

Fast and Robust Expectation Propagation MIMO Detection via Preconditioned Conjugated Gradient

Luca Schmid*, Dominik Sulz[†], and Laurent Schmalen*

*Communications Engineering Lab, Karlsruhe Institute of Technology, 76187 Karlsruhe, Germany

[†]Mathematical Institute, University of Tübingen, 72076 Tübingen, Germany

Email: *first.last@kit.edu, [†]dominik.sulz@uni-tuebingen.de

Abstract—We study the expectation propagation (EP) algorithm for symbol detection in massive multiple-input multiple-output (MIMO) systems. The EP detector shows excellent performance but suffers from a high computational complexity due to the matrix inversion, required in each EP iteration to perform marginal inference on a Gaussian system. We propose an inversion-free variant of the EP algorithm by treating inference on the mean and variance as two separate and simpler subtasks: We study the preconditioned conjugate gradient algorithm for obtaining the mean, which can significantly reduce the complexity and increase stability by relying on the Jacobi preconditioner that proves to fit the EP characteristics very well. For the variance, we use a simple approximation based on linear regression of the Gram channel matrix. Numerical studies on the Rayleigh-fading channel and on a realistic 3GPP channel model reveal the efficiency of the proposed scheme, which offers an attractive performance-complexity tradeoff and even outperforms the original EP detector in high multi-user inference cases where the matrix inversion becomes numerically unstable.

Index Terms—Expectation propagation, massive MIMO detector, conjugate gradient, low complexity, 6G.

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) is discussed as a key technology for future cellular communication systems due to its ability to serve multiple users on the same time and frequency resource with a high spectral efficiency by leveraging diversity and multiplexing gains whose advantages scale with the number of antennas [1]. Due to strong multi-user interference (MUI), efficient symbol detection becomes a crucial bottleneck in achieving the promising theoretical guarantees in practice. Optimal detection based on the maximum likelihood (ML) criterion quickly becomes intractable for growing system sizes. On the other hand, linear schemes with reduced complexity, like the linear minimum mean squared error (LMMSE) detector, show poor performance.

Many proposed MIMO detectors with a promising performance-complexity tradeoff are based on iterative algorithms that can be described as message passing (MP) on probabilistic graphical models [2]. A high-accuracy MP detector is based on the expectation propagation (EP) algorithm [3], [4]. To obtain global consistency between local approximations,

the EP detector iteratively constructs a Gaussian approximation to the posterior distribution of the transmitted MIMO symbols to match the moments in each MIMO dimension. It thereby achieves excellent performance for various antenna configurations [4].

A major drawback of the EP detector is the cubic complexity of the required matrix inversion in each EP iteration, which becomes a prohibitive computational burden for future massive MIMO systems with 32-256 antennas. To avoid the expensive computation of the matrix inverse, the authors in [5] reformulate the EP update equations using a fixed point approximation. The proposed approximation is quite coarse and the resulting matrix inversion-free detector suffers from slow convergence rate and severe performance degradation [5]. Therefore, the authors suggest falling back to the matrix inverse only for the initialization of the algorithm. The improved detector EPANet is proposed in [6], where deep learning techniques are considered to optimize the detector in an end-to-end manner. The EPANet algorithm works very well for low-MUI scenarios but the performance gap to the EP detector increases in the presence of high MUI. On the other hand, various works in literature study methods for approximating the matrix inverse, e.g., based on the Neumann-series [7]. However, the complexity reduction is limited if a good approximation is intended.

In contrast to the works mentioned above, we argue that the EP detector does *not* require the explicit computation of the full matrix inverse. Instead, the problem can be divided into the simpler subtasks of finding the diagonal elements of the inverse as well as solving a linear system of equations. The latter is a rigorously studied problem and there exists a variety of efficient approximative methods such as the Gauss-Seidel (GS) or the conjugated gradient (CG) algorithm [8]. These methods were already considered for the MIMO detection problem to approximate the LMMSE detector [9], [10] and in [11], the error of the CG algorithm was studied theoretically in the context of EP for the asymptotic case.

In this work, we compare the suitability of different approximative linear solvers in relieving the computational burden in the EP detector. Leveraging an intrinsic property of EP, we propose a well-suited preconditioner to substantially reduce the required number of CG iterations. For the diagonal elements of the inverse, we rely on a simple, yet effective approximation based on linear regression of the Gram channel

This work has received funding in part from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101001899) and in part from the German Federal Ministry of Education and Research (BMBF) within the project Open6GHub (grant agreement 16KISK010).

matrix. Numerical studies with varying MUI scenarios on the Rayleigh fading channel as well as on a more realistic 3GPP 3D MIMO urban macrocellular (UMa) channel model demonstrate the effectiveness of the novel method and provide valuable insights into both applied approximations. The proposed algorithm proves to be numerically more stable while achieving competitive results to the original EP detector with lower complexity.

