Matrix-Inversion-Free Expectation Propagation for Massive Connectivity

Rui Ma, Zheng Wang, Yongming Huang
School of Information Science and Engineering
Southeast University, Nanjing, China
Email: ruima@seu.edu.cn, wznuaa@gmail.com, huangym@seu.edu.cn

Zhen Gao
School of Information and Electronics
Beijing Institute of Technology, Beijing, China
Email: gaozhen16@bit.edu.cn

***Abstract*—This paper presents expectation propagation-conjugate gradient (EP-CG), a novel matrix-inversion-free EP framework designed for activity detection in massive connectivity scenarios. Conventional EP requires repeated inversions of large covariance matrices, becoming computationally prohibitive as the number of devices increases. In contrast, EP-CG computes the posterior mean via a few preconditioned CG iterations and estimates the required marginal variances using a Hutchinson diagonal estimator with Rademacher probe vectors. This approach reduces the cost per-iteration from $\mathcal{O}(N^3)$ to $\mathcal{O}(ULNR)$. By demonstration, we further theoretically specify the number of probe vectors required to achieve a desired level of estimation accuracy. Numerical results confirm our analysis: EP–CG achieves detection accuracy comparable to standard EP with a remarkable complexity reduction, even when the number of users is very large.**

***Index Terms*—Massive connectivity, expectation propagation, conjugate gradient, low complexity.**

I. Introduction

Massive connectivity is a defining feature of beyond-5G and 6G systems that support massive machine-type communications (mMTC) in smart cities, industrial Internet of things (IoT), logistics, and wearable/healthcare networks [1]. In these scenarios, a base station (BS) simultaneously serves a very large population of potential devices, while only a small and sporadically changing subset is active in each coherence block. Grant-free pilot transmission with compressed-sensing (CS) style recovery has therefore become popular for activity detection, where different methods have been developed [2]–[4]. Among them, expectation propagation (EP) is prospective due to its superior detection performance.

However, the conventional EP requires repeatedly constructing and inverting large covariance matrices, such that its computational complexity scales cubically with the dimension N. Moreover, the sensing matrix in massive connectivity scenarios is typically wide and under-determined, such that beneficial conditions like favorable propagation conditions are no longer hold [5]. Consequently, the off-diagonal entries of the corresponding Gram matrix become non-negligible, thereby making the complexity reduction of EP particularly challenging. Although [6] introduces a low complexity EP (LC-EP) method and finds statistical convergence property to achieve complexity reduction for the under-determined setting, yet those guarantees require both an independent and identically distributed (i.i.d.) Gaussian sensing matrix and an increased pilot length, which restricts its applicability in more general scenarios.

In this paper, we propose expectation propagation-conjugate gradient (EP-CG), a matrix-inversion-free variant of EP that replaces explicit inversions with a stochastic diagonal estimator together with a small number of preconditioned CG solves. Specifically, the diagonal of the posterior covariance is estimated via a Hutchinson estimator driven by a few probe vectors, while the EP posterior mean is obtained by solving the associated linear systems with CG. This design eliminates explicit matrix inversions, thereby dramatically lowering computational cost in the massive connectivity scenario. The per-iteration cost of EP–CG scales as $\mathcal{O}(ULNR)$ rather than the $\mathcal{O}(N^3)$ inversion cost of conventional EP with $U \ll N$, $L \ll N$, and $R \ll N$. Simulation results show that the proposed method achieves performance comparable to standard EP with a significant complexity reduction.

II. EP Algorithm for User Activity Detection

We consider a single-cell uplink with a single-antenna BS serving N potential devices, of which K are active. The channel from device n to the BS is $\tilde{h}_n \sim \mathcal{CN}(0, \beta_n)$, where β_n models path loss. Each active user sends a pilot of length L. The received signal $\tilde{\mathbf{y}} \in \mathbb{C}^{L\times 1}$ is

$$\tilde{\mathbf{y}} = \sum_{n=1}^{N} \tilde{\phi}_n a_n \tilde{h}_n + \tilde{\mathbf{w}} = \tilde{\mathbf{\Phi}}\tilde{\mathbf{x}} + \tilde{\mathbf{w}}, \tag{1}$$

where $\tilde{\phi}_n \in \mathbb{C}^{L\times 1}$ is the pilot sequence of user n, with i.i.d. entries $\tilde{\phi}_{l,n} \sim \mathcal{CN}(0, \frac{1}{L})$. $\tilde{\mathbf{\Phi}} = \left[\tilde{\phi}_1, \ldots . \tilde{\phi}_N\right] \in \mathbb{C}^{L\times N}$ is the pilot matrix. User activity is indicated by $a_n \in \{0, 1\}$, and the composite vector is $\tilde{\mathbf{x}} = \mathbf{a} \circ \tilde{\mathbf{h}} = [\tilde{x}_1, \ldots, \tilde{x}_N]^H$ with $\tilde{x}_n = a_n \tilde{h}_n$ ($\circ$ denotes the Hadamard product). $\tilde{\mathbf{w}} \in \mathbb{C}^{L\times 1} \sim \mathcal{CN}(\mathbf{0}, \sigma_{\tilde{w}}^2 \mathbf{I})$ is additive white Gaussian noise (AWGN). The transmit power of all devices is ρ. The complex-valued system (1) can be converted into an equivalent real-valued representation

