Optimality of Large MIMO Detection via Approximate Message Passing

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

***Abstract*—Optimal data detection in multiple-input multiple-output (MIMO) communication systems with a large number of antennas at both ends of the wireless link entails prohibitive computational complexity. In order to reduce the computational complexity, a variety of sub-optimal detection algorithms have been proposed in the literature. In this paper, we analyze the optimality of a novel data-detection method for large MIMO systems that relies on approximate message passing (AMP). We show that our algorithm, referred to as individually-optimal (IO) large-MIMO AMP (short IO-LAMA), is able to perform IO data detection given certain conditions on the MIMO system and the constellation set (e.g., QAM or PSK) are met.**

I. Introduction

We consider the problem of recovering the M_T-dimensional data vector $\mathbf{s}_0 \in \mathcal{O}^{M_\text{T}}$ from the noisy multiple-input multiple-output (MIMO) input-output relation $\mathbf{y} = \mathbf{H}\mathbf{s}_0 + \mathbf{n}$, by performing individually-optimal (IO) data detection [2], [3]

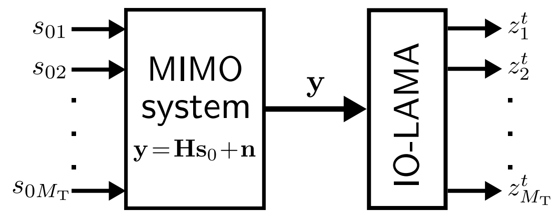
$$\text{(IO)} \quad s_\ell^{\text{IO}} = \underset{\tilde{s}_\ell \in \mathcal{O}}{\arg\max}\; p(\tilde{s}_\ell \,|\, \mathbf{y}, \mathbf{H}).$$

Here, s_ℓ^{IO} denotes the ℓ-th IO estimate, $\mathcal{O}$ is a finite constellation (e.g., QAM or PSK), $p(\tilde{s}_\ell \,|\, \mathbf{y}, \mathbf{H})$ is a probability density function assuming i.i.d. zero-mean complex Gaussian noise for the vector $\mathbf{n} \in \mathbb{C}^{M_\text{R}}$ with variance N_0 per complex dimension, M_T and M_R denotes the number of transmit and receive antennas, respectively, $\mathbf{y} \in \mathbb{C}^{M_\text{R}}$ is the receive vector, and $\mathbf{H} \in \mathbb{C}^{M_\text{R} \times M_\text{T}}$ is the (known) MIMO system matrix. In what follows, we assume that the entries of the MIMO system matrix $\mathbf{H}$ are i.i.d. zero-mean complex Gaussian with variance $1/M_\text{R}$, and we define the so-called *system ratio* as $\beta = M_\text{T}/M_\text{R}$.

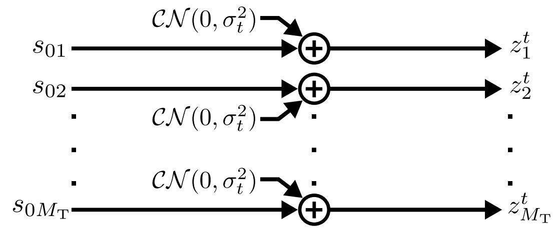
Although IO detection achieves the minimum symbol error-rate [4], the combinatorial nature of the (IO) problem [2], [3] requires prohibitive computational complexity, especially in large (or massive) MIMO systems [4], [5]. In order to enable data detection in such high-dimensional systems, a large number of low-complexity but sub-optimal algorithms have been proposed in the literature (see, e.g., [6]–[8]).

A. Contributions

In this paper, we propose and analyze a novel, computationally efficient data-detection algorithm, referred to as IO-LAMA (short for IO large MIMO approximate message passing). We show that IO-LAMA decouples the noisy MIMO system into a set of independent additive white Gaussian noise (AWGN) channels with equal signal-to-noise ratio (*SNR*); see Fig. 1 for an illustration of this decoupling property. The state-evolution (SE) recursion of AMP enables us to track the effective noise variance σ_t^2 of each decoupled AWGN channel at every algorithm iteration t. Using these results, we provide precise conditions on the MIMO system matrix, the system ratio β, the noise variance N_0, and the modulation scheme for which IO-LAMA *exactly* solves the (IO) problem.



(a) MIMO system with IO-LAMA as the data detector.



(b) Equivalent decoupled system with effective noise variance σ_t^2.

Fig. 1. IO-LAMA decouples large MIMO systems (a) into a set of parallel and independent AWGN channels with equal noise variance; (b) equivalent system in the large-system limit, i.e., for $\beta = M_\text{T}/M_\text{R}$ with $M_\text{T} \to \infty$.

B. Relevant Prior Art

Initial results for IO data detection in large MIMO systems reach back to [9] where Verdú and Shamai analyzed the achievable rates under optimal data detection in randomly-spread CDMA systems. Tanaka [10] derived expressions for the error-rate performance and the multi-user efficiency for IO detection using the replica method. While Tanaka's results were limited to BPSK constellations, Guo and Verdú extended his results to arbitrary discrete input distributions [3], [11]. All these results study the fundamental performance of IO data detection in the large-system limit, i.e., for $\beta = M_\text{T}/M_\text{R}$ with $M_\text{T} \to \infty$. Corresponding practical detection algorithms have been proposed for BPSK constellations [12], [13]—to the best of our knowledge, no computationally efficient algorithms for general constellation sets and complex-valued systems have been proposed in the open literature.

