continuous and bounded, then the MMSE dimension $\mathbf{D}$ as defined below will have the value of 1, i.e., $\mathbf{D}(x_\ell,Z)=\lim_{\sigma^2\to 0}\frac{\Psi(\sigma^2,\sigma^2)}{\sigma^2}=1$. Using this equation along with the definition of $\beta^{\max}$ in Definition 3, we obtain

$$\beta^{\max}=\min_{\sigma^2>0}\left\{\left(\frac{\Psi(\sigma^2,\sigma^2)}{\sigma^2}\right)^{-1}\right\}\leq\left(\lim_{\sigma^2\to 0}\frac{\Psi(\sigma^2,\sigma^2)}{\sigma^2}\right)^{-1}=1. \quad (62)$$

From (61) and (62), we have $\beta^{\max}=1$. Additionally by [8, Lem. 4], $\beta^{\min}\leq\beta^{\max}=1$ which completes the proof.

Appendix C
Proof of Theorem 5

To simplify notation, we denote the PDF of all distributions by p. Now to characterize the error probability of IO data detector, we start with the hard-soft assumption

$$\mathbb{P}(\hat{s}_\ell^{\text{IO}}\neq s_\ell)=\mathbb{P}(D(\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))\neq s_\ell). \quad (63)$$

Based on this assumption, we have to characterize the joint distribution of $(s_\ell,\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))$. Note that in [43] the limiting distribution of $(x_\ell,\mathbb{E}(x_\ell|\mathbf{y},\mathbf{H}))$ has been characterized. We will use this limiting distribution to study $(s_\ell,\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))$.

From (1), we have that $s_\ell\to x_\ell\to\mathbf{y}$ is a Markov chain. This implies that the random variable $q_\ell=\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H})$ which is a function of $\mathbf{y}$ and $\mathbf{H}$ is independent of s_ℓ given x_ℓ. Hence, instead of $(s_\ell,\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))$, we can characterize the limiting distribution of $(s_\ell,x_\ell,\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))$ which can be written as:

$$p(s_\ell,x_\ell,\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))=p(s_\ell,x_\ell)p(\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H})|x_\ell,s_\ell) \quad (64)$$

$$=p(s_\ell,x_\ell)p(\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H})|x_\ell). \quad (65)$$

Since the joint distribution $p(s_\ell, x_\ell)$ is known, characterizing the distribution of $(s_\ell, x_\ell, \mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))$ reduces to characterizing the distribution of $(\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H})|x_\ell)$ (or equivalently $(x_\ell, \mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))$). Let us compute $\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H})$, which is

$$\mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}) = \int s_\ell p(s_\ell|\mathbf{y},\mathbf{H})\mathrm{d}s_\ell \tag{66}$$

$$= \int \mathbb{E}(s_\ell|x_\ell) p(x_\ell|\mathbf{y},\mathbf{H})\mathrm{d}x_\ell. \tag{67}$$

Define $L(x_\ell) \triangleq \mathbb{E}(s_\ell|x_\ell)$. Thus, our original problem of characterizing the limiting distribution of $(s_\ell, \mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))$ is simplified to characterizing the limiting distribution of $(x_\ell, \mathbb{E}(L(x_\ell)|\mathbf{y},\mathbf{H}))$. This latter problem can be solved by the replica method as explained in [43]. Assuming that the assumptions underlying the replica symmetry in [43] are correct, we can argue from claim 1 in this paper that

$$(x_\ell, \mathbb{E}(L(x_\ell)|\mathbf{y},\mathbf{H}) \xrightarrow{d} (X, \mathbb{E}(L(X)|X+\tilde{\sigma}Z)), \tag{68}$$

where $S \sim p(s_\ell)$, $X|S \sim p(x_\ell|s_\ell)$, $Z \sim N(0,1)$ is independent of S and X, and $\tilde{\sigma}$ satisfies the fixed point equation

$$\tilde{\sigma}^2 = N_0 + \beta\Psi(\tilde{\sigma}^2). \tag{69}$$

Note that $\mathbb{E}(L(X)|X+\tilde{\sigma}Z) = \mathbb{E}(\mathbb{E}(S|X)|X+\tilde{\sigma}Z) = \mathbb{E}(S|X+\tilde{\sigma}Z)$. In other words, (68) can be written as

$$(x_\ell, \mathbb{E}(L(x_\ell)|\mathbf{y},\mathbf{H}) \xrightarrow{d} (X, \mathbb{E}(S|X+\tilde{\sigma}Z)). \tag{70}$$

