On the Performance of Mismatched Data Detection in Large MIMO Systems

Charles Jeon, Arian Maleki, and Christoph Studer

Abstract—We investigate the performance of mismatched data detection in large multiple-input multiple-output (MIMO) systems, where the prior distribution of the transmit signal used in the data detector differs from the true prior. To minimize the performance loss caused by this prior mismatch, we include a tuning stage into our recently-proposed large MIMO approximate message passing (LAMA) algorithm, which allows us to develop mismatched LAMA algorithms with optimal as well as sub-optimal tuning. We show that carefully-selected priors often enable simpler and computationally more efficient algorithms compared to LAMA with the true prior while achieving near-optimal performance. A performance analysis of our algorithms for a Gaussian prior and a uniform prior within a hypercube covering the QAM constellation recovers classical and recent results on linear and non-linear MIMO data detection, respectively.

I. INTRODUCTION

Data detection in multiple-input multiple-output (MIMO) systems deals with the recovery of the transmit data vector $\mathbf{s}_0 \in \mathcal{O}^{M_\mathrm{T}}$, where $\mathcal{O}$ is a finite constellation (e.g., QAM or PSK), from the noisy input-output relation $\mathbf{y} = \mathbf{H}\mathbf{s}_0 + \mathbf{n}$. In what follows, M_T and M_R denotes the number of transmit and receive antennas, respectively, $\mathbf{y} \in \mathbb{C}^{M_\mathrm{R}}$ is the receive vector, $\mathbf{H} \in \mathbb{C}^{M_\mathrm{R} \times M_\mathrm{T}}$ is the MIMO system matrix, and $\mathbf{n} \in \mathbb{C}^{M_\mathrm{R}}$ is i.i.d. circularly symmetric complex Gaussian noise. To minimize the symbol-error rate, we are interested in solving the individually-optimal (IO) data detection problem [2]–[4]

$$(\text{IO}) \quad s_\ell^{\mathrm{IO}} = \arg \max_{\tilde{s}_\ell \in \mathcal{O}} p(\tilde{s}_\ell \,|\, \mathbf{y}, \mathbf{H}), \; \ell = 1, \ldots, M_\mathrm{T}.$$

Here, s_ℓ^{IO} denotes the ℓ-th IO estimate and $p(\tilde{s}_\ell \,|\, \mathbf{y}, \mathbf{H})$ is the conditional probability density function of $\tilde{s}_\ell \in \mathcal{O}$ given the receive vector $\mathbf{y}$ and the channel matrix $\mathbf{H}$.

The problem (IO) is known to be of combinatorial nature [2]–[4], and the use of an exhaustive search or sphere-decoding methods results in prohibitive complexity for systems where M_T is large [5]. In contrast, our recently proposed algorithm referred to as large MIMO approximate message passing (LAMA) [6], achieves IO performance using a simple iterative procedure in the large-system limit, i.e., where we fix the system ratio $\beta = M_\mathrm{T}/M_\mathrm{R}$ and let $M_\mathrm{T} \to \infty$. For finite-dimensional systems, LAMA was shown to deliver near-IO performance at low computational complexity [6]. Despite all these advantages, LAMA requires repeated computations of transcendental functions that exhibit a high dynamic range.

C. Jeon and C. Studer are with the School of ECE at Cornell University, Ithaca, NY; e-mails: jeon@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.

These computations render the design of corresponding high-throughput hardware designs that rely on finite precision (e.g., fixed-point) arithmetic a challenging task.

A. Contributions

We propose a mismatched version of the LAMA algorithm (short M-LAMA), which enables the design of hardware-friendly data detectors that achieve near-IO performance. We first develop a mismatched version of the complex Bayesian approximate message passing (cB-AMP) algorithm [7] that includes a tuning stage to minimize the performance loss caused by the mismatch in the signal prior. We propose the mismatched state-evolution (SE) framework to enable a performance analysis in the large-system limit. We then apply our framework to mismatched data detection in large MIMO systems by considering two mismatched prior distributions: (i) a Gaussian prior and (ii) a uniform prior within a hypercube covering the QAM constellation. We analyze the performance of the resulting mismatched algorithms in the large-system limit and demonstrate their efficacy in finite-dimensional systems.

B. Relevant prior art

The performance of zero forcing (ZF) and minimum mean-square error (MMSE) detection, which are both well-known instances of mismatched detection algorithms, has been investigated in [8]–[10] for the large-system limit. The use of a uniform prior within a hypercube leads to an alternative mismatched detector for antipodal (e.g., BPSK) signals [11]–[13]. Corresponding theoretical results in [14], [15] for noiseless systems revealed that a system ratio of $\beta < 2$ enables perfect signal recovery. The error-rate performance for the noisy case was derived recently in [13]. The analysis of the mismatched algorithms presented in our paper recovers all these results.

The proposed M-LAMA algorithm relies upon approximate message passing (AMP) [16]–[18], developed for sparse signal recovery. The case of mismatched estimation of sparse signals via AMP was studied in [19], [20], where the performance of AMP was analyzed when the true prior is unknown. This AMP algorithm includes a tuning stage for automated, optimal parameter selection that minimizes the output mean-squared error (MSE). The key differences between M-LAMA and the results in [19], [20] are that (i) we consider MIMO data detection and (ii) we know the true signal prior and intentionally select a mismatched prior in order to design hardware-friendly data detection algorithms that enable near-IO performance.