An extended version of this paper including all proofs is in preparation [1].

C. Jeon, R. Ghods, and C. Studer are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY; e-mail: jeon@csl.cornell.edu, rghods@csl.cornell.edu, studer@cornell.edu.

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

This work was supported in part by Xilinx Inc. and by the US National Science Foundation (NSF) under grants ECCS-1408006 and CCF-1420328.

Our data-detection method, IO-LAMA, builds upon approximate message passing (AMP) [14]–[16], which was initially developed for the recovery of sparse signals. AMP has been generalized to arbitrary signal priors in [17]–[19] and enables a precise performance analysis via the SE recursion [14], [15]. Recently, AMP-related algorithms have been proposed for data detection [20]–[22]; these algorithms, however, lack of a theoretical performance analysis.

C. Notation

Lowercase and uppercase boldface letters designate vectors and matrices, respectively. For a matrix $\mathbf{H}$, we define its conjugate transpose to be $\mathbf{H}^{\mathrm{H}}$. The ℓ-th column of $\mathbf{H}$ is denoted by $\mathbf{h}_\ell^{\mathrm{c}}$. We use $\langle\cdot\rangle$ to write $\langle\mathbf{x}\rangle = \frac{1}{N}\sum_{k=1}^{N} x_k$. A multivariate complex-valued Gaussian probability density function (pdf) is denoted by $\mathcal{CN}(\mathbf{m},\mathbf{K})$, where $\mathbf{m}$ is the mean vector and $\mathbf{K}$ the covariance matrix. $\mathbb{E}_X[\cdot]$ and $\mathbb{V}\mathrm{ar}_X[\cdot]$ denotes the expectation and variance operator with respect to the pdf of the random variable X, respectively.

II. IO-LAMA: Large-MIMO Detection using AMP

We now present IO-LAMA and the SE recursion, which is used in Section III for our optimality analysis.

A. The IO-LAMA Algorithm

We assume that the transmit symbols s_ℓ, $\ell = 1,\ldots,M_{\mathrm{T}}$, of the transmit data vector $\mathbf{s}$ are taken from a finite set $\mathcal{O} = \{a_j : j = 1,\ldots,|\mathcal{O}|\}$ with constellation points a_j chosen, e.g., from a QAM or PSK alphabet. We assume an i.i.d. prior $p(\mathbf{s}) = \prod_{\ell=1}^{M_{\mathrm{T}}} p(s_\ell)$, with the following distribution for each transmit symbol s_ℓ:

$$p(s_\ell) = \textstyle\sum_{a\in\mathcal{O}} p_a \delta(s_\ell - a). \tag{1}$$

Here, p_a designates the (known) prior probability of each constellation point $a \in \mathcal{O}$ and $\delta(\cdot)$ is the Dirac delta function; for uniform priors, we have $p_a = |\mathcal{O}|^{-1}$.

The IO-LAMA algorithm summarized below is obtained by using the prior distribution in (1) within complex Bayesian AMP. A detailed derivation of the algorithm is given in [1].

Algorithm 1. *Initialize* $\hat{s}_\ell^1 = \mathbb{E}_S[S]$ *for* $\ell = 1,\ldots,M_T$, $\mathbf{r}^1 = \mathbf{y}$, *and* $\tau^1 = \beta\,\mathbb{V}\mathrm{ar}_S[S]/N_0$. *Then, for every IO-LAMA iteration* $t = 1,2,\ldots,$ *compute the following steps:*

$$\begin{aligned}
\mathbf{z}^t &= \hat{\mathbf{s}}^t + \mathbf{H}^{\mathrm{H}}\mathbf{r}^t\\
\hat{\mathbf{s}}^{t+1} &= \mathsf{F}(\mathbf{z}^t, N_0(1+\tau^t))\\
\tau^{t+1} &= \tfrac{\beta}{N_0}\langle \mathsf{G}(\mathbf{z}^t, N_0(1+\tau^t))\rangle\\
\mathbf{r}^{t+1} &= \mathbf{y} - \mathbf{H}\hat{\mathbf{s}}^{t+1} + \tfrac{\tau^{t+1}}{1+\tau^t}\mathbf{r}^t.
\end{aligned}$$

The functions $\mathsf{F}(s_\ell,\tau)$ *and* $\mathsf{G}(s_\ell,\tau)$ *correspond to the message mean and variance, and are computed as follows:*

$$\begin{aligned}
\mathsf{F}(\hat{s}_\ell,\tau) &= \textstyle\int_{s_\ell} s_\ell f(s_\ell|\hat{s}_\ell,\tau)\mathrm{d}s_\ell\\
\mathsf{G}(\hat{s}_\ell,\tau) &= \textstyle\int_{s_\ell} |s_\ell|^2 f(s_\ell|\hat{s}_\ell,\tau)\mathrm{d}s_\ell - |\mathsf{F}(\hat{s}_\ell,\tau)|^2.
\end{aligned} \tag{2}$$

