

Message Passing Receiver Design for Uplink Grant-Free SCMA

Fan Wei and Wen Chen

Department of Electronic Engineering, Shanghai Jiao Tong University, China

Email: {weifan89; wenchen}@sjtu.edu.cn

Abstract—In this paper, we consider the uplink grant-free SCMA detection where the base station blindly decodes the users' data without knowing the user activity and their channel state information. In order to shorten the pilots overhead, a message passing receiver that jointly performs the channel estimation, user detection and data decoding is proposed. We formulate the belief propagation (BP) framework based on the factor graph of SCMA. However, the direct use of BP for multi-variable problem is prohibitively complex. Motivated by the idea of approximate inference, expectation propagation is used to project the intractable distributions into Gaussian families such that a linear complexity decoder is obtained. Simulation results show that the proposed algorithm has a better performance compared with the existing methods.

Index Terms—SCMA, grant-free, message passing, joint channel estimation and data detection, expectation propagation.

I. INTRODUCTION

The fifth generation (5G) wireless networks require a bandwidth-efficient multiuser communication system to meet the massive connectivity requirements. Sparse code multiple access (SCMA) is such a non-orthogonal multiple access technique that can address the challenges in the future network [1]. Benefits from the shaping gain of the multi-dimensional codewords, SCMA has a better BER performance compared with other low code rate spreading techniques such as low density signature (LDS). At the receiver, a modest complexity message passing algorithm (MPA) is utilized for decoding and several recent works are proposed to further reduce the decoding complexity of SCMA [5], [6].

To reduce the latency and pilots overhead, the concept of contention based grant-free multiple access for SCMA is proposed in [2], [3]. In the contention based grant-free multiple access, the active users transmit their data without grant from the base station (BS) and the BS blindly decodes the users' data without knowing the user activity and their channel state information (CSI). Meanwhile, according to the statistical data of mobile traffic [4], the number of simultaneously active users in a network is limited compared with the massive connected users. Consequently, the user activity detection can be regarded as a sparse signals recovery problem.

By formulating the factor graph of SCMA, we propose a message passing receiver that performs joint channel estimation and data/user detection. Since the receiver explores not only the pilot symbols but also the transmitted data symbols in estimating the CSI and the user activity, the pilots overhead are reduced. To reduce the complexity of belief propagation (BP),

an approximate inference based on expectation propagation (EP) [7] is proposed to approximate the intractable distribution into some simpler forms such as Gaussian families. The EP based receiver has a linear complexity and thus results in a low latency. We note that the recent work [6] on low complexity receiver design is also based on EP. However, in their work, perfect CSI is available and all users are assumed to be active simultaneously. In this paper, the latter two information are not available at the receiver and our task is therefore more challenging.

Notations: Lowercase letters x , bold lowercase letters \mathbf{x} and bold uppercase letters \mathbf{X} denote scalars, column vectors and matrices, respectively. We use $(\cdot)^*$ and $(\cdot)^T$ to denote complex conjugate, matrix transpose. $\mathcal{CN}(x; \tau, v)$ denotes the complex Gaussian distribution with mean τ and variance v . $\text{diag}(\mathbf{x})$ is the diagonal matrix with the diagonal entries being vector \mathbf{x} . $\xi \setminus k$ means the set ξ with element k being excluded and $\langle f(x) \rangle_{g(x)} = E_{g(x)}\{f(x)\}$.

II. SYSTEM MODEL

A. SCMA Transceiver Structure

Consider an uplink grant-free SCMA system with K transmitted users. In the encoder, the coded bits \mathbf{c}_k for user k are directly mapped to the multi-dimensional SCMA codewords \mathbf{x}_k where $\mathbf{x}_k = (x_{1k}, x_{2k}, \dots, x_{Nk})^T$ is an N dimensional signal constellation point. The symbols in each dimension are transmitted to one OFDMA subcarrier shared by the other users in the system. To reduce the multiuser interference in each subcarrier, the codewords \mathbf{x}_k are constructed to be sparse such that only d_v of N dimensions have non-zero symbols while the remaining ones are set to be zero.

On the receiver side, the received signal at time slot t in the n th subcarrier can be written as,

$$y_{tn} = \sum_{k=1}^K h_{nk} x_{tnk} + z_n, \quad (1)$$

where x_{tnk} are the transmitted signals from user k which comprising both known pilot symbols and unknown SCMA codewords, h_{nk} is the unknown channel impulse response (CIR) to be estimated for user k in the n th subcarrier and we assume that h_{nk} is the frequency-selective block-fading channel in this paper. Finally, z_n is the additive complex Gaussian white noise with distribution $\mathcal{CN}(0, \sigma^2)$. Notice that due to the sparse structure of the SCMA codewords, the

symbols x_{tnk} from some users may be zero. Therefore, only d_c instead of the whole K users are collided with each other in each subcarrier.

Writing the signals from N subcarriers in a matrix form, we have,

$$\mathbf{y}_t = \sum_{k=1}^K \text{diag}(\mathbf{h}_k) \mathbf{x}_{tk} + \mathbf{z}, \quad (2)$$

where $\mathbf{y}_t = (y_{t1}, y_{t2}, \dots, y_{tN})^T$, $\mathbf{h}_k = (h_{1k}, h_{2k}, \dots, h_{Nk})^T$ and $\mathbf{z} = (z_1, z_2, \dots, z_N)^T$. The transmitted SCMA codewords \mathbf{x}_{tk} in each time slot are chosen randomly from a predefined alphabet set \mathcal{X} with size M . As the user activity detection is also an interest in this paper, we introduce an augmented alphabet set $\mathcal{X}^+ = \mathcal{X} \cup \{0\}$ with size $|\mathcal{X}^+| = M + 1$ since an inactive user is equivalent to be transmitting zero symbols all the time.