C. Notation

Lowercase and uppercase boldface letters designate vectors and matrices, respectively. We define the adjoint of matrix $\mathbf{H}$ as

$\mathbf{H}^{\mathsf{H}}$ and use $\langle \cdot \rangle$ to abbreviate $\langle \mathbf{x} \rangle = \frac{1}{N} \sum_{k=1}^{N} x_k$. 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 mean and variance with respect to the random variable X, respectively.

II. MISMATCHED COMPLEX BAYESIAN AMP

We start by presenting a mismatched version of the complex Bayesian approximate message passing (cB-AMP) algorithm [7] (short mcB-AMP), which enables the use a different prior distribution $\tilde{p}(\tilde{\mathbf{s}})$ than the true signal prior $p(\mathbf{s}_0)$.

A. The mismatched cB-AMP (mcB-AMP) algorithm

Given an i.i.d. prior distribution $p(\mathbf{s}_0) = \prod_{\ell=1}^{N} p(s_{0\ell})$ of the true signal $\mathbf{s}_0$ and a mismatched prior $\tilde{p}(\tilde{\mathbf{s}}) = \prod_{\ell=1}^{N} \tilde{p}(\tilde{s}_\ell)$, the proposed mcB-AMP algorithm corresponds to

$$\tilde{\sigma}_t^2 = \frac{1}{M_{\mathsf{R}}} \|\mathbf{r}^t\|_2^2, \tag{1}$$

$$\tau^t = \arg\min_{\tau \geq 0} \mathbb{E}_{S_0, Z}\left[\left| \mathsf{F}^{\mathsf{mm}}(S_0 + \tilde{\sigma}_t Z, \tau) - S_0 \right|^2 \right], \tag{2}$$

$$\mathbf{s}^{t+1} = \mathsf{F}^{\mathsf{mm}}\left(\mathbf{s}^t + \mathbf{H}^{\mathsf{H}}\mathbf{r}^t, \tau^t\right), \tag{3}$$

$$\mathbf{r}^{t+1} = \mathbf{y} - \mathbf{H}\mathbf{s}^{t+1} + \beta\mathbf{r}^t\langle \mathsf{F}'^{\mathsf{mm}}(\mathbf{s}^t + \mathbf{H}^{\mathsf{H}}\mathbf{r}^t, \tau^t)\rangle,$$

which is carried out for $t_{\max}$ iterations $t = 1, \ldots, t_{\max}$. The algorithm is initialized by $s_\ell^1 = \mathbb{E}_{S_0}[S_0]$ for all $\ell = 1, \ldots, M_{\mathsf{T}}$, where $S_0 \sim p(s_0)$, $\mathbf{r}^1 = \mathbf{y} - \mathbf{H}\mathbf{s}^1$, and $\mathsf{F}'^{\mathsf{mm}}$ is the derivative of F^{mm} taken by its first argument. The function

$$\mathsf{F}^{\mathsf{mm}}(s_\ell, \tau) = \mathbb{E}_{\tilde{S}}[\tilde{S}|s_\ell] = \int_{\mathbb{C}} \tilde{s}\tilde{p}(\tilde{s}|s_\ell, \tau)\mathrm{d}\tilde{s}, \tag{4}$$

is the posterior mean with respect to the mismatched prior $\tilde{p}(\tilde{s}_\ell)$ and the variance parameter τ, $p(\tilde{s}|s_\ell, \tau) = \frac{1}{C}p(s_\ell|\tilde{s}, \tau)\tilde{p}(\tilde{s})$, C is a normalization constant, and $p(s_\ell|\tilde{s}, \tau) \sim \mathcal{CN}(\tilde{s}, \tau)$. The functions $\mathsf{F}^{\mathsf{mm}}(s_\ell, \tau)$ and $\mathsf{F}'^{\mathsf{mm}}$ operate element-wise on vectors and the expectation in (2) is taken with respect to the true prior distribution of $S_0 \sim p(s_0)$ and $Z \sim \mathcal{CN}(0, 1)$.

The mcB-AMP algorithm differs from the original cB-AMP algorithm derived in [7] by the additional steps (1) and (2). In every iteration, step (1) estimates the so-called decoupled noise variance σ_t^2 (see Section II-B) and step (2) tunes the variance parameter τ^t based on the estimate σ_t^2. The tuning stage (2) ensures that mcB-AMP converges to the solution that minimizes σ_t^2 for every iteration (see Section II-C).

B. Decoupling property of AMP-based algorithms

We will frequently use the following definition.