Here, $f(s_\ell|\hat{s}_\ell,\tau)$ *is the posterior pdf defined by* $f(s_\ell|\hat{s}_\ell,\tau) = \frac{1}{Z}p(\hat{s}_\ell|s_\ell,\tau)p(s_\ell)$ *with* $p(\hat{s}_\ell|s_\ell,\tau) \sim \mathcal{CN}(s_\ell,\tau)$ *and a normalization constant* Z. *Both functions* $\mathsf{F}(\hat{s}_\ell,\tau)$ *and* $\mathsf{G}(\hat{s}_\ell,\tau)$ *operate element-wise on vectors.*

In order to analyze the performance of IO-LAMA in the large-system limit, we next summarize the SE recursion. The SE recursion in the following theorem enables us to track the effective noise variance σ_t^2 for the decoupled MIMO system for every iteration t (cf. Fig. 1), which is key for the optimality analysis in Section III. A detailed derivation is given in [1].

Theorem 1. *Fix the system ratio* $\beta = M_T/M_R$ *and the constellation set* $\mathcal{O}$, *and let* $M_T \to \infty$. *Initialize* $\sigma_1^2 = N_0 + \beta\,\mathbb{V}\mathrm{ar}_S[S]$. *Then, the effective noise variance* σ_t^2 *of IO-LAMA at iteration* t *is given by the following recursion:*

$$\sigma_t^2 = N_0 + \beta\Psi(\sigma_{t-1}^2). \tag{3}$$

The so-called mean-squared error (MSE) function is defined by

$$\Psi(\sigma_{t-1}^2) = \mathbb{E}_{S,Z}\left[\left|\mathsf{F}\big(S + \sigma_{t-1}Z, \sigma_{t-1}^2\big) - S\right|^2\right],$$

where F *is given in (2) and* $Z \sim \mathcal{CN}(0,1)$.

B. IO-LAMA Decouples Large MIMO Systems

In the large-system limit and for every iteration t, IO-LAMA computes the marginal distribution of s_ℓ, $\ell = 1,\ldots,M_{\mathrm{T}}$, which corresponds to a Gaussian distribution centered around the original signal $s_{0\ell}$ with variance σ_t^2. These properties follow from [16, Sec. 6], which shows that $\mathbf{z}^t = \hat{\mathbf{s}}^t + \mathbf{H}^{\mathrm{H}}\mathbf{r}^t$ is distributed according to $\mathcal{CN}(\mathbf{s}_0, \sigma_t^2\mathbf{I}_{M_{\mathrm{T}}})$. Hence, the input-output relation for each transmit stream $z_\ell^t = \hat{s}_\ell^t + (\mathbf{h}_\ell^{\mathrm{c}})^{\mathrm{H}}\mathbf{r}_\ell^t$ is equivalent to the following single-stream AWGN channel:

$$z_\ell^t = s_{0\ell} + n_\ell^t.$$

Here, $s_{0\ell}$ is the ℓ-th original transmitted signal and n_ℓ^t is AWGN with variance σ_t^2 per complex entry. As a consequence, IO-LAMA decouples the MIMO system into M_{T} parallel and independent AWGN channels with equal noise variance σ_t^2 in the large-MIMO limit; see Fig. 1(b) for an illustration.

III. Optimality of IO-LAMA

We now provide conditions for which IO-LAMA *exactly* solves the (IO) problem.

A. Fixed points of IO-LAMA's State Evolution

For $t \to \infty$, the SE recursion in Theorem 1 converges to the following fixed-point equation [1], [15]:

$$\sigma_{\mathrm{IO}}^2 = N_0 + \beta\Psi(\sigma_{\mathrm{IO}}^2), \tag{4}$$

which coincides with the "fixed-point equation" developed for IO detection by Guo and Verdú using the replica method in [3, Eq. (34)]. We note that (4) may have multiple fixed-point solutions. In the case of such non-unique fixed points, Guo and Verdú choose the solution that minimizes the "free energy" [3, Sec. 2-D], whereas IO-LAMA converges, in general, to the fixed-point solution with the largest effective noise variance σ^2. We note that if the fixed-point solution to (4) is unique, then IO-LAMA recovers the solution with minimal effective noise variance σ^2 and thus, performs IO detection. However, if there are multiple fixed-points solutions to (4), IO-LAMA is, in general, sub-optimal and does not necessarily converge to the fixed-point solution with the minimal "free energy."[1] We next

[1] Convergence to another fixed-point solution is possible if IO-LAMA is initialized sufficiently close to such a fixed point; see [1], [23] for the details.

provide conditions for which there is exactly one (unique) fixed point with minimum effective noise variance σ^2 and—as a consequence—IO-LAMA solves the (IO) problem.