$$\begin{bmatrix} \Re(\tilde{\mathbf{y}}) \\ \Im(\tilde{\mathbf{y}}) \end{bmatrix} = \begin{bmatrix} \Re(\tilde{\mathbf{\Phi}}) & -\Im(\tilde{\mathbf{\Phi}}) \\ \Im(\tilde{\mathbf{\Phi}}) & \Re(\tilde{\mathbf{\Phi}}) \end{bmatrix} \begin{bmatrix} \Re(\tilde{\mathbf{x}}) \\ \Im(\tilde{\mathbf{x}}) \end{bmatrix} + \begin{bmatrix} \Re(\tilde{\mathbf{w}}) \\ \Im(\tilde{\mathbf{w}}) \end{bmatrix}, \tag{2}$$

which can be further expressed as

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{w}. \tag{3}$$

Corresponding author: Zheng Wang (email: wznuaa@gmail.com)

The goal of BS is to detect active users by recovering $\mathbf{x}$ from the noisy observation $\mathbf{y}$, for which the maximum a posteriori (MAP) probability estimator is commonly employed. In particular, the output of the MAP estimator is given by

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}\in\mathbb{R}^{2N}} p(\mathbf{x}|\mathbf{y}) = \arg\max_{\mathbf{x}\in\mathbb{R}^{2N}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \tag{4}$$

with

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\mathbf{x}, \frac{\sigma_{\tilde{w}}^2}{2}\mathbf{I}), \tag{5}$$

$$p(\mathbf{x}) = \prod_{n=1}^{2N}\left[(1-\epsilon)\delta(x_n) + \epsilon\mathcal{N}(x_n|0, \frac{\beta_n}{2})\right]. \tag{6}$$

Here, $p(\mathbf{x})$ is the prior distribution and ϵ is the activity probability. However, the optimization problem in (4) is computationally difficult due to the discrete nature of the activity vector $\mathbf{a}$. To address this, EP is adopted to approximate the target distribution $p(\mathbf{x}|\mathbf{y})$ by a tractable Gaussian $q(\mathbf{x})$ [7].

$$q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\tilde{\mathbf{m}}, \tilde{\mathbf{V}}) \tag{7}$$

with

$$\tilde{\mathbf{V}} = \left(\sigma_w^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \tilde{\mathbf{v}}_1^{-1}\right)^{-1}, \tag{8}$$

$$\tilde{\mathbf{m}} = \tilde{\mathbf{V}}\left(\sigma_w^{-2}\mathbf{\Phi}^\top\mathbf{y} + \tilde{\mathbf{v}}_1^{-1}\tilde{\mathbf{m}}_1\right), \tag{9}$$

where $\tilde{\mathbf{m}}_1$ and $\tilde{\mathbf{v}}_1$ belong to the approximated prior distribution $q_{\text{prior}}(\mathbf{x}) \triangleq \mathcal{N}(\mathbf{x}|\tilde{\mathbf{m}}_1, \tilde{\mathbf{v}}_1)$. EP updates the pairs $(\tilde{\mathbf{m}}_1^t, \tilde{\mathbf{v}}_1^t)$ in each t-th iteration via moment matching. Specifically, at iteration t and for user n, form the cavity marginal

$$q_{\backslash\text{prior},n}^t(x_n) = \frac{q_n^t(x_n)}{q_{\text{prior},n}^t(x_n)} = \mathcal{N}(x_n|\tilde{m}_{\backslash 1,n}^t, \tilde{v}_{\backslash 1,n}^t), \tag{10}$$

where

$$\tilde{v}_{\backslash 1,n}^t = \left[\left(\tilde{V}_{nn}^t\right)^{-1} - \left(\tilde{v}_{1,n}^t\right)^{-1}\right]^{-1}, \tag{11}$$

$$\tilde{m}_{\backslash 1,n}^t = \tilde{v}_{\backslash 1,n}^i\left[\left(\tilde{V}_{nn}^t\right)^{-1}\tilde{m}_n^t - \left(\tilde{v}_{1,n}^t\right)^{-1}\tilde{m}_{1,n}^t\right]. \tag{12}$$

Next, compute the mean $E_q^t[x_n]$ and variance $V_q^t[x_n]$ of $\hat{q}_n^t(x_n)$, where $\hat{q}_n^t(x_n) = p(x_n)q_{\backslash\text{prior},n}^t(x_n)$.