Definition 1. *Assume a MIMO system with M_{T} transmit and M_{R} receive antennas, and let the entries of $\mathbf{H}$ be i.i.d. $\mathcal{CN}(0, 1/M_{\mathsf{R}})$. We define the* large-system limit *by fixing the system ratio $\beta = M_{\mathsf{T}}/M_{\mathsf{R}}$ and letting $M_{\mathsf{T}} \to \infty$.*

As shown in [7], [17], [21], AMP-based algorithms decouple the MIMO system into parallel AWGN channels in the large-system limit, i.e., the quantity $\mathbf{z}^t = \mathbf{s}^t + \mathbf{H}^{\mathsf{H}}\mathbf{r}^t$ can be expressed equivalently as $\mathbf{s}_0 + \mathbf{w}^t$, where $\mathbf{w}^t \sim \mathcal{CN}(0, \sigma_t^2 \mathbf{I}_{M_{\mathsf{T}}})$ and σ_t^2 is the decoupled noise variance. A key property of AMP-based algorithms is that the decoupled noise variance σ_t^2 can be tracked *exactly* by the state evolution (SE) framework. We formalize the *mismatched* SE framework for mcB-AMP

in Theorem 1. This result is a specific instance of the SE framework in [17] and has been adapted to mcB-AMP.

Theorem 1. *Assume the large-system limit and that F^{mm} is Lipschitz continuous. Then, the decoupled noise variance σ_{t+1}^2 after t mcB-AMP iterations is given by the coupled recursion*

$$\gamma_t^2 = \arg\min_{\gamma^2 \geq 0} \Psi^{mm}(\sigma_t^2, \gamma^2), \tag{5}$$

$$\sigma_{t+1}^2 = N_0 + \beta\Psi^{mm}(\sigma_t^2, \gamma_t^2), \tag{6}$$

which is initialized by $\sigma_1^2 = N_0 + \beta \mathbb{V}\mathrm{ar}_{S_0}[S_0]$. Here, $S_0 \sim p(s_0)$ and the MSE function is defined by $\Psi^{mm}(\sigma_t^2, \gamma_t^2) = \mathbb{E}_{S_0, Z}\left[\left|\mathsf{F}^{mm}(S_0 + \sigma_t Z, \gamma_t^2) - S_0\right|^2\right]$, where the expectation is taken with respect to S_0 and $Z \sim \mathcal{CN}(0, 1)$.

If the true prior is identical to the mismatched prior, i.e., $p(\mathbf{s}_0) = \tilde{p}(\tilde{\mathbf{s}})$, we can show that mcB-AMP delivers the same decoupled noise variance as cB-AMP without prior mismatch.

Lemma 2. *If $p(\mathbf{s}_0) = \tilde{p}(\tilde{\mathbf{s}})$, then the decoupled noise variance σ_{t+1}^2 of mcB-AMP is equivalent to σ_{t+1}^2 of cB-AMP [7].*

The proof of Lemma 2 follows by noting that the MSE function $\Psi^{\mathsf{mm}}(\sigma_t^2, \gamma^2)$ in (6) is minimized by $\gamma^2 = \sigma_t^2$ in (5), and reduces to the conditional variance [22]. Therefore, (6) reduces to the conventional SE recursion of cB-AMP [7].

C. Optimal tuning of the variance parameter τ

Before we discuss the tuning stage (2) in detail, we formalize what we mean by optimal tuning of the variance parameter τ^t. For all iterations $t = 1, \ldots, t_{\max}$, our goal is to minimize the decoupled noise variance $\sigma_{t_{\max}+1}^2$ given by Theorem 1, as the smallest $\sigma_{t_{\max}+1}^2$ minimizes the error probability of our algorithm. Hence, optimal tuning tries to identify a sequence of variance parameters $\{\tau^1, \ldots, \tau^{t_{\max}}\}$ so that mcB-AMP ultimately leads to the smallest $\sigma_{t_{\max}+1}^2$; sub-optimal choices of τ^t either lead to a higher $\sigma_{t_{\max}+1}^2$ or cause slower convergence to the smallest $\sigma_{t_{\max}+1}^2$. We will use the following definition [20].

Definition 2. *Assume the large-system limit and denote the decoupled noise variance of mcB-AMP obtained from the sequence $\{\tau^1, \ldots, \tau^{t_{\max}}\}$ as $\sigma_{t_{\max}+1}^2(\tau^1, \ldots, \tau^{t_{\max}})$. A sequence of parameters $\{\tau_\star^1, \ldots, \tau_\star^{t_{\max}}\}$ is* optimally tuned *at iteration $t_{\max}$, if and only if for all $\{\tau^1, \ldots, \tau^{t_{\max}}\}$ with $\tau^t \in [0, \infty)$ we have*

$$\sigma_{t_{\max}+1}^2(\tau_\star^1, \ldots, \tau_\star^{t_{\max}}) \leq \sigma_{t_{\max}+1}^2(\tau^1, \ldots, \tau^{t_{\max}}). \tag{7}$$

We next show that the tuning stage in (2), which is carried out at every iteration of mcB-AMP, achieves the smallest $\sigma_{t_{\max}+1}^2$, i.e., optimally tunes the variance parameters τ^t. We omit the proof details as it follows closely that in [20, Sec. 4.4].