B. Exact Recovery Thresholds (ERTs)

We start by analyzing IO-LAMA in the noiseless setting. We provide conditions on the system ratio β and the constellation set $\mathcal{O}$, which guarantee exact recovery of an unknown transmit signal $\mathbf{s}_0 \in \mathcal{O}^{M_\mathrm{T}}$ in the large-system limit, i.e., β is fixed and $M_\mathrm{T} \to \infty$. In particular, we show that if $\beta < \beta_\mathcal{O}^{\max}$, where $\beta_\mathcal{O}^{\max}$ is the so-called *exact recovery threshold (ERT)*, then IO-LAMA perfectly recovers $\mathbf{s}_0$; for $\beta \geq \beta_\mathcal{O}^{\max}$, perfect recovery is not guaranteed, in general.[2] To make this behavior explicit, we need the following technical result; the proof is given in Appendix A.

Lemma 2. *Fix the constellation set $\mathcal{O}$. If $\mathbb{V}\mathrm{ar}_S[S]$ is finite, then there exists a non-negative gap $\sigma^2 - \Psi(\sigma^2) \geq 0$ with equality if and only if $\sigma^2 = 0$. As $\sigma^2 \to 0$, the MSE $\Psi(\sigma^2) \to 0$ and as $\sigma^2 \to \infty$, MSE $\Psi(\sigma^2) \to \mathbb{V}\mathrm{ar}_S[S]$.*

For all $\sigma^2 > 0$, Lemma 2 guarantees that $\Psi(\sigma^2) < \sigma^2$. Suppose that for some $\beta > 1$, $\beta\Psi(\sigma^2) < \sigma^2$ also holds for all $\sigma^2 > 0$. Then, as long as $\beta > 1$ is not too large to also ensure $\beta\Psi(\sigma^2) < \sigma^2$ for all $\sigma^2 > 0$, there will only be a *single* fixed point at $\sigma^2 = 0$. Therefore, LAMA can still perfectly recover the original signal $\mathbf{s}_0$ by Theorem 1 since $\Psi(\sigma^2) = 0$. Leveraging the gap between $\Psi(\sigma^2)$ and σ^2 will allow us to find the exact recovery threshold (ERT) of LAMA for values of $\beta > 1$. For the fixed (discrete) constellation set $\mathcal{O}$, the largest β that ensures $\beta\Psi(\sigma^2) < \sigma^2$ is precisely the ERT defined next.

Definition 1. *Fix $\mathcal{O}$ and let $N_0 = 0$. Then, the exact recovery threshold (ERT) that enables perfect recovery of the original signal $\mathbf{s}_0$ using IO-LAMA is given by*

$$\beta_\mathcal{O}^{\max} = \min_{\sigma^2 > 0} \left\{ \left(\frac{\Psi(\sigma^2)}{\sigma^2} \right)^{-1} \right\}. \tag{5}$$

With Definition 1, we state Theorem 3, which establishes optimality in the noiseless case; the proof is given in Appendix B.

Theorem 3. *Let $N_0 = 0$ and fix a discrete set $\mathcal{O}$. If $\beta < \beta_\mathcal{O}^{\max}$, then IO-LAMA perfectly recovers the original signal $\mathbf{s}_0$ from $\mathbf{y} = \mathbf{H}\mathbf{s}_0 + \mathbf{n}$ in the large system limit.*

Note that for a given constellation set $\mathcal{O}$, the ERT $\beta_\mathcal{O}^{\max}$ can be computed numerically using (5). Furthermore, the signal variance, $\mathbb{V}\mathrm{ar}_S[S]$, has no impact on the ERT as the MSE function $\Psi(\sigma^2)$ and σ^2 scale linearly with $\mathbb{V}\mathrm{ar}_S[S]$. Table I summarizes ERTs $\beta_\mathcal{O}^{\max}$ for common QAM and PSK constellation sets.

C. Optimality Conditions for IO-LAMA

We now study the optimality of IO-LAMA in the presence of noise, where *exact* recovery is no longer guaranteed. In particular, we provide conditions for which IO-LAMA converges to the fixed point with minimal effective noise variance σ^2, which corresponds to solving the (IO) problem.

[2] We assume the initialization in Algorithm 1. IO-LAMA may recover the original signal for $\beta \geq \beta_\mathcal{O}^{\max}$ if initialized appropriately; see, e.g., [23].

TABLE I

ERTs $\beta_\mathcal{O}^{\mathrm{MAX}}$, MRTs $\beta_\mathcal{O}^{\mathrm{MIN}}$, and critical noise levels $N_0^{\min}(\beta_\mathcal{O}^{\min})$ and $N_0^{\max}(\beta_\mathcal{O}^{\max})$ of IO-LAMA for common constellation sets

Constellation	$\beta_\mathcal{O}^{\min}$	$N_0^{\min}(\beta_\mathcal{O}^{\min})$	$\beta_\mathcal{O}^{\max}$	$N_0^{\max}(\beta_\mathcal{O}^{\max})$
BPSK	2.9505	$2.999 \cdot 10^{-1}$	4.1709	$2.432 \cdot 10^{-1}$
QPSK	1.4752	$1.499 \cdot 10^{-1}$	2.0855	$1.216 \cdot 10^{-1}$
16-QAM	0.9830	$3.000 \cdot 10^{-2}$	1.3629	$2.454 \cdot 10^{-2}$
64-QAM	0.8424	$7.144 \cdot 10^{-3}$	1.1573	$5.868 \cdot 10^{-3}$
8-PSK	1.4576	$4.440 \cdot 10^{-2}$	1.8038	$3.826 \cdot 10^{-2}$
16-PSK	1.4728	$1.143 \cdot 10^{-2}$	1.8005	$9.953 \cdot 10^{-3}$