$$E_q^t[x_n] = \frac{X_{1,n}^t}{X_{0,n}^t}, \tag{13}$$

$$V_q^t[x_n] = \frac{X_{2,n}^t}{X_{0,n}^t} - \left|E_q^t[x_n]\right|^2, \tag{14}$$

where $X_{m,n}^t$ (with $m = 0,1,2$) represents the moments of $\hat{q}_n^t(x_n)$, which can be computed as follows

$$X_{0,n}^t = (1-\epsilon)\mathcal{N}\left(0|\tilde{m}_{\backslash 1,n}^t, \tilde{v}_{\backslash 1,n}^t\right) + \epsilon\mathcal{N}\left(0|\tilde{m}_{1,n}^t, \beta_n + \tilde{v}_{\backslash 1,n}^t\right), \tag{15}$$

$$X_{1,n}^t = \epsilon\mathcal{N}\left(0 \mid \tilde{m}_{\backslash 1,n}^t, \beta_n + \tilde{v}_{\backslash 1,n}^t\right)\frac{\tilde{m}_{\backslash 1,n}^t\beta_n}{\beta_n + \tilde{v}_{\backslash 1,n}^t}, \tag{16}$$

$$X_{2,n}^t = \epsilon\mathcal{N}\left(0|\tilde{m}_{\backslash 1,n}^t, \beta_n + \tilde{v}_{\backslash 1,n}^t\right)\left(\left|\frac{\tilde{m}_{\backslash 1,n}^t\beta_n}{\beta_n + \tilde{v}_{\backslash 1,n}^t}\right|^2 + \frac{\beta_n\tilde{v}_{\backslash 1,n}^t}{\beta_n + \tilde{v}_{\backslash 1,n}^t}\right). \tag{17}$$

Finally, update the pairs $(\tilde{\mathbf{m}}_1^t, \tilde{\mathbf{v}}_1^t)$ such that $\hat{q}_n^{t+1}(x_n)$ has the same mean and variance with $q_n^{t+1}(x_n)$

$$\tilde{v}_{1,n}^{t+1} = \left[V_q^t[x_n]^{-1} - \left(\tilde{v}_{\backslash 1,n}^t\right)^{-1}\right]^{-1}, \tag{18}$$

$$\tilde{m}_{1,n}^{t+1} = \tilde{v}_{1,n}^{t+1}\left[V_q^t[x_n]^{-1}E_q^t[x_n] - \left(\tilde{v}_{\backslash 1,n}^t\right)^{-1}\tilde{m}_{\backslash 1,n}^t\right]. \tag{19}$$

III. EP-CG Algorithm for User Activity Detection

The key to avoiding the explicit inversion in EP is to bypass a direct formulation of (8). Since the values of $\tilde{\mathbf{V}}$ in (8) are only used in the mean update (9) and the site updates (11)-(12), explicit inversion can be entirely eliminated by reformulating these steps respectively.

A. Efficient Posterior Mean Update of $\tilde{\mathbf{m}}$ in (9) via CG

The posterior mean update in equation (9) depends only on the action of $\tilde{\mathbf{V}}$ on a vector, which is mathematically equivalent to solving a linear system. In particular, (9) can be rewritten as

$$\mathbf{A}\tilde{\mathbf{m}} = \mathbf{b}, \tag{20}$$

where $\mathbf{A} = \tilde{\mathbf{V}}^{-1} = \sigma_w^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \tilde{\mathbf{v}}_1^{-1}$ and $\mathbf{b} = \sigma_w^{-2}\mathbf{\Phi}^\top\mathbf{y} + \tilde{\mathbf{v}}_1^{-1}\tilde{\mathbf{m}}_1$. In this formulation, $\tilde{\mathbf{m}}$ can be efficiently obtained by solving (20) using CG method. This iterative solver avoids the explicit formation in (8), mitigating the high computational complexity associated with the covariance update.

B. Efficient Estimation of $\tilde{V}_{nn}$ for (11) and (12) via Hutchinson's Estimator

The cavity updates in (11)-(12) require only the diagonal entries of $\tilde{\mathbf{V}}$ since both expressions rely on the elements $\tilde{V}_{nn}^t$. This motivates us to approximate $\tilde{V}_{nn}^t$ with the estimates s_n^t for user n by exploiting the standard Hutchinson estimator [8].

Specifically, the Hutchinson construction works by generating R probe vectors $\mathbf{p}_r \in \mathbb{R}^N$ for $r = 1,\ldots,R$ with i.i.d. Rademacher entries (±1 with equal probability). For each probe vector, the matrix-vector product $\mathbf{z}_r = \tilde{\mathbf{V}}\mathbf{p}_r$ is computed. Here, we point out that each $\mathbf{z_r}$ can be obtained by solving a linear system

$$\mathbf{A}\mathbf{z_r} = \mathbf{p_r} \tag{21}$$

with $\mathbf{A} = \tilde{\mathbf{V}}^{-1}$, following the same approach as (20) to avoid explicit formation of $\tilde{\mathbf{V}}$. The approximation is then formed as

$$\mathbf{s}^t = \frac{1}{R}\sum_{r=1}^{R}\mathbf{p}_r \odot \mathbf{z}_r. \tag{22}$$

To quantify the accuracy of the proposed estimator, following the analysis in [9], we define the t-th iteration column–wise ratio $\rho_n^t = \frac{\left\|\tilde{\mathbf{v}}_n^t\right\|_2^2 - (\tilde{V}_{n,n}^t)^2}{(\tilde{V}_{n,n}^t)^2} = \frac{\sum_{n\neq j}(\tilde{V}_{nj}^t)^2}{(\tilde{V}_{nn}^t)^2}$ and denote its upper bound by $\bar{\rho}_n^t$, and let R^t denote the required number of probe vectors needed for the t-th iteration in EP, then we can arrive at the following Theorem.