II. SYSTEM MODEL AND PRELIMINARIES

A. System Model

We consider the multi-user MIMO channel, where N_t single-antenna users simultaneously transmit information to a base station that is equipped with N_r antennas. The observation $\tilde{\mathbf{y}} \in \mathbb{C}^{N_r}$ at the receiver is modeled by

$$\tilde{\mathbf{y}} = \tilde{\mathbf{H}}\tilde{\mathbf{x}} + \tilde{\mathbf{n}},$$

where $\tilde{\mathbf{H}} \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix, $\tilde{\mathbf{x}} \in \tilde{\mathcal{M}}^{N_t}$ is the vector of transmitted symbols and $\tilde{\mathbf{n}} \sim \mathcal{CN}(\tilde{\mathbf{n}} : \mathbf{0}, 2\sigma^2 \mathbf{I}_{N_r})$ is an additive circular-symmetric complex Gaussian noise vector. The transmit symbols are independently and uniformly sampled from a quadrature amplitude modulation (QAM) constellation $\tilde{\mathcal{M}} \subset \mathbb{C}$ of size $|\tilde{\mathcal{M}}| = M$. We define the signal-to-noise ratio (SNR) in dB at the receiver as

$$\text{snr} := 10 \log_{10} \left(\frac{\mathbb{E}[\|\tilde{\mathbf{H}}\tilde{\mathbf{x}}\|_2^2]}{\mathbb{E}[\|\tilde{\mathbf{n}}\|_2^2]} \right).$$

The complex-valued system can be decomposed into an equivalent real-valued representation

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \quad \mathbf{H} = \begin{pmatrix} \text{Re}\{\tilde{\mathbf{H}}\} & -\text{Im}\{\tilde{\mathbf{H}}\} \\ \text{Im}\{\tilde{\mathbf{H}}\} & \text{Re}\{\tilde{\mathbf{H}}\} \end{pmatrix} \in \mathbb{R}^{2N_r \times 2N_t}$$

with $\mathbf{y} = (\text{Re}\{\tilde{\mathbf{y}}\}, \text{Im}\{\tilde{\mathbf{y}}\})^T$, $\mathbf{x} = (\text{Re}\{\tilde{\mathbf{x}}\}, \text{Im}\{\tilde{\mathbf{x}}\})^T$ and $\mathbf{n} = (\text{Re}\{\tilde{\mathbf{n}}\}, \text{Im}\{\tilde{\mathbf{n}}\})^T$, which will become convenient later. The complex-valued QAM constellation $\tilde{\mathcal{M}}$ is accordingly transformed to a real-valued pulse amplitude modulation (PAM) constellation $\mathcal{M}$.

The goal of MIMO detection is to infer the transmit symbols $\mathbf{x}$ from the observation $\mathbf{y}$ at the receiver. The likelihood distribution can be expressed as

$$p(\mathbf{y}|\mathbf{x}) \prod_{n=1}^{2N_t} \mathbb{1}_{\{x_n \in \mathcal{M}\}}, \quad \mathbf{x} \in \mathbb{R}^{2N_t} \quad (1)$$

where the indicator functions $\mathbb{1}_{\{x_n \in \mathcal{M}\}}$ constrain the real-valued variables x_n to take discrete values of the constellation $\mathcal{M}$. Optimal detection with respect to the symbol error rate (SER) is based on the symbol-wise ML estimate, i.e., on the marginal distributions of eq. (1):

$$\hat{x}_{n,\text{ML}} = \underset{x_n \in \mathbb{R}}{\text{argmax}} p(\mathbf{y}|x_n) \mathbb{1}_{\{x_n \in \mathcal{M}\}}. \quad (2)$$

We assume the availability of channel state information (CSI), i.e., the receiver has perfect knowledge of $\mathbf{H}$ and σ^2 . For the modeling of $\mathbf{H}$, we follow two approaches:

1) *Rayleigh-fading Channel Model*: Each element of $\tilde{\mathbf{H}}$ is independently sampled from a circular complex standard Gaussian distribution $\mathcal{CN}(0, 1)$. Albeit the rather unrealistic assumption of uncorrelated channel coefficients in practice, this channel model is well-studied, enabling a good comparison of this work to other methods in the literature, like [4], [5], [12].

2) *3GPP 3D MIMO UMa NLOS Channel Model*: For a more realistic model, we consider an UMa scenario from the 3GPP 3D MIMO channel model [13] with a non-line-of-sight (NLOS) propagation. We define a configuration similar to [12]: N_t single-antenna users are randomly dropped within a radius of 10-500 m from a base station in a 120°-cell sector and follow a linear trajectory with a constant speed of 30 km/h (8.3 m/s). The base station is equipped with $N_r/2$ dual-polarized antennas that are installed in a rectangular planar array at 25 m height. The center frequency is set to 2.53 GHz, and the bandwidth of 20 MHz is divided among 1024 sub-carriers. As opposed to [12], this work does not focus on the exploitation of correlations in the frequency and time domain. To sample more independent channel realizations, we subsample the sub-carriers with a rate of 32 and only sample 1 OFDM symbol per channel realization. Furthermore, we assume perfect power control, i.e.,

$$\frac{1}{N_r} \sum_{i=1}^{N_r} \|\tilde{\mathbf{H}}_{ij}\|_2^2 = 1, \quad j = 1, \dots, N_t.$$

B. Expectation Propagation for MIMO Detection

Exact ML detection in (2) is typically intractable because marginalization on $p(\mathbf{x}|\mathbf{y})$ becomes prohibitive for $N_t \gg 1$. Expectation propagation (EP) is an approximate inference method that simplifies exact Bayesian inference by approximating posterior distributions with members of an exponential family [3]. In the remainder of this section, we discuss the application of the EP method for MIMO detection as it was proposed by Céspedes et al. in [4]. For an excellent in-depth treatment of the general EP method, we refer the reader to [14] and [15].