TABLE II

Summary of (sub-)optimality regimes of IO-LAMA

	$\beta \leq \beta_\mathcal{O}^{\min}$	$\beta_\mathcal{O}^{\min} < \beta < \beta_\mathcal{O}^{\max}$	$\beta_\mathcal{O}^{\max} \leq \beta$
$N_0 < N_0^{\min}(\beta)$	*optimal*	*optimal*	suboptimal
$N_0^{\min}(\beta) \leq N_0 \leq N_0^{\max}(\beta)$	*optimal*	(sub-)optimal[3]	suboptimal
$N_0^{\max}(\beta) < N_0$	*optimal*	*optimal*	*optimal*

Note that such a minimum free-energy solution is also the fixed point for the IO detector in [3, Eq. (34)]. We call the fixed point with minimum effective noise variance *optimal fixed point*; other fixed points are called *suboptimal fixed points*.

We identify three different operation regimes for IO-LAMA depending on the system ratio β (see Table II). To make these three regimes explicit, we need the following definition.

Definition 2. *Fix the constellation set $\mathcal{O}$. Then, the minimum recovery threshold (MRT) $\beta_\mathcal{O}^{\min}$ is defined by*

$$\beta_\mathcal{O}^{\min} = \min_{\sigma^2 > 0} \left\{ \left(\frac{\mathrm{d}\Psi(\sigma^2)}{\mathrm{d}\sigma^2} \right)^{-1} \right\}. \tag{6}$$

The definition of MRT shows that for all system ratios $\beta \leq \beta_\mathcal{O}^{\min}$, the fixed point of (4) is unique. The following lemma establishes a fundamental relationship between MRT and ERT; the proof is given in Appendix C.

Lemma 4. *The MRT never exceeds the ERT.*

We next define the minimum critical and maximum guaranteed noise variance, $N_0^{\min}(\beta)$ and $N_0^{\max}(\beta)$, that determine boundaries for the optimality regimes when $\beta > \beta_\mathcal{O}^{\min}$.

Definition 3. *Fix $\beta \in (\beta_\mathcal{O}^{\min}, \beta_\mathcal{O}^{\max})$. Then, the* minimum critical noise *$N_0^{\min}(\beta)$ that ensures convergence to the optimal fixed point is defined by*

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

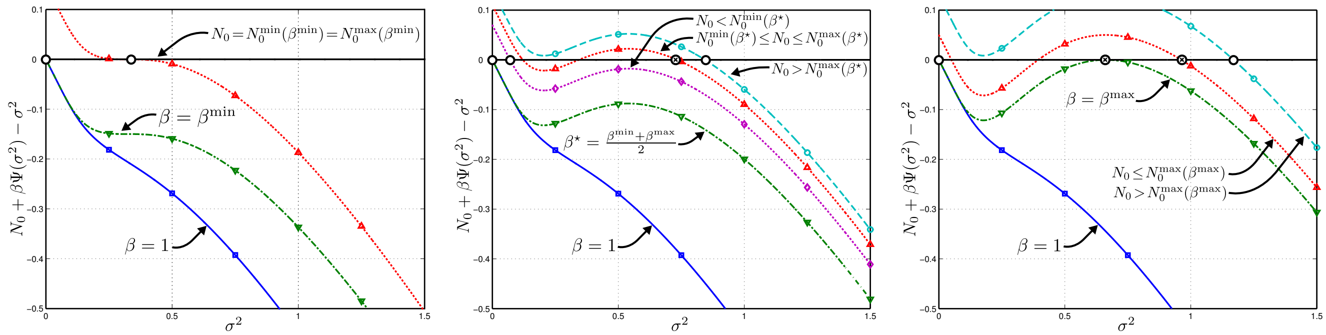
Definition 4. *Fix $\beta > \beta_\mathcal{O}^{\min}$. Then, the* maximum guaranteed noise *$N_0^{\max}(\beta)$ that ensures convergence to the optimal fixed point is defined by*

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

We recall that all the zero crossings of the function

$$g(\sigma^2, \beta, N_0)_\mathcal{O} = N_0 + \beta\Psi(\sigma^2) - \sigma^2 \tag{7}$$

[3] For certain constellation sets (e.g., 16-PSK), there exist sub-intervals in $[N_0^{\min}(\beta), N_0^{\max}(\beta)]$ where IO-LAMA is still optimal; see [1] for the details.



(a) $\beta \leq \beta_{\mathcal{O}}^{\min}$: IO-LAMA always converges to the unique, optimal fixed point (FP) irrespective of N_0.

(b) $\beta \in (\beta_{\mathcal{O}}^{\min}, \beta_{\mathcal{O}}^{\max})$: IO-LAMA converges to the optimal FP if $N_0 < N_0^{\min}(\beta)$ or $N_0 > N_0^{\max}(\beta)$.

(c) $\beta \geq \beta_{\mathcal{O}}^{\max}$: IO-LAMA converges to the optimal fixed point if $N_0 > N_0^{\max}(\beta)$.