Theorem 1. *For any desired accuracy $\varepsilon > 0$ and confidence level $\xi \in (0,1)$, the estimator's relative error holds*

$$\Pr\left(\left|\tilde{V}_{n,n} - s_n\right| \leq \varepsilon\left|\tilde{V}_{n,n}\right|\right) \geq 1 - \xi, \tag{23}$$

if the number of probe vectors R^t satisfies

$$R^t \;\geq\; \bar{\rho}_n^t \,\frac{2\ln(2/\xi)}{\varepsilon^2} \tag{24}$$

with $\bar{\rho}_n^t = \frac{1}{\sigma_w^4}\sum_{j\neq n}\left(\tilde{v}_{1,j}^t\right)^2$.

Proof. We begin the proof by analyzing the EP posterior covariance at iteration t. Applying the Woodbury matrix identity to $\tilde{\mathbf{V}}^t = (\sigma_w^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + (\tilde{\mathbf{v}}_1^t)^{-1})^{-1}$, we obtain the equivalent form

$$\tilde{\mathbf{V}}^t = \tilde{\mathbf{v}}_1^t - \tilde{\mathbf{v}}_1^t\mathbf{\Phi}^\top\left(\sigma_w^2\mathbf{I} + \mathbf{\Phi}\tilde{\mathbf{v}}_1^t\mathbf{\Phi}^\top\right)^{-1}\mathbf{\Phi}\tilde{\mathbf{v}}_1^t. \tag{25}$$

Define $\mathbf{\Sigma}^t = \sigma_w^2\mathbf{I} + \mathbf{\Phi}\tilde{\mathbf{v}}_1^t\mathbf{\Phi}^\top$, (25) simplifies to $\tilde{\mathbf{V}}^t = \tilde{\mathbf{v}}_1^t - \tilde{\mathbf{v}}_1^t\mathbf{\Phi}^\top(\mathbf{\Sigma}^t)^{-1}\mathbf{\Phi}\tilde{\mathbf{v}}_1^t$. Extracting the (n,j)-th entry of $\tilde{\mathbf{V}}^t$ yields the component-wise expression:

$$\tilde{V}_{nj}^t = \begin{cases} \tilde{v}_{1,n}^t - (\tilde{v}_{1,n}^t)^2\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}^t)^{-1}\boldsymbol{\phi}_n, & n = j \\ -\tilde{v}_{1,n}^t\tilde{v}_{1,j}^t\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}^t)^{-1}\boldsymbol{\phi}_j, & n \neq j \end{cases}. \tag{26}$$

In order to calculate ρ_n^t, it can be reformulated as

$$\rho_n^t = \frac{\sum_{j\neq n}(\tilde{v}_{1,n}^t)^2(\tilde{v}_{1,j}^t)^2\left(\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}^t)^{-1}\boldsymbol{\phi}_j\right)^2}{\left(\tilde{v}_{1,n}^t - (\tilde{v}_{1,n}^t)^2\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}^t)^{-1}\boldsymbol{\phi}_n\right)^2}. \tag{27}$$

$\mathbf{\Sigma}^t$ can be reformulated as $\mathbf{\Sigma}^t = \sigma_w^2\mathbf{I} + \mathbf{\Phi}\tilde{\mathbf{v}}_1^t\mathbf{\Phi}^\top = \sigma_w^2\mathbf{I} + \sum_j \tilde{v}_{1,j}^t\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top$, define $\mathbf{\Sigma}_{\backslash n}^t = \sigma_w^2\mathbf{I} + \sum_{j\neq n}\tilde{v}_{1,j}^t\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top$. The relationship between them is

$$\mathbf{\Sigma}^t = \mathbf{\Sigma}_{\backslash n}^t + \tilde{v}_{1,n}^t\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\top. \tag{28}$$

Apply Sherman–Morrison formula to (28) and obtain

$$(\mathbf{\Sigma}^t)^{-1} = (\mathbf{\Sigma}_{\backslash n}^t)^{-1} - \frac{\tilde{v}_{1,n}^t(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}}{1 + \tilde{v}_{1,n}^t\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_n}. \tag{29}$$