Next, we use this result to characterize the joint limiting distribution of $(s_\ell, x_\ell, \mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))$. If we define $q_\ell = \mathbb{E}(s_\ell|\mathbf{y},\mathbf{H})$ and $Q = X + \tilde{\sigma}Z$, then for every $s, x, q \in \mathbb{R}$ we have $p_{q_\ell,x_\ell}(q,x) \to p_{Q,X}(q,x)$, and, furthermore,

$$f_{s_\ell,x_\ell,q_\ell}(s,x,q) = p_{s_\ell|x_\ell}(s|x)p_{q_\ell,x_\ell}(q,x) \tag{71}$$

$$= p_{S|X}(s|x)p_{q_\ell,x_\ell}(q,x), \tag{72}$$

which will converge to $p_{S|X}(s|x)p_{Q,X}(q,x)$. Consequently, $(s_\ell, x_\ell, \mathbb{E}(s_\ell|\mathbf{y},\mathbf{H}))$ converges to $(S, X, \mathbb{E}(S|X+\tilde{\sigma}Z))$ in distribution, which along with markov chain $s_\ell \to x_\ell \to \mathbb{E}(s_\ell|\mathbf{y},\mathbf{H})$ leads to the result $(s_\ell, q_\ell) \xrightarrow{d} (S, Q)$, or equivalently,

$$p_{q_\ell|s_\ell}(q|s) \xrightarrow{d} p_{Q|S}(q|s). \tag{73}$$

Next, we will use this result to characterize the error probability of IO data detector $\mathbb{P}(\hat{s}_\ell^{\mathrm{IO}} \neq s_\ell)$. To simplify the rest of the proof we make several assumptions that are correct for systems MIMO. Suppose $s_\ell \in \mathcal{O}$ and that the cardinality of this set is finite. From (63), the IO error probability can be written as $\mathbb{P}(D(q_\ell) \neq s_\ell)$ which converges as follows for a given $s_\ell = s$:

$$\mathbb{P}(D(q_\ell) \neq s_\ell|s_\ell = s) = 1 - \mathbb{P}(q_\ell \in D^{-1}(s)|s_\ell = s) \tag{74}$$

$$\to 1 - \mathbb{P}(Q \in D^{-1}(s)|S = s). \tag{75}$$

The last claim is a consequence of [50, Lem. 2.2] which connects convergence in distribution of (73) to the convergence in probability above. This relation holds due to the fact that the boundary of D^{-1} has Lebesgue measure zero. Averaging over all values of $s_\ell \in \mathcal{O}$, we obtain

$$\mathbb{P}(D(q_\ell) \neq s_\ell) = \sum_{s\in\mathcal{O}} \mathbb{P}(D(q_\ell) \neq s_\ell|s_\ell = s)p(s_\ell = s) \tag{76}$$

$$\to \sum_{s\in\mathcal{O}} \mathbb{P}(D(Q) \neq S|S = s)p(S = s) = \mathbb{P}(D(Q) \neq S). \tag{77}$$

Hence, we have $\mathbb{P}(\hat{s}_\ell^{\mathrm{IO}} \neq s_\ell) \to \mathbb{P}(D(Q) \neq S)$.

APPENDIX D
PROOF OF THEOREM 6

Throughout this section, we assume that the random variables $S \sim p(s_\ell), X|S \sim p(x_\ell|s_\ell)$ and $Z \sim \mathcal{CN}(0,1)$ are independent of X and S. We start with the following lemma.

Lemma 16. *$\mathbb{P}(S \neq D(X+\sigma Z))$ is continuous in σ.*

Note that $\mathbb{P}(S \neq D(X+\sigma Z)) = \sum_{s\in\mathcal{O}} \mathbb{P}(S \neq D(X+\sigma Z)|S = s)p(S = s)$. Hence, if we prove that $\mathbb{P}(S \neq D(X+\sigma Z)|S = s)$ is continuous then, so is $\mathbb{P}(S \neq D(X+\sigma Z))$. Furthermore, $\mathbb{P}(S \neq D(X+\sigma Z)|S = s) = \mathbb{P}(X+\sigma Z \in D^{-1}(s)|S = s)$. It is straightforward to write the probability in its integral form and confirm that it is continuous in σ.