Fig. 2. Plot of the function (7) for three regimes (a) $\beta \leq \beta_{\mathcal{O}}^{\min}$, (b) $\beta \in (\beta_{\mathcal{O}}^{\min}, \beta_{\mathcal{O}}^{\max})$, and (c) $\beta \geq \beta_{\mathcal{O}}^{\max}$ for QPSK modulation, uniform priors, and $\mathbb{V}\text{ar}_S[S] = E_s = 1$. The optimal fixed points are designated by $\circ$; suboptimal fixed points are designated by $\otimes$.

correspond to all fixed points of the SE recursion of IO-LAMA; we use this function to study the algorithm's optimality.

Figure 2 illustrates our optimality analysis for a large-MIMO system with QPSK constellations. We show (7) depending on the effective noise variance σ^2 and for different system ratios β. The regimes $\beta \leq \beta_{\mathcal{O}}^{\min}$, $\beta \in (\beta_{\mathcal{O}}^{\min}, \beta_{\mathcal{O}}^{\max})$, and $\beta \geq \beta_{\mathcal{O}}^{\max}$ are shown in Fig. 2(a), Fig. 2(b), and Fig. 2(c), respectively. The special case for $\beta = 1$ with $N_0 = 0$ corresponds to the solid blue line, along with the corresponding (unique) fixed point at the origin. In the following three paragraphs, we discuss the three operation regimes of IO-LAMA in detail.

(i) $\beta \leq \beta_{\mathcal{O}}^{\min}$: In this region, the SE recursion of IO-LAMA always converges to the unique, optimal fixed point. For $\beta < \beta_{\mathcal{O}}^{\min}$, the slope of (7) for all σ^2 is strictly-negative. Hence, as (7) is always decreasing, there exists exactly one unique fixed point of the SE recursion regardless of the noise variance N_0. Thus, IO-LAMA converges to the optimal fixed point and consequently, solves the (IO) problem.

We emphasize that we still obtain exactly one fixed point even when β is equal to the MRT. Since $\beta = \beta_{\mathcal{O}}^{\min}$, there exists at least one $\sigma_\star^2$ that satisfies $\beta_{\mathcal{O}}^{\min} \frac{\mathrm{d}}{\mathrm{d}\sigma^2}\Psi(\sigma^2)\big|_{\sigma^2=\sigma_\star^2} = 1$. By definition of $\beta_{\mathcal{O}}^{\min}$, (7) at $\sigma_\star^2$ implies that $\sigma_\star^2$ is a saddle-point, so (7) has exactly one zero at $\sigma_\star^2$. We observe that if $\sigma_\star^2$ is unique, then $N_0^{\min}(\beta_{\mathcal{O}}^{\min}) = N_0^{\max}(\beta_{\mathcal{O}}^{\min})$. For all other $\sigma^2 \neq \sigma_\star^2$, the construction of $\sigma_\star^2$ implies that $\beta_{\mathcal{O}}^{\min} \frac{\mathrm{d}}{\mathrm{d}\sigma^2}\Psi(\sigma^2) < 1$, so the fixed point of (7) remains to be unique.

The green, dash-dotted and red, dotted line in Fig. 2(a) shows (7) for $\beta = \beta_{\mathcal{O}}^{\min}$ with $N_0 = 0$ and $N_0 = N_0^{\min}(\beta_{\mathcal{O}}^{\min}) = N_0^{\max}(\beta_{\mathcal{O}}^{\min})$, respectively. In both cases, we see that the SE recursion of IO-LAMA converges to the unique fixed point.

(ii) $\beta_{\mathcal{O}}^{\min} < \beta < \beta_{\mathcal{O}}^{\max}$: In this region, the SE recursion of IO-LAMA converges to the unique, optimal fixed point if $N_0 < N_0^{\min}(\beta)$ or $N_0 > N_0^{\max}(\beta)$.

The green, dash-dotted line, cyan, dashed line, and magenta, dotted line in Fig. 2(b) shows (7) for $\beta^\star = (\beta_{\mathcal{O}}^{\min} + \beta_{\mathcal{O}}^{\max})/2$ with $N_0 = 0$, $N_0 > N_0^{\max}(\beta^\star)$ and $N_0 < N_0^{\min}(\beta^\star)$, respectively. We note that for the three cases, the fixed point is unique, labeled in Fig. 2(b) by a circle. On the other hand, the red, dotted line in Fig. 2(b) shows (7) with $\beta^\star$ under noise $N_0 \in [N_0^{\min}(\beta^\star), N_0^{\max}(\beta^\star)]$. In this case, however, we observe that SE recursion of IO-LAMA converges to the rightmost suboptimal fixed point labeled by the crossed circle $\otimes$. Hence, IO-LAMA does not, in general, solve the (IO) problem when $N_0^{\min}(\beta) \leq N_0 \leq N_0^{\max}(\beta)$.