To simplify the distribution in (1), we approximate $\mathbb{1}_{\{x_n \in \mathcal{M}\}}$ by a 1-dimensional Gaussian in canonical form

$$\mathbb{1}_{\{x_n \in \mathcal{M}\}} \approx \exp\left(\gamma_n x_n - \frac{1}{2} \Lambda_n x_n^2\right) =: t_n(x_n),$$

with free parameters γ_n and $\Lambda_n > 0$. This significantly facilitates inference, since the resulting distribution

$$q(\mathbf{x}) := \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma^2 \mathbf{I}_{2N_t}) \prod_{n=1}^{2N_t} t_n(x_n)$$

belongs to the Gaussian exponential family $\mathcal{F}_G$ for which inference is tractable [14]. In this context, the meaning of tractable inference is twofold: First, we can compute the

moment parameters associated with the sufficient statistics of $q(\mathbf{x})$, i.e., the covariance matrix and mean

$$\Sigma = (\sigma^{-2} \mathbf{H}^T \mathbf{H} + \text{diag}(\Lambda))^{-1}, \quad (3)$$

$$\mu = \Sigma (\sigma^{-2} \mathbf{H}^T \mathbf{y} + \gamma), \quad (4)$$

where we have introduced the vector notations $\Lambda = (\Lambda_1, \dots, \Lambda_{2N_t})^T$ and $\gamma = (\gamma_1, \dots, \gamma_{2N_t})^T$. Second, marginal inference for $\mathcal{F}_G$ is tractable and is directly given by $q(x_n) = \mathcal{N}(x_n : \mu_n, \Sigma_{nn})$. EP is an iterative method to determine the parameters γ_n and $\Lambda_n > 0$, $n = 1, \dots, 2N_t$ such that the Kullback-Leibler (KL) divergence $D_{\text{KL}}(p(\mathbf{x}|\mathbf{y})||q(\mathbf{x}))$ is minimized. As shown in [14], this is equivalent to matching the expected sufficient statistics of $p(\mathbf{x}|\mathbf{y})$ and $q(\mathbf{x})$, also known as *moment matching*. The EP algorithm approaches the moment matching condition by sequentially matching the moments of marginal distributions in the factorization in (1). In each iteration ℓ , the parameter updates $\gamma_n^{(\ell+1)}$ and $\Lambda_n^{(\ell+1)}$ are chosen such that the moments of the marginal Gaussian distribution

$$q^{(\ell)}(x_n) \frac{\exp\left(\gamma_n^{(\ell+1)} x_n - \frac{1}{2} \Lambda_n^{(\ell+1)} x_n^2\right)}{\exp\left(\gamma_n^{(\ell)} x_n - \frac{1}{2} \Lambda_n^{(\ell)} x_n^2\right)}$$

match with the mean $\mu_{p_n}^{(\ell)}$ and variance $\sigma_{p_n}^{2(\ell)}$ of the marginal tiled distribution

$$q^{(\ell)}(x_n) \frac{\mathbb{1}_{\{x_n \in \mathcal{M}\}}}{\exp\left(\gamma_n^{(\ell)} x_n - \frac{1}{2} \Lambda_n^{(\ell)} x_n^2\right)},$$

for all $n = 1, \dots, 2N_t$, respectively. We denote by $q^{(\ell)}(x_n) = \mathcal{N}(x_n : \mu_n^{(\ell)}, \sigma_n^{2(\ell)})$ the marginals of $q^{(\ell)}(\mathbf{x})$ in iteration ℓ . Solving the moment matching conditions for $\gamma_n^{(\ell+1)}$ and $\Lambda_n^{(\ell+1)}$ leads to the EP update equations

$$\Lambda_n^{(\ell+1)} = \beta \left(\frac{1}{\sigma_{p_n}^{2(\ell)}} - \frac{1 - \sigma_n^{2(\ell)} \Lambda_n^{(\ell)}}{\sigma_n^{2(\ell)}} \right) + (1 - \beta) \Lambda_n^{(\ell)}, \quad (5)$$

$$\gamma_n^{(\ell+1)} = \beta \left(\frac{\mu_{p_n}^{(\ell)}}{\sigma_{p_n}^{2(\ell)}} - \frac{\mu_n^{(\ell)}}{\sigma_n^{2(\ell)}} + \gamma_n^{(\ell)} \right) + (1 - \beta) \gamma_n^{(\ell)},$$

where $\beta \in [0, 1]$ is a smoothing parameter. After convergence or a fixed number of EP iterations L , the obtained approximation $q(\mathbf{x}) \approx p(\mathbf{x}|\mathbf{y})$ significantly simplifies the MIMO detection problem in (2) to a nearest neighbor decision

$$\hat{x}_{n,\text{EP}} = \underset{x_n \in \mathcal{M}}{\text{argmin}} \|x_n - \mu_n^{(L)}\|_2^2.$$