Theorem 3. *Suppose $\{\tau_\star^1, \ldots, \tau_\star^{t_{\max}}\}$ are optimally-tuned for iteration $t_{\max}$. Then, for any $t < t_{\max}$, $\{\tau_\star^1, \ldots, \tau_\star^t\}$ are also optimally-tuned for iteration t. Thus, one can obtain $t_{\max}$ optimally-tuned variance parameters by optimizing $\tau_\star^1$ at $t = 1$, and then, proceeding iteratively by optimizing $\tau_\star^t$ until $t = t_{\max}$.*

In general, the exact value of the decoupled noise variance σ_t^2 that is needed for the tuning stage in step (2) to select $\tau_\star^t$ is unknown. We therefore use the estimate $\hat{\sigma}_t^2 = \frac{1}{M_{\mathsf{R}}}\|\mathbf{r}^t\|^2$ in

step (1), which was shown to converge to the true decoupled noise variance σ_t^2 in the large-system limit [18].

D. Decomposition of complex-valued systems

We now briefly discuss properties of mcB-AMP in complex-valued systems that will be necessary for our analysis of mcB-AMP in MIMO systems. In particular, we show that for special constellations, the complex-valued set $\mathcal{O}$ can be exactly characterized by a suitably-chosen real-valued set $\text{Re}\{\mathcal{O}\}$.

Definition 3. *For all $s \in \mathcal{O}$, express s as $s = a + ib$, where $a \in \text{Re}\{\mathcal{O}\}$, $b \in \text{Im}\{\mathcal{O}\}$. Then, the constellation $\mathcal{O}$ is* separable *if $p(s) = p(a)p(b)$ holds for all $s \in \mathcal{O}$ and $\text{Re}\{\mathcal{O}\} = \text{Im}\{\mathcal{O}\}$.*

An example of a separable constellation is M^2-QAM with equally likely transmit symbols. For such a separable set $\mathcal{O}$, the following lemma (with proof in [7]) provides the equivalence of the complex-valued and real-valued SE framework.

Lemma 4. *For a separable constellation set $\mathcal{O}$, the mismatched SE recursion in Theorem 1 can be expressed equivalently by:*

$$\sigma_{t+1}^2 = N_0 + 2\beta \min_{\gamma^2 \geq 0} \mathbb{E}_{S_0^R, Z^R}\left[\left(\mathsf{F}_R^{mm}(S_0^R + \sigma_t Z^R, \gamma^2) - S_0^R\right)^2\right],$$

where F_R^{mm} is posterior mean function with respect to the constellation set $\mathcal{O}^R = \text{Re}\{\mathcal{O}\}$, $S_0^R \in \mathcal{O}^R$ and $Z^R \sim \mathcal{N}(0, 1/2)$.

E. Fixed-point analysis

While the performance of mcB-AMP at every iteration can be characterized by the mismatched SE recursion equations in Theorem 1, we are interested in analyzing the performance of mcB-AMP for $t_{\max} \to \infty$. In this case, the mismatched SE in Theorem 1 converges to the following fixed-point equation:

$$\sigma_\star^2 = N_0 + \beta \min_{\gamma^2 \geq 0} \Psi^{mm}(\sigma_\star^2, \gamma^2) = N_0 + \beta \Psi_\star^{mm}(\sigma_\star^2). \quad (8)$$

Thus, as $t_{\max} \to \infty$, σ_{t+1}^2 in Theorem 1 converges to $\sigma_\star^2$ as given by (8). In general, if there are multiple fixed points, then mcB-AMP converges to the largest fixed-point, which ultimately leads to a higher probability of error than that of the smallest fixed-point solution. To provide conditions that ensure a unique fixed-point solution to (8) (see Section III), we use the following lemma (the proof is given in [7]).

Lemma 5. *Fix $p(\mathbf{s}_0)$ and $\tilde{p}(\tilde{s})$. The* minimum recovery threshold *(MRT) $\beta^{\min}$ for M-LAMA is defined by:*

$$\beta^{\min} = \min_{\sigma^2 \geq 0} \left(\frac{\mathrm{d}\Psi_\star^{mm}(\sigma^2)}{\mathrm{d}\sigma^2}\right)^{-1}. \quad (9)$$

For all system ratios $\beta < \beta^{\min}$ the fixed-point solution in (8) is unique. Furthermore, let $\sigma_\star^2 = \arg\min_{\sigma^2 \geq 0}(\mathrm{d}\Psi_\star^{mm}(\sigma^2)/\mathrm{d}\sigma^2)^{-1}$. If for any other $\sigma^2 \neq \sigma_\star^2$, $\beta^{\min}\mathrm{d}\Psi_\star^{mm}(\sigma^2)/\mathrm{d}\sigma^2 < 1$, then M-LAMA also has a unique fixed point at $\beta = \beta^{\min}$.

III. Mismatched Data Detection with Tuning

We now apply the mismatched cB-AMP framework to mismatched data detection in large MIMO systems, and refer to the algorithm as mismatched large MIMO AMP (M-LAMA).