(iii) $\beta \geq \beta_{\mathcal{O}}^{\max}$: In this region, the SE recursion of IO-LAMA converges to the unique, optimal fixed point when $N_0 > N_0^{\max}(\beta)$. As $\beta \to \beta_{\mathcal{O}}^{\max}$, the low noise $N_0 < N_0^{\min}(\beta)$ (or high *SNR*) region of optimality disappears because $N_0^{\min}(\beta) \to 0$ as $\beta \to \beta_{\mathcal{O}}^{\max}$ from (5).

The green, dash-dotted line and red, dotted line in Fig. 2(c) shows (7) for $\beta = \beta_{\mathcal{O}}^{\max}$ with $N_0 = 0$ and $0 < N_0 \leq N_0^{\max}(\beta)$, respectively. We observe that the SE recursion of IO-LAMA converges to the suboptimal fixed point when $\beta = \beta_{\mathcal{O}}^{\max}$ even with $N_0 = 0$. On the other hand, the cyan, dashed line refers to (7) for $\beta = \beta_{\mathcal{O}}^{\max}$ with $N_0 > N_0^{\max}(\beta)$. While the noiseless case resulted the SE recursion of IO-LAMA to converge to the suboptimal fixed point, we observe that for strong noise (or equivalently low *SNR*), the SE recursion of IO-LAMA actually recovers the IO solution. Therefore, when $\beta \geq \beta_{\mathcal{O}}^{\max}$, IO-LAMA solves the (IO) problem when the noise is greater than the maximum guaranteed noise $N_0^{\max}(\beta)$.

As a final remark, we note that the ERT $\beta_{\mathcal{O}}^{\max}$ and MRT $\beta_{\mathcal{O}}^{\min}$ in Table I do not depend on $\mathbb{V}\text{ar}_S[S]$; the critical noise levels $N_0^{\min}(\beta)$ and $N_0^{\max}(\beta)$, however, depend on $\mathbb{V}\text{ar}_S[S]$.

D. ERT, MRT, and Critical Noise Levels

The ERT, MRT, as well as $N_0^{\min}(\beta)$ and $N_0^{\max}(\beta)$ for common constellations are summarized in Table I. We assume equally likely priors with the transmit signal normalized to $E_s = \mathbb{V}\text{ar}_S[S] = 1$.[4] We note that the calculations of ERT and MRT for the simplest case of BPSK constellations involve computations of a logistic-normal integral for which no closed-form expression is known [24]. Consequently, the following results were obtained via numerical integration for computing the MSE function $\Psi(\sigma^2)$. As noted in Table I for a QPSK system under complex-valued noise, the ERT is $\beta_{\text{QPSK}}^{\max} \approx 2.0855$, and the MRT is given as $\beta_{\text{QPSK}}^{\min} \approx 1.4752$.

[4]The critical noise levels depend linearly on E_s. Hence, we assume that $E_s = 1$ without loss of generality.

The MRTs for 16-QAM and 64-QAM indicate that small system ratios $\beta < 1$ are required to always guarantee that IO-LAMA solves the (IO) problem in the presence of noise. For instance, we require $\beta \leq \beta^{\min}_{\text{64-QAM}} \approx 0.8424$, i.e. $M_{\text{T}} \leq 0.8424 M_{\text{R}}$, to ensure that IO-LAMA solves the IO problem for 64-QAM in the large system limit. As $\beta \to \beta^{\max}_{\text{64-QAM}} \approx 1.1573$, IO-LAMA is only optimal for $N_0 > N_0^{\max}(\beta^{\max}_{\text{64-QAM}}) \approx 5.868 \cdot 10^{-3}$. From Table I, we see that IO-LAMA is a suitable candidate algorithm for the detection of higher-order QAM constellations in massive multi-user MIMO systems as one typically assumes $M_{\text{R}} \gg M_{\text{T}}$ [25].

IV. Conclusions

We have presented the IO-LAMA algorithm along with the state-evolution recursion. Using these results, we have established conditions on the MIMO system matrix, the noise variance N_0, and the constellation set for which IO-LAMA exactly solves the (IO) problem. While the presented results are exclusively for the large-system limit, our own simulations indicate that IO-LAMA achieves near-optimal performance in realistic, finite-dimensional systems; see [1] for more details.

Appendix A
Proof of Lemma 2

Since the variance of S is finite, denote $\mathbb{V}\text{ar}_S[S] = \sigma_s^2$. By [26, Prop. 5], we have the following upper bound:

$$\Psi(\sigma^2) \leq \frac{\sigma_s^2}{\sigma_s^2 + \sigma^2}\sigma^2 = \frac{1}{1 + \sigma^2/\sigma_s^2}\sigma^2. \tag{8}$$

Here, equality holds for all σ^2 if and only if S is complex normal with variance σ_s^2 [26]. Note that if $\sigma^2 = 0$, then (8) is achieved for any σ_s^2. If $\sigma^2 > 0$, then $\Psi(\sigma^2) < \sigma^2$ by (8).