Suppose that we run AMPI for t iterations and then apply D to $\mathbf{z}^t$ and σ_t to obtain the signal estimate $\hat{s}_\ell^t$. Then, according to Lemma 12, the asymptotic error probability of AMPI is

$$\mathbb{P}(s_\ell \neq \hat{s}_\ell^t) = \mathbb{P}(S \neq D(X+\sigma_t Z)). \tag{78}$$

Also, note that the effective noise variance σ_t is given by the cSE recursion in (16); i.e. for $t \to \infty$, σ_t converges to the solution of AMPI's fixed-point equation as given in (20). Now, since the fixed-point equation of AMPI in (20) coincides with fixed-point equation of the IO data detector in (69), we have $\sigma_t \to \tilde{\sigma}$. The rest of the proof is a simple continuity argument with two statements:

1) Since $\mathbb{P}(S \neq D(X+\sigma Z))$ is a continuous function in σ, for every $\epsilon > 0$ there exists $\Delta\sigma$ such that if $\bar{\sigma} \in (\tilde{\sigma} - \Delta\sigma, \tilde{\sigma} + \Delta\sigma)$, then $\mathbb{P}(S \neq D(X+\bar{\sigma}Z)) < \mathbb{P}(S \neq D(X+\tilde{\sigma}Z)) + \epsilon$.
2) Since $\sigma^t \to \tilde{\sigma}$ as $t \to \infty$, we know that there exists a t_0 such that for $t > t_0$, $\sigma^t < \tilde{\sigma} + \Delta\sigma$.

By combining these two statements, we conclude that for every $\epsilon > 0$, there exists a t_0 such that for $t > t_0$

$$\mathbb{P}(s_\ell \neq \hat{s}_\ell^{t_0}) \overset{(a)}{=} \mathbb{P}(S \neq D(X+\sigma_{t_0}Z)) \tag{79}$$

$$< \mathbb{P}(S \neq D(X+\tilde{\sigma}Z)) + \epsilon \overset{(b)}{=} \mathbb{P}(s_\ell \neq \hat{s}_\ell^{\mathrm{IO}}) + \epsilon. \tag{80}$$

Here, (a) and (b) follow from (78) and Theorem 6, respectively. The proof is complete by averaging over all $\ell = 1, \ldots, N$.

REFERENCES

[1] R. Ghods, C. Jeon, A. Maleki, and C. Studer, "Optimal large-MIMO data detection with transmit impairments," in *53rd Annual Allerton Conference on Communication, Control, and Computing*, Sept. 2015, pp. 1211–1218.

[2] J. G. Proakis and M. Salehi, *Communication systems engineering*, vol. 2.

[3] D. Guo and S. Verdú, "Randomly spread cdma: Asymptotics via statistical physics," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 1983–2010, 2005.

[4] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[5] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[6] M. Wu, B. Yin, A. Vosoughi, C. Studer, J. Cavallaro, and C. Dick, "Approximate matrix inversion for high-throughput data detection in the large-scale MIMO uplink," in *Proc. IEEE Int. Symp. Circuits and Syst. (ISCAS)*, May 2013, pp. 2155–2158.

[7] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimal data detection in large MIMO," *arXiv preprint arXiv:1811.01917*, 2018.

[8] ——, "Optimality of large MIMO detection via approximate message passing," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1227–1231.

[9] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser mimo-ofdm systems using approximate message passing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 902–915, 2014.

[10] D. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Natl. Academy of Sciences (PNAS)*, vol. 106, no. 45, pp. 18 914–18 919, Sept. 2009.

[11] C. Studer, M. Wenk, and A. Burg, "MIMO transmission with residual transmit-RF impairments," in *Int. ITG Workshop on Smart Antennas (WSA)*, Feb. 2010, pp. 189–196.

[12] M. Vehkaperä, T. Riihonen, M. A. Girnyk, E. Björnson, M. Debbah, L. K. Rasmussen, and R. Wichman, "Asymptotic analysis of SU-MIMO channels with transmitter noise and mismatched joint decoding," *IEEE Trans. Commun.*, vol. 32, no. 6, pp. 1065–1082, Mar. 2015.

[13] T. C. Schenk, *RF imperfections in high-rate wireless systems: impact and digital compensation*. Springer Netherlands, 2008.

[14] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[15] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[16] M. Davenport, J. Laska, J. Treichler, and R. Baraniuk, "The pros and cons of compressive sensing for wideband signal acquisition: noise folding versus dynamic range," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4628–4642, Sep. 2012.

[17] J. Treichler, M. Davenport, and R. Baraniuk, "Application of compressive sensing to the design of wideband signal acquisition receivers," *US/Australia Joint Work. Defense Apps. of Signal Processing (DASP), Lihue, Hawaii*, vol. 5, 2009.

[18] E. Arias-Castro and Y. C. Eldar, "Noise folding in compressed sensing," *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 478–481, 2011.

[19] S. Verdú, *Multiuser Detection*, 1st ed. Cambridge University Press, 1998.

[20] G. F. Cooper, "The computational complexity of probabilistic inference using bayesian belief networks," *Artificial intelligence*, vol. 42, no. 2-3, pp. 393–405, 1990.

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

[22] A. Montanari, *Graphical models concepts in compressed sensing, Compressed Sensing (Y.C. Eldar and G. Kutyniok, eds.)*. Cambridge University Press, 2012.