First, to evaluate the quadratic form $\boldsymbol{\phi}_n^\top\left(\mathbf{\Sigma}^t\right)^{-1}\boldsymbol{\phi}_n$ appearing in (26), we invoke (29) to rewrite $\left(\mathbf{\Sigma}^t\right)^{-1}$ and then yields

$$\begin{aligned}\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}^t)^{-1}\boldsymbol{\phi}_n &= \boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_n - \frac{\tilde{v}_{1,n}^t\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_n}{1 + \tilde{v}_{1,n}^t\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_n} \\ &= f_n^t - \frac{\tilde{v}_{1,n}^t(f_n^t)^2}{1 + \tilde{v}_{1,n}^t f_n^t} = \frac{f_n^t}{1 + \tilde{v}_{1,n}^t f_n^t},\end{aligned} \tag{30}$$

where $f_n^t = \boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_n$ is introduced for notational simplicity. We next compute the cross-term $\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}^t)^{-1}\boldsymbol{\phi}_j (n \neq j)$ in (26) and substitute (29) to it, we have

$$\begin{aligned}\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}^t)^{-1}\boldsymbol{\phi}_j &= \boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_j - \frac{\tilde{v}_{1,n}^t\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_j}{1 + \tilde{v}_{1,n}^t\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_n} \\ &= g_{nj}^t - \frac{\tilde{v}_{1,n}^t f_n^t g_{nj}^t}{1 + \tilde{v}_{1,n}^t t_i} = \frac{g_{nj}^t}{1 + \tilde{v}_{1,n}^t f_n^t},\end{aligned} \tag{31}$$

where $g_{nj}^t = \boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_j (n \neq j)$ is introduced for notational simplicity. Substitute (30) and (31) to (27), we can re-express ρ_n^t as

$$\rho_n^t = \sum_{j\neq n}(\tilde{v}_{1,j}^t)^2(g_{nj}^t)^2 = \sum_{j\neq n}(\tilde{v}_{1,j}^t)^2\left(\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_j\right)^2. \tag{32}$$

Since column has been normalized, i.e., $\|\phi_i\| = \|\phi_j\| = 1$, $\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_j$ is bounded by

$$\left|\boldsymbol{\phi}_n^\top(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\boldsymbol{\phi}_j\right| \leq \left\|(\mathbf{\Sigma}_{\backslash n}^t)^{-1}\right\|_2 \leq \frac{1}{\sigma_w^2}. \tag{33}$$

Substituting equation (33) into equation (32) yields the bound in equation (24). □

Algorithm 1 EP-CG Algorithm

```
Input: $\mathbf{y}$, $\mathbf{\Phi}$, $\beta_n$, $\sigma_w^2$, $\epsilon$, $\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_R}$, $\tilde{\mathbf{m}}_1^0$, $\tilde{\mathbf{v}}_1^0$
Output: $\tilde{\mathbf{V}}$, $\tilde{\mathbf{m}}$
for $t = 0, 1, 2, \ldots, T$ do
    Compute $\mathbf{A}^{t+1}$, $\mathbf{B}^{t+1}$ by (34) and (35)
    Evoke Function 1 to get $\mathbf{Z} = [\boldsymbol{z}_1\,|\boldsymbol{z}_2|\ldots|\boldsymbol{z}_R|\,\tilde{\mathbf{m}}]$
    Compute $s_n^t$ by (22)
    for $n = 1, 2, \ldots, N$ do
        Compute $\tilde{v}_{\backslash 1,n}^{t+1}$, $\tilde{m}_{\backslash 1,n}^{t+1}$ by (36) and (37)
        Compute $X_{0,n}^{t+1}$, $X_{1,n}^{t+1}$, $X_{2,n}^{t+1}$ by (15), (16) and (17)
        Compute $E_q^{t+1}[x_n]$, $V_q^{t+1}[x_n]$ by (14) (13)
        Compute $\tilde{v}_{1,n}^{t+1}$, $\tilde{m}_{1,n}^{t+1}$ by (18) and (19)
    end for
end for
```

This theorem provides a theoretical lower bound for R^t to achieve a desired estimation accuracy. Since ε and ξ are designed choices rather than known constants, (24) can serve as a design guideline. In practice, our simulations confirm that satisfactory performance is achieved with a moderate R^t (e.g., between 10 and 20), which is practically efficient.

C. Parallel Implementation and Enhancement

Since CG is invoked in both preceding subsections, it is natural to integrate a parallel CG solver into the overall EP–CG framework to solve the multiple linear systems jointly. To be specific, combining (20) and (21), we obtain the compact matrix equation $\mathbf{AZ} = \mathbf{B}$, where $\mathbf{A} \in \mathbb{R}^{N\times N}$ and $\mathbf{B} \in \mathbb{R}^{N\times(R+1)}$, which can be written as

$$\mathbf{A} = \sigma_w^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \tilde{\mathbf{v}}_1^{-1}, \tag{34}$$

$$\mathbf{B} = [\boldsymbol{p}_1\,|\boldsymbol{p}_2|\ldots|\boldsymbol{p}_R|\,\mathbf{b}]. \tag{35}$$