The first part follows directly from (8) as $\Psi(\sigma^2)$ is non-negative. The second part requires one to realize that $\sigma^2 \to \infty$ also implies $\mathsf{F}(\cdot, \sigma^2) \to \sum_{a \in \mathcal{O}} a p_a = \mathbb{E}_S[S]$, and hence,

$$\lim_{\sigma^2 \to \infty} \Psi(\sigma^2) \to \mathbb{E}_S\left[|S - \mathbb{E}_S[S]|^2\right] = \mathbb{V}\text{ar}_S[S].$$

Appendix B
Proof of Theorem 3

We assume the initialization in Algorithm 1. Since $N_0 = 0$, if LAMA perfectly recovers the original signal $\mathbf{s}_0$, then the fixed point in (4) is unique at $\sigma^2 = 0$. This happens if the system ratio is strictly less than the ERT $\beta^{\max}_{\mathcal{O}}$ because otherwise, i.e., $\beta \geq \beta^{\max}_{\mathcal{O}}$, there exists a non-unique fixed point to (4) for some $\sigma^2 > 0$ by Definition 1.

Appendix C
Proof of Lemma 4

We show that under a fixed constellation set $\mathcal{O}$, $\beta^{\min}_{\mathcal{O}} \leq \beta^{\max}_{\mathcal{O}}$. The proof is straightforward as,

$$\begin{aligned}\beta^{\min}_{\mathcal{O}} &\overset{(a)}{=} \min_{\sigma^2 > 0}\left\{\left(\frac{\mathrm{d}\Psi(\sigma^2)}{\mathrm{d}\sigma^2}\right)^{-1}\right\} \leq \left(\frac{\mathrm{d}\Psi(\sigma^2)}{\mathrm{d}\sigma^2}\right)^{-1}\Big|_{\sigma^2 = \beta^{\max}_{\mathcal{O}}\Psi(\sigma^2)} \\ &\overset{(b)}{=} \left(\frac{1}{\beta^{\max}_{\mathcal{O}}}\right)^{-1} = \beta^{\max}_{\mathcal{O}},\end{aligned}$$

where (a) and (b) follow from the MRT and ERT definitions.

References

[1] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimal data detection in large MIMO," *in preparation for a journal*.

[2] D. Guo and S. Verdú, "Multiuser detection and statistical mechanics," in *Commun., Inform. and Netw. Security*. Springer, 2003, pp. 229–277.

[3] ——, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983–2010, Jun. 2005.

[4] S. Verdú, *Multiuser Detection*, 1st ed. New York, NY, USA: Cambridge University Press, 1998.

[5] D. Seethaler, J. Jaldén, C. Studer, and H. Bölcskei, "On the complexity distribution of sphere decoding," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 5754–5768, Sept. 2011.

[6] T. Narasimhan and A. Chockalingam, "Channel hardening-exploiting message passing (CHEMP) receiver in large-scale MIMO systems," *IEEE J. Sel. Topics Sig. Proc. (J-STSP)*, vol. 8, no. 5, pp. 847–860, Oct. 2014.

[7] M. Čirkić and E. Larsson, "SUMIS: Near-optimal soft-in soft-out MIMO detection with low and fixed complexity," *IEEE Trans. Sig. Proc.*, vol. 62, no. 12, pp. 3084–3097, Jun. 2014.

[8] M. Wu, B. Yin, G. Wang, C. Dick, J. Cavallaro, and C. Studer, "Large-scale MIMO detection for 3GPP LTE: Algorithm and FPGA implementation," *IEEE J. Sel. Topics Sig. Proc. (J-STSP)*, vol. 8, no. 5, pp. 916–929, Oct. 2014.

[9] S. Verdú and S. Shamai, "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 622–640, Mar 1999.

[10] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

[11] D. Guo and C.-C. Wang, "Random sparse linear systems observed via arbitrary channels: A decoupling principle," in *Proc. IEEE Int'l. Symp. Inf. Theory (ISIT)*, June 2007, pp. 946–950.

[12] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," *J. Phys. A: Math. Gen.*, vol. 36, no. 43, pp. 11 111–11 121, Oct. 2003.

[13] G. Caire, R. Müller, and T. Tanaka, "Iterative multiuser joint decoding: optimal power allocation and low-complexity implementation," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1950–1973, Sept. 2004.

[14] 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.

[15] 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.

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

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

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

[19] A. Maleki, L. Anitori, Z. Yang, and R. Baraniuk, "Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP)," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4290–4308, July 2013.

[20] M. Nassar, P. Schniter, and B. Evans, "A factor graph approach to joint OFDM channel estimation and decoding in impulsive noise environments," *IEEE Trans. Sig. Proc.*, vol. 62, no. 6, pp. 1576–1589, Mar. 2014.

[21] P. Schniter, "A message-passing receiver for BICM-OFDM over unknown clustered-sparse channels," *Proc. Intl. Workshop IEEE Sig. Proc. Adv. Wireless Commun. (SPAWC)*, pp. 246–250, Jun. 2011.

[22] 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 J. Sel. Topics Sig. Proc. (J-STSP)*, vol. 8, no. 5, pp. 902–915, Oct. 2014.

[23] L. Zheng, A. Maleki, X. Wang, and T. Long, "Does ℓ_p-minimization outperform ℓ_1-minimization?" *arXiv preprint: 1501.03704*, Jan. 2015.

[24] E. Demidenko, *Generalized Linear Mixed Models*. John Wiley & Sons, Inc., 2005, pp. 329–430.

[25] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[26] 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.