In what follows, we assume that the true prior is taken from a discrete constellation set $\mathcal{O}$ with equally likely symbols, i.e.,

$p(s_{0\ell}) = \frac{1}{|\mathcal{O}|}\sum_{a \in \mathcal{O}}\delta(s_{0\ell} - a)$, where $|\mathcal{O}|$ is the cardinality of the set $\mathcal{O}$; we also define $\mathbb{E}_{S_0}[|S_0|^2] = E_s$. Note that in MIMO systems, the true signal prior is generally known. Hence, it is natural to ask why one should consider a mismatched prior, especially since the LAMA algorithm [6] that uses the true prior minimizes the error probability. To answer this question, consider the posterior mean function (4) of LAMA [6]

$$\mathsf{F}(r, \tau) = \frac{\sum_{a \in \mathcal{O}} a \exp\left(-\frac{1}{\tau}|r-a|^2\right)}{\sum_{a \in \mathcal{O}} \exp\left(-\frac{1}{\tau}|r-a|^2\right)}, \quad (10)$$

whose calculation requires high arithmetic precision. In fact, even the use of double-precision floating-point arithmetic becomes numerically unstable for small values of τ. Hence, the development of hardware designs for LAMA that use finite-precision arithmetic is challenging. In contrast, suitably-chosen mismatched priors can lead to hardware-friendly data detectors.

A. Optimally-tuned data detection with a Gaussian prior

We now derive an M-LAMA algorithm variant with a mismatched Gaussian prior. Without loss of generality, we assume a standard complex Gaussian distribution for the mismatched prior, i.e., $\tilde{p}(\tilde{s}_\ell) \sim \mathcal{CN}(0, 1)$ as the variance parameter τ^t will be scaled accordingly to E_s in the tuning stage (2). For the mismatched Gaussian prior, the posterior mean function (4) is given by $\mathsf{F}^{mm}(r, \tau) = \frac{E_s}{E_s + \tau}r$. In the large-system limit, we can derive the following mismatched SE recursion given in (6) using Theorem 1:

$$\sigma_{t+1}^2 = N_0 + \beta \min_{\gamma^2 \geq 0} \frac{E_s}{(E_s + \gamma^2)^2}(E_s\sigma_t^2 + \gamma^4). \quad (11)$$

Evidently, the mismatched SE recursion (11) only depends on the signal energy E_s and no other properties of the true prior $p(\mathbf{s}_0)$. This fact allows us to optimally tune the variance parameters only with knowledge of signal energy E_s. We note that the RHS of (11) is minimized by $\gamma^2 = \sigma_t^2$ and thus, the mismatched SE recursion becomes:

$$\sigma_{t+1}^2 = N_0 + \beta \frac{E_s}{E_s + \sigma_t^2}\sigma_t^2. \quad (12)$$

By Lemma 5, M-LAMA has a unique fixed point if $\beta \leq 1$. Interestingly, if we define signal-to-interference ratio (SIR) as $SIR = 1/\sigma^2$ and let $t_{\max} \to \infty$, then the fixed-point solution of (12) coincides to the SIR of the linear MMSE detector in the large-system limit [8]–[10]. Hence, for a mismatched Gaussian prior, M-LAMA achieves the same performance as the linear MMSE detector. We note that the proofs given in [8]–[10] use results from random matrix theory, whereas our analysis uses the mismatched SE recursion in Theorem 1. Furthermore, our result is constructive, i.e., M-LAMA is a novel, computationally efficient algorithm that implements linear MMSE detection.

B. Sub-optimal data detection with a Gaussian prior

We can replace the optimal tuning stage of τ^t in (2) for the M-LAMA algorithm by a fixed variance parameter choice, which leads to a sub-optimal, mismatched algorithm, referred to as sub-optimal M-LAMA (short SM-LAMA). We now show that this approach leads to other well-known linear data detectors. In particular, we obtain the following mismatched

SE recursions in the large-system limit for the following two choices of variance parameters $\gamma_t^2 \to 0$ and $\gamma_t^2 \to \infty$ in (5):

$$\text{(ZF)} \quad \sigma_{t+1}^2 = N_0 + \beta \lim_{\gamma_t^2 \to 0} \Psi^{\text{mm}}(\sigma_t^2, \gamma_t^2) = N_0 + \beta \sigma_t^2,$$

$$\text{(MF)} \quad \sigma_{t+1}^2 = N_0 + \beta \lim_{\gamma_t^2 \to \infty} \Psi^{\text{mm}}(\sigma_t^2, \gamma_t^2) = N_0 + \beta \mathbb{V}\text{ar}_{S_0}[S_0].$$

By Lemma 5, (ZF) and (MF) has a unique fixed point if $\beta < 1$ and for any finite β, respectively. If $\beta < 1$, then the solution to the fixed-point equation of (ZF) and (MF) coincides exactly to the SIR given by ZF and MF detector in the large-system limit [8]–[10], respectively. Hence, by choosing specific predefined and sub-optimal variance parameters γ_t^2, SM-LAMA can be used to perform ZF and MF data detection.