By applying parallel CG, we obtain the solution $\mathbf{Z} = [\mathbf{z}_1 \mid \mathbf{z}_2 \mid \cdots \mid \mathbf{z}_R \mid \tilde{\mathbf{m}}] \in \mathbb{R}^{N\times(R+1)}$, from which the desired posterior mean $\tilde{\mathbf{m}}$ is obtained directly as the last column, and the diagonal estimates $\mathbf{s}$ are computed via the estimator (22). Substituting these estimates into the original update equations leads to the following modified rules that replace equations (11) and (12).

$$\tilde{v}_{\backslash 1,n}^t = \left[\left(s_n^t\right)^{-1} - \left(\tilde{v}_{1,n}^t\right)^{-1}\right]^{-1}, \tag{36}$$

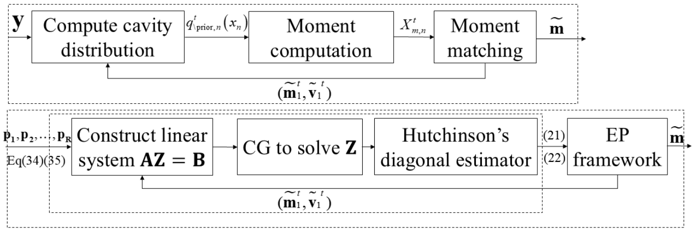
y
Compute cavity distribution
$q^t_{\backslash \text{prior},n}(x_n)$
Moment computation
$X^t_{m,n}$
Moment matching
$\widetilde{\mathbf{m}}$
$(\widetilde{\mathbf{m}}_1^t, \widetilde{\mathbf{v}}_1^t)$
$\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_R$
Eq(34)(35)
Construct linear system $\mathbf{AZ} = \mathbf{B}$
CG to solve $\mathbf{Z}$
Hutchinson's diagonal estimator
(21)
(22)
EP framework
$\widetilde{\mathbf{m}}$
$(\widetilde{\mathbf{m}}_1^t, \widetilde{\mathbf{v}}_1^t)$

Fig. 1. EP (upper) and proposed EP-CG method (bottom)

$$\tilde{m}^t_{\backslash 1,n} = \tilde{v}^i_{\backslash 1,n}\Big[\big(s^t_n\big)^{-1}\tilde{m}^t_n - \big(\tilde{v}^t_{1,n}\big)^{-1}\tilde{m}^t_{1,n}\Big]. \tag{37}$$

For faster convergence, we apply a diagonal preconditioner to the CG method. Specifically, the computational cost scales linearly with the number of CG iterations, denoted as U. Although $U \leq N$ is theoretically sufficient, practical scenarios involving large-scale systems or ill-conditioned matrices can significantly increase U [10]. To address this issue, a preconditioner $\mathbf{M} = \text{diag}(\sigma_w^{-2}\boldsymbol{I} + \tilde{\mathbf{v}}_1^{-1})$ is applied. Specifically, we solve the transformed system $\mathbf{A}'\mathbf{Z}' = \mathbf{B}', \mathbf{A}' = \mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}$ with $\mathbf{B}' = \mathbf{M}^{-1/2}\mathbf{B}$ and $\mathbf{Z}' = \mathbf{M}^{1/2}\mathbf{Z}$. Moreover, theoretical evidence shows that U is essentially independent of the system dimensions N and L [11]. Thus, even for high-dimensional systems, attaining a fixed residual tolerance does not require increasing U, thereby making CG particularly well-suited for large-scale inference. To summarize, the proposed EP-CG algorithm is outlined in Algorithm 1. Fig. 1 presents a comparative analysis of the computational workflows between the proposed EP-CG method and the conventional EP algorithm.

As for the complexity analysis, assume that the total number of iteration of EP is T. In the original EP algorithm, each iteration requires an explicit inversion of an $N \times N$ matrix, resulting in an overall computational complexity of $\mathcal{O}(TN^3)$. In contrast, the proposed EP–CG method replaces each matrix inversion with an iterative CG solve coupled with a Hutchinson stochastic diagonal estimator. The complexity of EP-CG is dominated by matrix-matrix multiplication, specifically the product of the sensing matrix $\mathbf{\Phi} \in L \times N$ with matrix $\mathbf{B} \in N \times (R+1)$. With CG performing U iterations per outer EP iteration, the overall complexity of the proposed EP–CG is therefore $\mathcal{O}(TULNR)$. EP-CG achieves lower computational complexity when the inequality $ULR < N^2$ holds. This condition is typically satisfied in massive connectivity scenarios because N is often in the hundreds or thousands, whereas realistic parameter choices such as $U \approx 100$, $L \ll N$, and $R \approx 10$ keep the left-hand side far below N^2. For a better understanding, the complexity comparisons among different algorithms is shown in Table I. Although the computational complexity of EP–CG remains higher than that of the LC-EP method in [6], EP–CG is applicable to sensing matrices with arbitrary distributions, whereas LC-EP is restricted to i.i.d. Gaussian sensing matrices.