III. THE EPICG ALGORITHM

In this section we give a detailed description of the proposed *Expectation Propagation integrated Conjugate Gradient* (EPICG) algorithm. To perform marginal inference on the multivariate Gaussian distribution $q(\mathbf{x})$, i.e., to obtain the mean $\mu_n^{(\ell)}$ and variance $\sigma_n^{2(\ell)}$ of the marginals $q^{(\ell)}(x_n)$, the EP detector computes the covariance matrix Σ in (3) for each iteration ℓ . The matrix inversion operation is unfavorable from a numerical point of view since it can lead to numerical

instabilities and further imposes a complexity of the order $\mathcal{O}((2N_t)^3)$. This is typically unaffordable for modern massive MIMO system sizes and thus inhibits widespread employment of the EP detector in practice. In addition, the solution can become less accurate for ill-conditioned channel matrices, limiting the detection performance [16].

We argue that marginal inference on $q(\mathbf{x})$ does *not* require the computation of the *full* matrix inverse. We only require the diagonal elements of the inverse to obtain the variance of the marginals $q^{(\ell)}(x_n)$; their mean is provided as the solution of a linear system of equations.¹ Based on this observation, we propose to divide the marginal inference problem into two separate and simpler sub-tasks. In the following, we discuss both subtasks and propose approximative solutions with reduced complexity. Note that although we are treating both problems separately, the intrinsic moment matching condition of the EP algorithm combines both tasks in each EP iteration ℓ .

A. Marginal Inference: Mean

As an alternative to (3), the mean values $\mu_n^{(\ell)}$ can be computed without the matrix inversion by solving the linear system of equations

$$\mathbf{A}^{(\ell)} \boldsymbol{\mu}^{(\ell)} = \mathbf{b}^{(\ell)}, \quad \ell = 0, \dots, L \quad (6)$$

with

$$\begin{aligned} \mathbf{A}^{(\ell)} &:= \sigma^{-2} \mathbf{H}^T \mathbf{H} + \text{diag}(\Lambda^{(\ell)}), \\ \mathbf{b}^{(\ell)} &:= \sigma^{-2} \mathbf{H}^T \mathbf{y} + \gamma^{(\ell)}. \end{aligned} \quad (7)$$

Solving a linear system is generally numerically more stable than computing the inverse, especially if the inverse is only part of a larger computation [17], as it is the case in (4). Several stable and rigorously studied numerical methods exist to solve linear systems. We focus on iterative methods since they provide low-complexity approximative solutions upon early termination. Besides the well-known GS method [8, Sec. 11.2], we consider Krylov subspace methods in the following.

1) *Krylov Subspace Methods*: In numerical linear algebra, Krylov subspace methods are a well-understood tool to iteratively find approximations to solutions of linear systems. The key idea is to find the best approximation of the solution in a subspace, the so-called Krylov space. In each iteration, this subspace is enlarged and a better approximation is constructed in the larger Krylov subspace. Standard algorithms for the Krylov subspace method are the CG algorithm and Lanczos/Arnoldi-based methods such as the full orthogonalization method (FOM) and biconjugate gradient (BiCG) [8, [18]. The latter methods involve a solution of a $k \times k$ linear system in the k^{th} step, which becomes inefficient for large k . The CG algorithm, which was originally published in [19], is specifically tailored to symmetric positive definite

¹Note that the Gaussian belief propagation (GaBP) algorithm can provide the mean and variance of the marginals in one MP algorithm and thus seems predestined for this problem. However, due to the underlying fully connected graph, we observed non-convergent behavior in most cases.

problems like the matrices in (7), and each iteration has the same computational complexity. However, the CG algorithm is known to suffer from slow convergence if the problem is badly conditioned [8, Theorem 11.3.3]. We can substantially accelerate convergence by applying a preconditioner to the CG that is specifically tailored to the characteristics of our considered problem. For more details on preconditioning, we refer the reader to [8, Sec. 11.5].

2) *Preconditioned CG (pCG) Method:* Suppose we have a linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ for a symmetric, positive definite matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ and a matrix $\mathbf{B}$, such that $\mathbf{B} \approx \mathbf{A}^{-1}$. The matrix $\mathbf{B}$ is called the preconditioner and has to be chosen adequately. For an arbitrary initialization $\mathbf{x}^{(0)}, \mathbf{g}^{(0)} = \mathbf{A}\mathbf{x}^{(0)} - \mathbf{b}$, $\mathbf{d}^{(0)} = \mathbf{B}\mathbf{g}^{(0)}$ and $\rho^{(0)} = \langle \mathbf{B}\mathbf{g}^{(0)}, \mathbf{g}^{(0)} \rangle$ the pCG algorithm reads for $k = 0, 1, \dots$ as