C. Optimally-tuned data detection with a hypercube prior

We now derive an M-LAMA algorithm variant using a mismatched uniform distribution within a hypercube around the true prior distribution of square constellations (e.g., QPSK and 16-QAM). For example, for a QPSK system with equally likely symbols, we use a mismatched prior distributed uniformly in the interval $[-1, +1]$ for both the real and imaginary part. For this mismatched prior, we use Lemma 4 to compute the posterior mean function independently for the real and imaginary part; the posterior mean function F^{mm} is given by

$$\mathsf{F}^{\text{mm}}(s_\ell, \tau) = s_\ell + \frac{\tau}{2}\left(\nu_-(\mathsf{s}_\ell^{\text{R}}, \tau/2) + i\nu_-(\mathsf{s}_\ell^{\text{I}}, \tau/2)\right), \quad (13)$$

where we use $\mathsf{s}_\ell^{\text{R}} = \text{Re}\{s_\ell\}$, $\mathsf{s}_\ell^{\text{I}} = \text{Im}\{s_\ell\}$, and

$$\nu_-(s_\ell, \tau) = \frac{e^{-\frac{1}{2\tau}(s_\ell+\alpha)^2} - e^{-\frac{1}{2\tau}(s_\ell-\alpha)^2}}{\sqrt{2\pi\tau}\left(\Phi\left(\frac{s_\ell+\alpha}{\sqrt{\tau}}\right) - \Phi\left(\frac{s_\ell-\alpha}{\sqrt{\tau}}\right)\right)},$$

with $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \mathrm{d}u$. The mismatched SE recursion can be obtained from Theorem 1 and evaluated numerically.

There are two disadvantages of this algorithm: (i) The computation of the posterior mean function (13) is not efficient from a hardware perspective as it involves transcendental functions. In fact, computing $\nu_-(s_\ell, \tau)$ requires—similar to that of the optimal LAMA algorithm (10)—excessively high numerical precision; (ii) the tuning stage (2) turns out to be non-trivial and requires numerical methods to find a minimum. Hence, corresponding hardware designs are impractical.

D. Sub-optimal data detection with a hypercube prior

Analogously to the ZF detector in Section III-B, we can derive a sub-optimal variant of M-LAMA (SM-LAMA) with the hypercube prior from Section III-C, where we replace the tuning stage in (2) by the fixed choice $\tau^t \to 0$.

This choice leads to a much simpler algorithm compared to the optimally-tuned M-LAMA algorithm and enables a detailed performance analysis. The posterior mean function reduces to

$$\lim_{\tau \to 0} \mathsf{F}^{\text{mm}}(s_\ell, \tau) = \mathsf{s}_\ell^{\text{R}} + \text{sign}(\mathsf{s}_\ell^{\text{R}}) \min\{\alpha - |\mathsf{s}_\ell^{\text{R}}|, 0\}$$
$$+ i\left(\mathsf{s}_\ell^{\text{I}} + \text{sign}(\mathsf{s}_\ell^{\text{I}}) \min\{\alpha - |\mathsf{s}_\ell^{\text{I}}|, 0\}\right),$$

and thus, $\lim_{\tau \to 0} \mathsf{F}^{\text{mm}}(s_\ell, \tau)$ and its derivative can be evaluated efficiently in hardware. Furthermore, by fixing $\tau \to 0$, the tuning stages in (1) and (2) are no longer required.

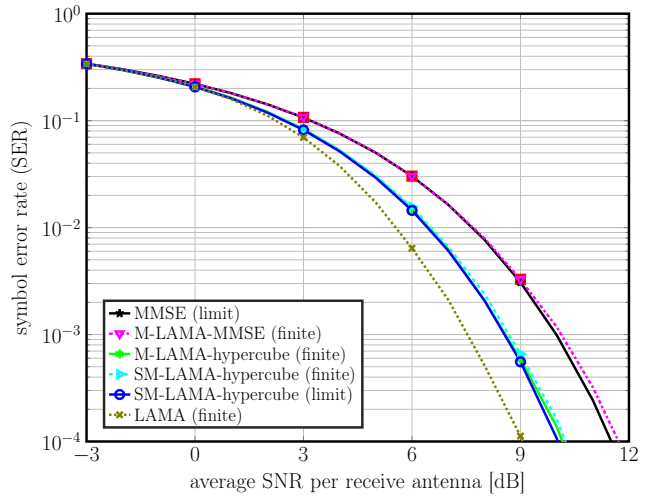Fig. 1. Symbol-error rate of M-LAMA and its variants for a 128×64 large-MIMO system with 10 iterations and QPSK; (finite) and (limit) denote the detectors that are simulated and computed by SE (in the large-system limit), respectively. SM-LAMA with the uniform hypercube prior performs within 1 dB from LAMA [6] that achieves IO performance in the large-system limit.

We now present conditions on the system ratio β for which SM-LAMA has a unique fixed point. The following Lemma 6, with proof in Appendix C, shows that the MRT of SM-LAMA for M^2-QAM is given by $\beta^{\text{min}} = (1 - 1/M)^{-1}$.

Lemma 6. *Assume a M^2-QAM prior for S_0 with equally likely symbols. Then, the MRT for SM-LAMA is given by $\beta^{\text{min}} = (1 - 1/M)^{-1}$. Moreover, SM-LAMA has a unique fixed point at $\beta = \beta^{\text{min}}$ regardless of the noise variance N_0.*

Using Lemma 6, we obtain the same MRT for M-PAM constellations by Lemma 4. We omit the proof and refer to [7].