[23] A. Maleki, "Approximate message passing algorithms for compressed sensing," Ph.D. dissertation, Stanford University, Jan. 2011.

[24] D. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Jan. 2010, pp. 1–5.

[25] ——, "Message passing algorithms for compressed sensing: II. Analysis and validation," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Jan. 2010, pp. 1–5.

[26] C. Studer, M. Wenk, and A. Burg, "System-level implications of residual transmit-RF impairments in MIMO systems," in *Proc. European Conf. on Antennas and Propagation (EUCAP)*, Apr. 2011, pp. 2686–2689.

[27] T. C. Schenk, P. F. Smulders, and E. R. Fledderus, "Performance of MIMO OFDM systems in fading channels with additive TX and RX impairments," in *Proc. IEEE BENELUX/DSP Valley Signal Process. Symp.*, Apr. 2005, pp. 41–44.

[28] B. Goransson, S. Grant, E. Larsson, and Z. Feng, "Effect of transmitter and receiver impairments on the performance of MIMO in HSDPA," in *Proc. IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Jul. 2008, pp. 496–500.

[29] H. Suzuki, T. V. A. Tran, I. B. Collings, G. Daniels, and M. Hedley, "Transmitter noise effect on the performance of a MIMO-OFDM hardware implementation achieving improved coverage," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 867–876, Aug. 2008.

[30] H. Suzuki, I. B. Collings, M. Hedley, and G. Daniels, "Practical performance of MIMO-OFDM-LDPC with low complexity double iterative receiver," in *Proc. IEEE Int. Symp. Personal, Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2009, pp. 2469–2473.

[31] J. P. González-Coma, P. M. Castro, and L. Castedo, "Impact of transmit impairments on multiuser MIMO non-linear transceivers," in *Int. ITG Workshop on Smart Antennas (WSA)*, Feb. 2011, pp. 1–8.

[32] ——, "Transmit impairments influence on the performance of MIMO receivers and precoders," in *Proc. European. Wireless Conf. – Sustainable Wireless Technol. (European Wireless)*, Apr. 2011, pp. 1–8.

[33] E. Bjornson, P. Zetterberg, M. Bengtsson, and B. Ottersten, "Capacity limits and multiplexing gains of MIMO channels with transceiver impairments," *IEEE Commun. Lett.*, vol. 17, no. 1, Jan. 2013.

[34] X. Zhang, M. Matthaiou, E. Bjornson, M. Coldrey, and M. Debbah, "On the MIMO capacity with residual transceiver hardware impairments," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 5299–5305.

[35] S. Peter, M. Artina, and M. Fornasier, "Damping noise-folding and enhanced support recovery in compressed sensing," *IEEE Transactions on Signal Processing*, vol. 63, no. 22, pp. 5990–6002, Nov 2015.

[36] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," *CoRR*, vol. abs/1010.5141, 2010.

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

[38] D. L. Donoho, A. Maleki, and A. Montanari, "How to design message passing algorithms for compressed sensing," *preprint*, 2011.

[39] Y. Weiss, "Correctness of local probability propagation in graphical models with loops," *Neural computation*, vol. 12, no. 1, pp. 1–41, 2000.

[40] 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.

[41] L. Zheng, A. Maleki, H. Weng, X. Wang, and T. Long, "Does ℓ_p-minimization outperform ℓ_1-minimization?" *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 6896–6935, 2017.

[42] J. Barbier, N. Macris, M. Dia, and F. Krzakala, "Mutual information and optimality of approximate message-passing in random linear estimation," *IEEE Transactions on Information Theory*, 2020.

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

[44] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[45] A. Mousavi, A. Maleki, R. G. Baraniuk *et al.*, "Consistent parameter estimation for lasso and approximate message passing," *The Annals of Statistics*, vol. 46, no. 1, pp. 119–148, 2018.

[46] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, pp. 1135–1151, 1981.

[47] R. Ghods, C. Jeon, A. Maleki, and C. Studer, "Optimal data detection and signal estimation in systems with input noise," *arXiv preprint arXiv:2008.02337*, 2020.

[48] E. Polak and G. Ribiere, "Note sur la convergence de méthodes de directions conjuguées," *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, vol. 3, no. R1, pp. 35–43, 1969.

[49] P. E. Frandsen, K. Jonasson, H. B. Nielsen, and O. Tingleff, "Unconstrained optimization," 1999.

[50] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[51] D. Guo, Y. Wu, S. Shamai, and S. Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2371–2385, Apr. 2011.

[52] Y. Wu and S. Verdú, "MMSE dimension," *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1463–1467, June 2010.