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} - \alpha^{(k)} \mathbf{d}^{(k)}, & \mathbf{g}^{(k+1)} &= \mathbf{g}^{(k)} - \alpha^{(k)} \mathbf{A} \mathbf{d}^{(k)}, \\ \mathbf{d}^{(k+1)} &= \mathbf{B}\mathbf{g}^{(k+1)} + \beta^{(k)} \mathbf{d}^{(k)}, & \rho^{(k+1)} &= \langle \mathbf{B}\mathbf{g}^{(k+1)}, \mathbf{g}^{(k+1)} \rangle, \end{aligned}$$

where $\alpha^{(k)} = \rho^{(k)} / \langle \mathbf{A} \mathbf{d}^{(k)}, \mathbf{d}^{(k)} \rangle$ and $\beta^{(k)} = \rho^{(k+1)} / \rho^{(k)}$. The computational cost per pCG step is dominated by the matrix-vector product $\mathbf{A} \mathbf{d}^{(k)}$, which leads to a global complexity in the order of $\mathcal{O}(kN^2)$. The pCG algorithm reaches (except for round-off errors) the exact solution of the linear system after $k = N$ iterations. For approximative solutions with $k < N$, a reasonable stopping criterion is $\sqrt{\rho^{(k)}} \leq \text{tol} \cdot \|\mathbf{b}\|_2$.

We apply the pCG algorithm to solve the system in (6). To obtain fast convergence of the pCG method, the choice of a suitable preconditioner is crucial. In this special setting, the convergence of the EP algorithm leads to decreasing marginal variances which implies large entries in $\Lambda^{(\ell)}$ (cf. (5)) and hence strong diagonal dominance of $\mathbf{A}^{(\ell)}$. This behavior is also observed in our experiments. Hence, it is natural to choose a diagonal, so-called, Jacobi preconditioner [8, Sec. 11.5]. For $\mathbf{A}^{(\ell)} = (A_{nm})_{n,m=1}^{2N_t}$, a good choice is $\mathbf{B} = \text{diag}(1/A_{nn})_{n=1}^{2N_t}$. This choice for $\mathbf{B}$ requires no matrix inversion and allows for very efficient matrix-vector products in the pCG algorithm.

B. Marginal Inference: Variance

The marginal variances $\sigma_n^{2(\ell)}$ are obtained in each iteration ℓ of the EP algorithm by extracting the diagonal elements of the covariance matrix $\Sigma^{(\ell)} = \mathbf{A}^{-1(\ell)}$ in (3). To avoid the full matrix inversions for reduced complexity, we rely on an approximation similar to [5], which is based on the Neumann-series expansion terminated after the first term:

$$\sigma_n^{2(\ell)} = \Sigma_{nn}^{(\ell)} = \frac{1}{A_{nn}^{(\ell)}} + \sum_{m \neq n} \frac{A_{nm}^{(\ell)} A_{mn}^{(\ell)}}{A_{mm}^{(\ell)}} \frac{1}{A_{nn}^{2(\ell)}} \approx \frac{1}{A_{nn}^{(\ell)}}. \quad (8)$$

Inspired by [6], we introduce correction parameters to improve the approximation in (8):

$$\tilde{\sigma}_n^{2(\ell)} := \max \left\{ \alpha_1^{(\ell)} \frac{1}{A_{nn}^{(\ell)}} + \alpha_2^{(\ell)}, \frac{1}{A_{nn}^{(\ell)}} \right\}.$$

Compared to [6], the parameters apply a linear transformation on $\frac{1}{A_{nn}}$ (instead of on A_{nn}), and we lower bound

$\tilde{\sigma}_n^{2(\ell)}$ by $\frac{1}{A_{nn}}$ instead of by a constant ($\epsilon = 10^{-12}$). Optimizing $\alpha_1^{(\ell)}$ and $\alpha_2^{(\ell)}$ in each iteration ℓ based on a labeled dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{A}^{(\ell)}, \mathbf{A}^{-1(\ell)})_d, d = 1, \dots, D\}$ of representative matrix-inverse pairs can be interpreted as linear regression² between the dependent variables $(\mathbf{A}^{-1(\ell)})_{nn} = \sigma_n^{2(\ell)}$ and the regressors $\frac{1}{A_{nn}^{(\ell)}}$. Following the method of least squares, we obtain

$$(\alpha_1^{(\ell)}, \alpha_2^{(\ell)}) = \underset{(\alpha_1, \alpha_2) \in \mathbb{R}^2}{\text{argmin}} \sum_{\mathcal{D}_{\text{train}}} \sum_{n=1}^{2N_t} \left(\alpha_1 \frac{1}{A_{nn}^{(\ell)}} + \alpha_2 - \sigma_n^{2(\ell)} \right)^2,$$

which can be minimized in closed form [14, Sec. 3]. We observe that this approximation works remarkably well for EP iterations $\ell > 0$ and for $\ell = 0$ if the Gram channel matrix $\mathbf{H}^T \mathbf{H}$ is strongly diagonal dominant. In the remaining cases, the initialization with the inverse $\Sigma^{(0)}$ has shown to significantly increase accuracy.