Function 1 Parallel Conjugate Gradient Method

```
Initialize $\mathbf{Z}$ as a $N \times (R+1)$ matrix of all zeros
Initialize $\mathbf{E} = \mathbf{B}$ and $\mathbf{P} = \mathbf{M}^{-1}\mathbf{E}$ and $\mathbf{W} = \mathbf{M}^{-1}\mathbf{B}$
compute $\rho_r = \sum_{j=1}^{D} R_{n,r} \cdot W_{n,r}$ for $r = 1, \dots, R+1$
**for** $u = 1, 2, \dots, U$ **do**
    compute $\mathbf{\Psi} = \mathbf{AP}$
    compute $\pi_r = \sum_{j=1}^{N} P_{n,r} \cdot \Psi_{n,r}$ for $r = 1, \dots, R+1$
    compute $\gamma_r = \rho_r / \pi_r$ for $q = 1, \dots, R+1$
    update $\mathbf{Z} = \mathbf{Z} + \mathbf{P\Gamma}$, where $\mathbf{\Gamma} = \text{diag}\{\gamma\}$
    update $\mathbf{E} = \mathbf{E} - \mathbf{\Psi\Gamma}$
    **if** $\delta = \|R\|_F / \|B\|_F \leq e_{\max}$ **then**
        **return** $\mathbf{Z}$
    **end if**
    compute $\mathbf{W} = \mathbf{M}^{-1}\mathbf{E}$
    let $\rho_r^{\text{old}} = \rho_r$ for $r = 1, \dots, R+1$
    compute $\rho_r = \sum_{n=1}^{N} NR_{n,r} \cdot W_{n,r}$
    compute $\eta_r = \rho_r / \rho_r^{\text{old}}$ for $r = 1, \dots, R+1$
    update $\mathbf{P} = \mathbf{E} + \mathbf{PH}$, where $\mathbf{H} = \text{diag}\{\eta\}$
**end for**
**return** $\mathbf{Z}$
```

IV. Numerical Results

We consider a single–cell uplink with $N = 500$ users, each assigned an L–dimensional pilot. The large–scale channel gain of user n is modeled in decibels as $\beta_n = -128.1 - 36.7\log_{10}(d_n)$, where d_n denotes the distance between user n and the BS. User locations are randomized by drawing d_n uniformly from $[300, 500]\,\text{m}$. The receiver noise spectral density is set to $-169\,\text{dBm/Hz}$ and the activity probability is $\epsilon = 0.1$. We use the activity error rate (AER), defined as the percentage of misclassified users over the total number of users. We compare the proposed detector against several baselines: AMP with a soft–threshold denoiser [12], MMSE–denoiser AMP [3], standard EP [7], and OMP [13].

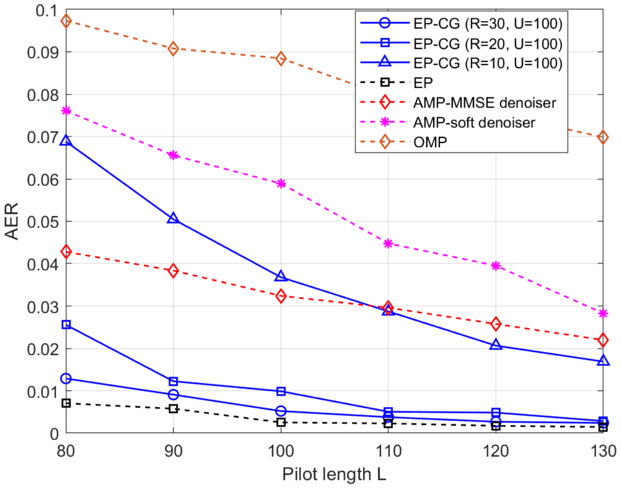

Fig. 2. Performance comparison of different methods based on AER versus increasing pilot length.

TABLE I

Computational Complexity Comparison for a $80 \times N$ massive connectivity system with $U = 100$ and $T = 5$

Algorithm	Complexity Order	$N = 500$	$N = 1000$	$N = 1500$	$N = 2000$
EP [7]	$\mathcal{O}\left(TN^3\right)$	$6.25 \cdot 10^8$	$5.00 \cdot 10^9$	$1.69 \cdot 10^{10}$	$4.00 \cdot 10^{10}$
LC-EP [6]	$\mathcal{O}\left(TN^2\right)$	$1.25 \cdot 10^6$	$5.00 \cdot 10^6$	$1.13 \cdot 10^7$	$2.00 \cdot 10^7$
EP-CG	$\mathcal{O}\left(TULNR\right), R = 10$	$2.00 \cdot 10^8$	$4.00 \cdot 10^8$	$6.00 \cdot 10^8$	$8.00 \cdot 10^8$
	$\mathcal{O}\left(TULNR\right), R = 20$	$4.00 \cdot 10^8$	$8.00 \cdot 10^8$	$1.20 \cdot 10^9$	$1.60 \cdot 10^9$

Fig. 2 illustrates the AER performance as the pilot-sequence length L varies from 80 to 130, with the transmit power fixed at $\rho = 10\,\text{dBm}$. Across all settings, the proposed EP-CG closely tracks the baseline EP. Moreover, as the number of probes R increases and the CG iterations U keeps constant, the performance steadily improves. This behavior occurs because a larger R reduces the variance of the estimation error from the Hutchinson estimator, thereby enhancing the precision of the diagonal approximation. Intuitively, with sufficiently large R, the stochastic estimation becomes accurate enough to make EP-CG's AER approach that of standard EP.