Corollary 7. *SM-LAMA has the same MRT for M^2-QAM and M-PAM in a complex and real-valued system, respectively.*

We now show that this computationally-efficient SM-LAMA variant achieves the same performance as a well-known relaxation of the maximum likelihood data detection problem [11]–[13]. The algorithm, which is known as box relaxation (BOX) detector, solves the following convex problem:

$$\hat{\mathbf{s}} = \arg\min_{\tilde{\mathbf{s}} \in \mathbb{C}^{M_{\text{T}}}} \|\mathbf{y} - \mathbf{H}\tilde{\mathbf{s}}\|_2 \quad \text{subject to } \|\tilde{\mathbf{s}}\|_\infty \leq 1 \quad (14)$$

and slices the individual entries of $\hat{\mathbf{s}}$ onto the QPSK (or BSPK) constellations. The next result shows that SM-LAMA achieves the same error rate performance as the BOX detector for $\beta < 2$, while providing a simple and computationally efficient algorithm. The proof is given in Appendix B.

Lemma 8. *Assume $\beta < 2$ and the large system limit. Then, for a complex-valued MIMO system with QPSK (or a real-valued MIMO system with BPSK), SM-LAMA achieves the same error-rate performance as the BOX algorithm in (14).*

IV. NUMERICAL RESULTS

We now compare the error-rate performance of the M-LAMA algorithm variants proposed in this paper. While the mismatched

SE framework in Theorem 1 enables an exact performance analysis in the large-system limit, we also carry out Monte–Carlo simulations in a finite-dimensional large MIMO system with 128 receive and 64 transmit antennas. Fig. 1 shows the symbol error-rate performance of LAMA, M-LAMA, and SM-LAMA. The optimally-tuned M-LAMA and sub-optimal SM-LAMA with the uniform hypercube prior perform within 1 dB of the LAMA algorithm [6], which achieves IO performance in the large-system limit. These results demonstrate that carefully-selected mismatched priors enable near-IO performance in finite-dimensional systems and in a hardware-friendly way.

APPENDIX A
DERIVATION OF THE MSE FUNCTION $\Psi^{\mathrm{QAM}}(\sigma^2)$

We will compute the MSE function for $\Psi^{\mathrm{QAM}}(\sigma^2)$ by first computing the MSE function $\Psi^{\mathrm{PAM}}(\sigma^2)$ for a real-valued M-PAM system with equally likely symbols and then, use Lemma 4 to express $\Psi^{\mathrm{QAM}}(\sigma^2)$ for M^2-QAM. We start by defining $\mathsf{F}^\alpha(s_\ell) = s_\ell + \mathrm{sign}(s_\ell)\min\{\alpha - |s_\ell|, 0\}$. Note that for equally likely symbols, the M-PAM constellation can be expressed by $p(s_\ell) = \frac{1}{M}\sum_{k=-M/2+1}^{M/2} \delta(s_\ell - (2k-1))$. Thus, by using symmetry for equally likely M-PAM symbols:

$$\Psi^{\mathrm{PAM}}(\sigma^2) = \frac{2}{M}\sum_{k=1}^{M/2}\Psi_k^{\mathrm{PAM}}(\sigma^2), \tag{15}$$

where we introduced $\Psi_k^{\mathrm{PAM}}(\sigma^2)$ that is defined by:

$$\begin{aligned}\Psi_k^{\mathrm{PAM}}(\sigma^2) &= \mathbb{E}_Z[(\mathsf{F}^\alpha((2k-1)+\sigma Z)-(2k-1))^2]\\ &= \sigma^2 + (\bar{\alpha}_k^2 - \sigma^2)Q\left(\frac{\bar{\alpha}_k}{\sigma}\right) + (\alpha_k^2 - \sigma^2)Q\left(\frac{\alpha_k}{\sigma}\right)\\ &\quad - \frac{\sigma}{\sqrt{2\pi}}\bar{\alpha}_k e^{-\frac{\bar{\alpha}_k^2}{2\sigma^2}} - \frac{\sigma}{\sqrt{2\pi}}\alpha_k e^{-\frac{\alpha_k^2}{2\sigma^2}},\end{aligned}$$

with $\bar{\alpha}_k = \alpha - (2k-1)$, $\alpha_k = \alpha + (2k-1)$, $Z \sim \mathcal{N}(0,1)$, and $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-u^2/2}\mathrm{d}u$. By Lemma 4, Ψ^{QAM} can be computed by $\Psi^{\mathrm{QAM}}(\sigma^2) = 2\Psi^{\mathrm{PAM}}(\sigma^2/2)$ with $\alpha = M-1$.

APPENDIX B
PROOF OF LEMMA 8

We start by stating the following result from [13, Thm. 2.1] that establishes the error-rate performance of BOX detector.