C. The Proposed EPiCG Algorithm

Combining the discussions in Sec. III-A and III-B, we propose the EPiCG algorithm with two different initializations as our main contribution of this work. We apply the original EP detector (cf. Sec. II-B) with the following modifications:

- EPiCG: Compute the means $\mu_n^{(\ell)}$ by solving the linear system in (6) with the pCG method and Jacobi-preconditioning. Approximate the variances $\sigma_n^{2(\ell)}$ by the strategy proposed in Sec. III-B.
- EPiCG- $\Sigma^{(0)}$: For $\ell = 0$, initialize $\mu_n^{(0)}$ and $\sigma_n^{2(0)}$ via the inverse $\mathbf{A}^{-1(0)} = \Sigma^{(0)}$ as in the original EP detector. Otherwise for $\ell > 0$, apply the EPiCG algorithm.

IV. EXPERIMENTAL STUDY

We evaluate the EPiCG algorithm for MIMO detection and focus in particular on the analysis and distinction between the two proposed approximations. In the following numerical experiments, we use a 16-QAM constellation and fix the number of receive antennas to $N_r = 128$. To cover different MUI scenarios, we study both $N_t = 64$ and $N_t = 128$ single-antenna users. The results are based on 5000 channel matrices $\tilde{\mathbf{H}}$ that were randomly sampled from the channel models introduced in Sec. II-A, respectively.

A. Approximation of the Marginal Mean

We start by evaluating the suitability of different approximative methods for solving the linear system (6) in each EP iteration ℓ . Figure 1 plots the relative residual error (RRE) $\frac{\|\mathbf{A}^{(\ell)} \hat{\mathbf{x}}^{(k)} - \mathbf{b}^{(\ell)}\|_2}{\|\mathbf{b}^{(\ell)}\|_2}$ over the steps k of the GS, CG and pCG method, respectively. The results are shown for the Rayleigh fading channel with $N_t = 128$ and $\text{snr} = 19$ dB. We compare the methods with two non-iterative approaches from literature: solving the linear system using the inverse $\hat{\mathbf{x}} = \mathbf{A}^{-1(\ell)} \mathbf{b}^{(\ell)}$ as proposed in the original EP detector [4],

²In numerical studies, we found that a piecewise exponential regression with two intervals can fit the data much better. However, the overall gains in terms of EP detection performance were insignificant.

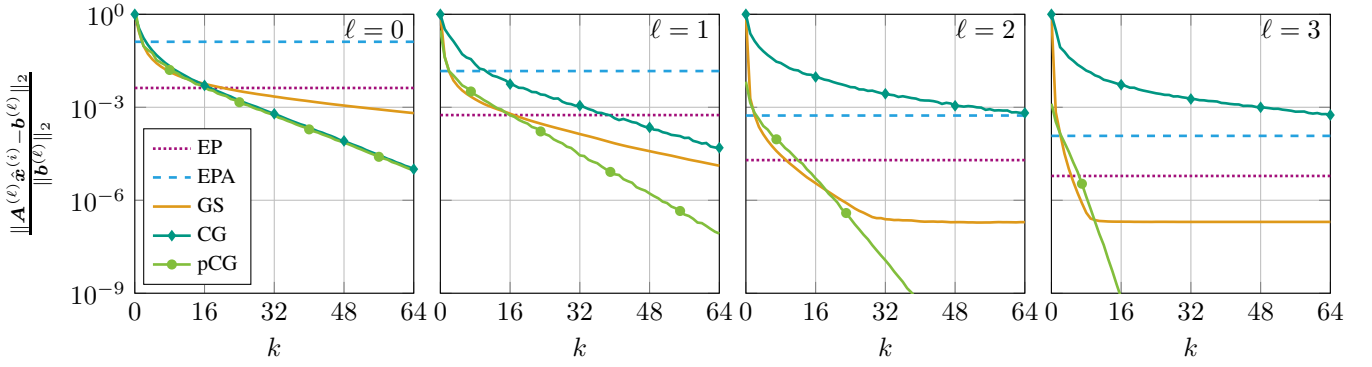


Fig. 1: Median RRE of various approximative solvers over the steps k . The plots vary the EP iteration ℓ , in which the linear system (6) is solved.

and the approximation used in the EPA algorithm, i.e., avoiding to solve the linear system completely by assuming $\hat{x}_n = \mu_{p_n}^{(\ell)}$. However, this assumption is not correct, but only applies when the fixed points of EP are reached [5].

In EP iteration $\ell = 0$, the GS solver is outperformed by the CG and pCG method for $k > 16$. The latter behave almost identically. The benefit of the Jacobi preconditioner becomes apparent for the subsequent EP iterations $\ell > 0$, where the matrix $A^{(\ell)}$ becomes more and more diagonal-dominant. While the conventional CG solver converges similarly or slower compared to $\ell = 0$, the pCG method significantly increases the convergence rate in each subsequent EP iteration ℓ . Since the computational complexity scales quadratically for each step of the considered solvers, the plots indicate a precision-complexity tradeoff. For instance in $\ell = 1$, the pCG solver only requires $k = 16$ steps to reach the same RRE $= 10^{-4}$ as the inverse-based solution of the EP detector. Note that the RRE is inherently available within the pCG steps which offers great flexibility for controlling this tradeoff.

The approximation of the EPA algorithm is very coarse in $\ell = 0$ but quickly improves with the convergence of the EP algorithm, i.e., with increasing ℓ . It approximatively requires $k = 4$ iterations of the pCG algorithm to reach the RRE of the EPA algorithm in each EP iteration ℓ .