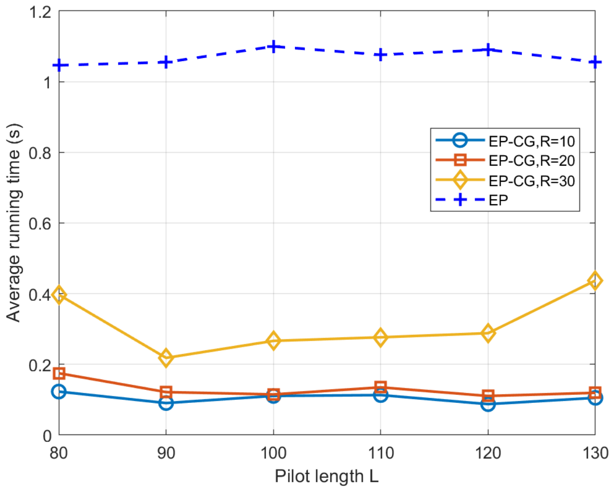

Fig. 3. Algorithm running time versus increasing pilot length.

Fig. 3 illustrates the running time versus the pilot length L with $N = 500, U = 100$. While EP maintains a high computational baseline due to explicit matrix inversions, EP-CG maintains a low and stable cost across varying L. These results confirm that replacing explicit solves with the proposed CG-based framework significantly improves efficiency.

V. Conclusion

This paper presents a novel EP-CG algorithm, a matrix-inversion-free variant of expectation propagation tailored for massive connectivity. By replacing explicit dense solves with a small number of preconditioned CG iterations and estimating only the required marginal variances via a Hutchinson diagonal estimator, EP-CG eliminates cubic-time inversions. Simulation results confirm that the proposed EP-CG achieves low complexity while preserving detection accuracy.

Acknowledgment

This work was supported in part by the National Key R&D Program of China under Grants No.2023YFC2205501, in part by National Science and Technology Major Projects of China under Grant 2025ZD1301800, and in part by National Natural Science Foundation of China under Grants No.62371124.

References

[1] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. De Carvalho, “Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of things,” *IEEE Signal Processing Mag.*, vol. 35, no. 5, pp. 88–99, 2018.

[2] J. Huang, L. Xiao, S. Li, J. Zhou, and T. Jiang, “Joint activity and data detection for asynchronous grant-free access in NTN,” *Chinese J. Electron.*, vol. 34, no. 4, pp. 1226–1232, 2025.

[3] L. Liu and W. Yu, “Massive connectivity with massive MIMO—part I: Device activity detection and channel estimation,” *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, 2018.

[4] Y. Cai, C. Dong, Z. Zhang, W. Xu, and K. Niu, “EP-based turbo receiver for single-carrier index modulation: A vector-wise implementation,” *IEEE Trans. Commun.*, vol. 73, no. 5, pp. 3214–3226, May 2025.

[5] L. Liu and W. Yu, “Massive connectivity with massive MIMO—part II: Achievable rate characterization,” *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2947–2959, 2018.

[6] R. Ma, Y. Ma, J. Zhu, Z. Wang, and Y. Cai, “A low complexity expectation propagation algorithm for active user detection for massive connectivity,” in *2025 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Abu Dhabi, United Arab Emirates, Jul. 2025, pp. 1114–1118.

[7] J. Ahn, B. Shim, and K. B. Lee, “EP-based joint active user detection and channel estimation for massive machine-type communications,” *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5178–5189, 2019.

[8] P. Dharangutte and C. Musco, “A tight analysis of Hutchinson's diagonal estimator,” in *Proc. Symp. Simplicity in Algorithms (SOSA)*, Florence, Italy, Jan. 2023, pp. 353–364.

[9] E. Hallman, I. C. F. Ipsen, and A. K. Saibaba, “Monte carlo methods for estimating the diagonal of a real symmetric matrix,” *SIAM J. Matrix Anal. Appl.*, vol. 44, no. 1, pp. 240–269, 2023.

[10] M. R. Hestenes and E. Stiefel, “Methods of conjugate gradients for solving linear systems,” *J. Research of the National Bureau of Standards*, vol. 49, no. 6, pp. 409–436, 1952.

[11] A. Lin, A. H. Song, B. Bilgic, and D. Ba, “Covariance-free sparse Bayesian learning,” *IEEE Trans. Signal Process.*, vol. 70, pp. 3818–3831, 2022.

[12] G. Hannak, M. Mayer, A. Jung, G. Matz, and N. Goertz, “Joint channel estimation and activity detection for multiuser communication systems,” in *2015 IEEE International Conference on Communications Workshops (ICCW)*, London, United Kingdom, Jun. 2015, pp. 2086–2091.

[13] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec 2007.