Theorem 9 (Theorem 2.1 [13]). *Assume a real-valued BPSK system with $\beta < 2$. The bit-error rate in the large-system limit converges to $Q(1/\tau_\star)$, where $\tau_\star$ is the unique solution to $\tau_\star = \arg\min_{\tau>0} g(\tau)$, where $g(\tau)$ is defined by*

$$g(\tau) = \frac{\tau}{2}\left(\frac{1}{\beta}-\frac{1}{2}\right) + \frac{N_0}{2\beta\tau} + \frac{\tau}{2}\int_{\frac{\tau}{2}}^\infty \left(x - \frac{2}{\tau}\right)^2 \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x.$$

We note that as $\tau_\star$ is the unique, minimal solution to $g(\tau)$, we have $g'(\tau_\star) = 0$. Rearranging terms in $g'(\tau_\star) = 0$ results in the fixed-point equation $\tau_\star^2 = N_0 + \beta\Psi(\tau_\star^2)$ with

$$\Psi(\tau_\star^2) = \frac{1}{2}\tau_\star^2 - \frac{2\tau_\star}{\sqrt{2\pi}}e^{-\frac{2}{\tau_\star^2}} + Q\left(\frac{2}{\tau_\star}\right)(4-\tau_\star^2). \tag{16}$$

The function $\Psi(\tau_\star^2)$ in (16) is identical to (15) for a BPSK system, i.e., $M = 2$ with $\sigma^2 = \tau_\star^2$. This implies that the BOX-relaxed method in [13] and SM-LAMA with the uniform hypercube prior achieves the same fixed-point (8). Moreover, due to the decoupling property of SM-LAMA detailed in

Section II-B, the error-rate of a real-valued BPSK system in the large-system limit is given by $Q(1/\tau_\star)$. The result of Theorem 9 can be generalized to complex-valued QPSK systems and is equal to fixed-point solution of SM-LAMA by Lemma 4.

APPENDIX C
PROOF OF LEMMA 6

As shown in Appendix A, we compute $\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\Psi^{\mathrm{QAM}}(\sigma^2)$ and observe that $\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\Psi^{\mathrm{QAM}}(\sigma^2) \le 1 - 1/M$, where equality is achieved only when $\sigma^2 \to 0$. Thus, $\beta^{\min} = (1-1/M)^{-1}$. To show that M-LAMA also has a unique fixed point when $\beta = \beta^{\min}$, we use Lemma 5 and observe that no other $\sigma_\star^2 > 0$ satisfies $\beta^{\min} = (\mathrm{d}\Psi(\sigma^2)/\mathrm{d}\sigma^2)^{-1}$ at $\sigma^2 = \sigma_\star^2$.

REFERENCES

[1] C. Jeon, A. Maleki, and C. Studer, "Mismatched data detection in large MIMO systems," *in preparation*.
[2] S. Verdú, *Multiuser Detection*. Cambridge University Press, 1998.
[3] D. Guo and S. Verdú, "Multiuser detection and statistical mechanics," in *Commun., Inf. and Netw. Security*. Springer, 2003, pp. 229–277.
[4] ——, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983–2010, Jun. 2005.
[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] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1227–1231.
[7] ——, "Optimal data detection in large MIMO," *in preparation*.
[8] 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.
[9] D. Tse and S. Hanly, "Linear multiuser receivers: effective interference, effective bandwidth and user capacity," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 641–657, Mar. 1999.
[10] S. Shamai and S. Verdu, "The impact of frequency-flat fading on the spectral efficiency of CDMA," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1302–1327, May 2001.
[11] P. H. Tan, L. K. Rasmussen, and T. J. Lim, "Constrained maximum-likelihood detection in CDMA," *IEEE Trans. Commun.*, vol. 49, no. 1, pp. 142–153, Jan. 2001.
[12] A. Yener, R. D. Yates, and S. Ulukus, "CDMA multiuser detection: A nonlinear programming approach," *IEEE Trans. Commun.*, vol. 50, no. 6, pp. 1016–1024, Jun. 2002.
[13] C. Thrampoulidis, E. Abbasi, W. Xu, and B. Hassibi, "BER analysis of the box relaxation for BPSK signal recovery," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016.
[14] D. L. Donoho and J. Tanner, "Counting the faces of randomly-projected hypercubes and orthants, with applications," *Discrete Comput. Geometry*, vol. 43, no. 3, pp. 522–541, Apr. 2010.
[15] O. Mangasarian and B. Recht, "Probability of unique integer solution to a system of linear equations," *Eur. J. Oper. Res.*, vol. 214, no. 1, pp. 27–30, Oct. 2011.
[16] 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.
[17] 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.
[18] A. Montanari, *Graphical models concepts in compressed sensing, Compressed Sensing (Y.C. Eldar and G. Kutyniok, eds.)*. Cambridge University Press, 2012.
[19] A. Mousavi, A. Maleki, and R. G. Baraniuk, "Parameterless Optimal Approximate Message Passing," *arXiv:1311.0035 [cs.IT]*, Oct. 2013.
[20] ——, "Consistent Parameter Estimation for LASSO and Approximate Message Passing," *arXiv:1511.01017 [math.ST]*, Nov. 2015.
[21] A. Maleki, "Approximate message passing algorithms for compressed sensing," Ph.D. dissertation, Stanford University, Jan. 2011.
[22] 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.