B. Detection Performance

We compare the performance of the proposed EPiCG algorithm to the LMMSE detector, the EPANet algorithm [6] as well as to the original EP detector [4]. We fix $\beta = 0.1$ for the EP and EPiCG algorithm and we perform $L = 10$ iterations for all iterative detection schemes. We also consider the initialization variant of the EPiCG- $\Sigma^{(0)}$ algorithm for the EPANet algorithm as suggested in [6], accordingly denoted by EPANet- $\Sigma^{(0)}$. We fix the stopping criterion of the pCG algorithm to $\text{tol}(\ell) = (10^{-3}, 10^{-4}, 10^{-5}, 10^{-5}, \dots, 10^{-5})$. For the Rayleigh-fading channel, the precision is thereby similar to the matrix inverse (cf. Fig. 1) which enables a good comparison since the main difference between the EPiCG algorithm and the EP detector stems from the variance approximation. To create $\mathcal{D}_{\text{train}}$, we randomly generate $D = 100$ transmission scenarios with snr values sampled from a uniform

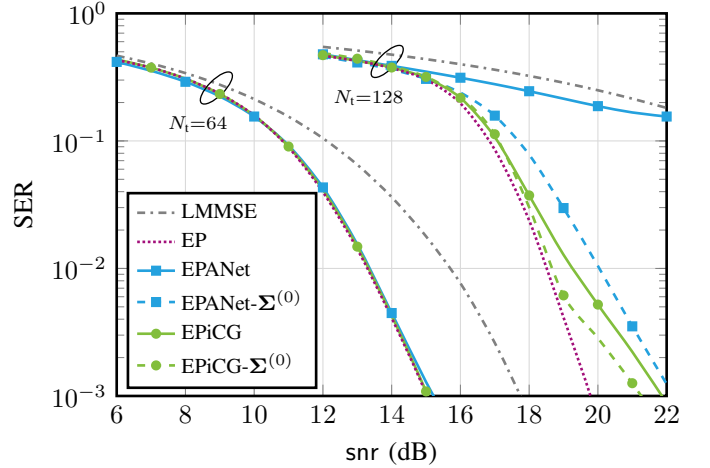


Fig. 2: SER over snr for various MIMO detectors on the Rayleigh-fading channel with $N_t = 64, 128$ and $N_r = 128$.

distribution with bounds chosen such that the SER of the resulting EPiCG detector performance is in the target SER interval $[10^{-2}, 10^{-1}]$.

Figure 2 evaluates the SER over the SNR for the Rayleigh-fading channel. In the low-MUI scenario with $N_t = 64$, both the EPANet and the EPiCG algorithm approach the EP detector performance. In the high-MUI scenario with $N_t = N_r$, the EPANet detector behaves similarly to the LMMSE detector and only reaches competitive performance when being initialized with $\Sigma^{(0)}$. The EPiCG detector follows the waterfall behavior of the EP detector with an snr offset of 0.75 dB for $\text{SER} = 10^{-2}$. We observed that the remaining gap between the EPiCG algorithm compared to the EP detector comes from the variance approximation in iteration $\ell = 0$ as the dominant source of error. The EPiCG- $\Sigma^{(0)}$ algorithm consequently closes this gap.

We further evaluate the performance of the considered detectors on the 3GPP 3D MIMO UMa NLOS channel model in Fig. 3. The significantly increased correlation in this channel model and the consequently decreased diagonal dominance of the Gram channel matrices generally degrades the performance of both the EPANet detector and the EPiCG algorithm. For

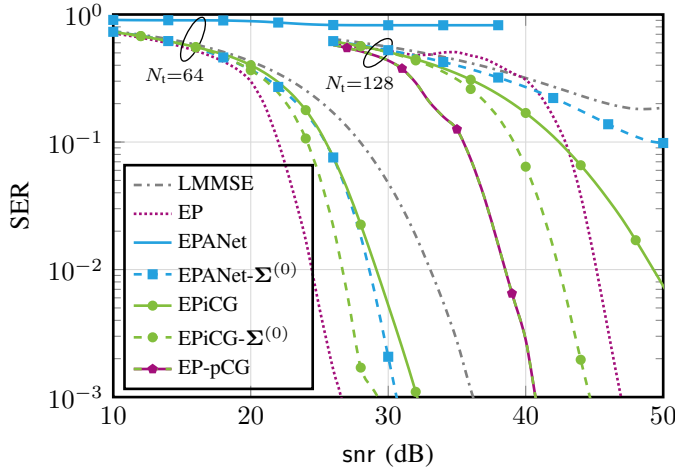


Fig. 3: SER over snr for various MIMO detectors on the 3GPP 3D MIMO UMa NLOS channel with $N_t = 64, 128$ and $N_r = 128$.

$N_t = 64$, the EPiCG- $\Sigma^{(0)}$ algorithm reduces the snr gap to EP detection from 4 dB to 2 dB at an $\text{SER} = 10^{-2}$, compared to the EPiCG and EPANet- $\Sigma^{(0)}$ detectors. In the high-MUI scenario with $N_t = 128$, the EPiCG- $\Sigma^{(0)}$ algorithm outperforms the EP detector. We conjecture that the performance of the EP detector degrades due to an unstable solution of eq. (4) via the matrix inverse. To verify this, we replace the computation of the mean in the EP detector by the pCG algorithm with $\text{tol} = 10^{-5}$ and only use the diagonal elements of the matrix inverse to obtain the variances. The result is denoted by EP-pCG in Fig. 3 and confirms the conjecture.

C. Complexity Analysis

Finally, we evaluate the proposed EPiCG detection algorithms from a computational complexity perspective. Table I provides the total number of pCG steps that are required within the EPiCG algorithm over all EP iterations, i.e., to solve L linear systems of dimension $2N_t \times 2N_t$. Since we use a precision-based stopping criterion instead of a fixed number of pCG steps, the numbers represent rounded mean values over all 5000 evaluation samples. The EPiCG- $\Sigma^{(0)}$ variant uses the matrix inverse for the computation of the marginal means in EP iteration $\ell = 0$ and requires approximately half of the pCG steps compared with the EPiCG algorithm. Note that optimizing the distribution of the pCG steps over the EP iterations might have the potential to improve further the performance-complexity tradeoff, which we leave for future work.

V. CONCLUSION

We propose the EPiCG algorithm as a faster and more robust method for massive MIMO detection. Replacing the matrix inversion for marginal inference in the EP algorithm by the application of the pCG method reduces the computational complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(kN^2)$ operations in each EP iteration. This offers a flexible performance-complexity tradeoff, while the accuracy is only mildly reduced due to

TABLE I: TOTAL NUMBER OF PCG STEPS IN ALL $L = 10$ EP ITERATIONS OF THE EPiCG ALGORITHM FOR DIFFERENT TRANSMISSION SCENARIOS.

Algorithm N_t	EPiCG		EPiCG- $\Sigma^{(0)}$	
	64	128	64	128
Rayleigh	40	125	28	94
3GPP UMa NLOS	187	652	98	323

an approximation of the marginal variance. Simulations on realistic MIMO channels illustrate that the pCG algorithm furthermore stabilizes the computation compared to the matrix inversion for highly-correlated channels with strong MUI.

REFERENCES

- [1] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1941–1988, 2015.
- [2] 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 Signal Process.*, vol. 8, no. 5, pp. 902–915, Oct. 2014.
- [3] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. UAI*, Seattle, WA, USA, 2001, pp. 362–369.
- [4] J. Céspedes, P. M. Olmos, M. Sánchez-Fernández, and F. Perez-Cruz, "Expectation propagation detection for high-order high-dimensional MIMO systems," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2840–2849, Aug. 2014.
- [5] X. Tan, Y.-L. Ueng, Z. Zhang, X. You, and C. Zhang, "A low-complexity massive MIMO detection based on approximate expectation propagation," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7260–7272, Aug. 2019.
- [6] Y. Ge, X. Tan, Z. Ji, Z. Zhang, X. You, and C. Zhang, "Improving approximate expectation propagation massive MIMO detector with deep learning," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2145–2149, 2021.
- [7] Y. Zhang, Z. Wu, C. Li, Z. Zhang, X. You, and C. Zhang, "Expectation propagation detection with Neumann-series approximation for massive MIMO," in *Proc. IEEE SiPS*, Cape Town, South Africa, 2018.
- [8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. Philadelphia, PA: Johns Hopkins University Press, 2013.
- [9] L. Dai, X. Gao, X. Su, S. Han, C.-L. I, and Z. Wang, "Low-complexity soft-output signal detection based on Gauss-Seidel method for uplink multiuser large-scale MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, 2015.
- [10] Y. Wei, M.-M. Zhao, M. Hong, M.-j. Zhao, and M. Lei, "Learned conjugate gradient descent network for massive MIMO detection," *IEEE Trans. Signal Process.*, vol. 68, pp. 6336–6349, 2020.
- [11] K. Takeuchi and C.-K. Wen, "Rigorous dynamics of expectation-propagation signal detection via the conjugate gradient method," in *Proc. IEEE SPAWC*, Sapporo, Japan, Jul. 2017.
- [12] M. Khani, M. Alizadeh, J. Hoydis, and P. Fleming, "Adaptive neural signal detection for massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5635–5648, Aug. 2020.
- [13] *Study on 3D Channel Model for LTE*, 3GPP TR, 2015.
- [14] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
- [15] M. W. Seeger, "Expectation propagation for exponential families," University of California, Berkeley, CA, USA, Technical Report, 2006.
- [16] A. Kosasih, V. Onasis, V. Miloslavskaya, W. Hardjawana, V. Andrean, and B. Vucetic, "Graph neural network aided MU-MIMO detectors," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2540–2555, Sep. 2022.
- [17] J. J. D. Croz and N. J. Higham, "Stability of methods for matrix inversion," *IMA J. of Numerical Analysis*, vol. 12, no. 1, pp. 1–19, 1992.
- [18] V. Simoncini and D. B. Szyld, "Recent computational developments in Krylov subspace methods for linear systems," *Numerical Linear Algebra with Applications*, vol. 14, no. 1, pp. 1–59, 2007.
- [19] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," *Journal of Research of the National Bureau of Standards*, vol. 49, no. 1, 1952.