B. Factor Graph Representation

From (2), the joint probability density function (pdf) is given by $p(\mathbf{C}, \mathbf{X}, \mathbf{H}, \mathbf{y})$ (time index t is dropped here for notation simplification), where $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K]$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K]$ and $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_K]$ are the collections of coded bits, SCMA codewords and CIRs from all users, respectively.

Based on the observation signal \mathbf{y} as well as pilot symbols, the SCMA decoder tries to find the *maximum a posterior* (MAP) estimation for each bit c_{kl} ,

$$\hat{c}_{kl} = \arg \max p(c_{kl} | \mathbf{y}), \quad (3)$$

where c_{kl} is the l th coded bits for user k and $p(c_{kl} | \mathbf{y})$ is given by,

$$p(c_{kl} | \mathbf{y}) \propto \sum_{\mathbf{C} \setminus c_{kl}, \mathbf{X}} \int p(\mathbf{C}, \mathbf{X}, \mathbf{H}, \mathbf{y}) d\mathbf{H}, \quad (4)$$

A direct computation of (4) involves marginalization of discrete variables \mathbf{C} and \mathbf{X} which is prohibitively complex when the number of users K is large.

With the observation that $\mathbf{C} \rightarrow \mathbf{X} \rightarrow \mathbf{y}$ forms a Markov chain and the CIRs \mathbf{H} are independent of \mathbf{C} and \mathbf{X} , the joint pdf can be factorized as follows,

$$p(\mathbf{C}, \mathbf{X}, \mathbf{H}, \mathbf{y}) = p(\mathbf{C})p(\mathbf{X} | \mathbf{C})p(\mathbf{y} | \mathbf{X}, \mathbf{H})p(\mathbf{H}). \quad (5)$$

In (5), $p(\mathbf{C})$ denotes the *priori* distribution of users' coded bits and $p(\mathbf{X} | \mathbf{C})$ is given by,

$$p(\mathbf{X} | \mathbf{C}) = \prod_k p(\mathbf{x}_k | \mathbf{c}_k), \quad (6)$$

where each term $p(\mathbf{x}_k | \mathbf{c}_k)$ represents the mapping function of SCMA encoder. Based on (2), the likelihood function $p(\mathbf{y} | \mathbf{X}, \mathbf{H})$ can be written as,

$$p(\mathbf{y} | \mathbf{X}, \mathbf{H}) \propto \exp \left\{ -\frac{1}{\sigma^2} |\mathbf{y} - \sum_{k=1}^K \text{diag}(\mathbf{h}_k) \mathbf{x}_k|^2 \right\}. \quad (7)$$

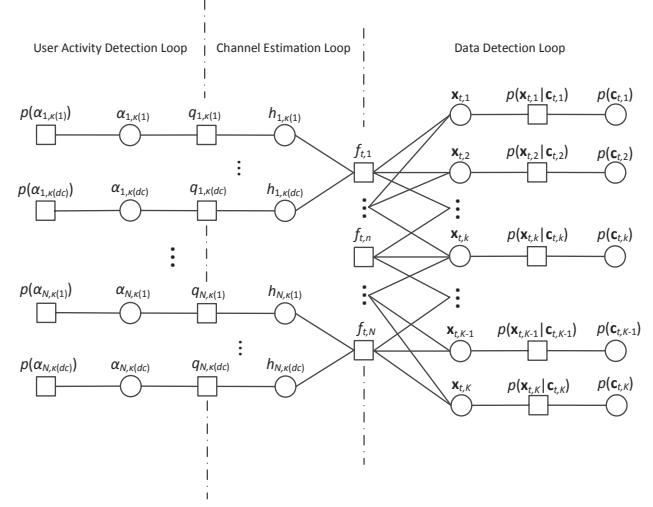


Fig. 1: Factor graph representation of the SCMA system

For channel \mathbf{H} , a non-stationary zero mean complex Gaussian *priori* distribution is considered in this paper,

$$\begin{aligned} p(\mathbf{H}) &= \prod_k q(\mathbf{h}_k | \boldsymbol{\alpha}_k) \\ &= \prod_k \prod_n \mathcal{CN}(h_{nk}; 0, \alpha_{nk}^{-1}), \end{aligned} \quad (8)$$

where α_{nk} is precision parameter and modeled by a Gamma distribution given by,

$$p(\alpha_{nk}; a, b) = \text{Gama}(\alpha_{nk} | a, b). \quad (9)$$

After integrating the variable α_{nk} , $p(h_{nk}; a, b)$ can be written as,

$$\begin{aligned} p(h_{nk}; a, b) &= \int q(h_{nk} | \alpha_{nk}) p(\alpha_{nk}; a, b) d\alpha_{nk} \\ &= \text{St}(h_{nk}; \mu, \nu, \lambda), \end{aligned} \quad (10)$$

where the Student's t-distribution $\text{St}(h_{nk}; \mu, \nu, \lambda)$ is given by,

$$\begin{aligned} \text{St}(h_{nk}; \mu, \nu, \lambda) &= \frac{\Gamma(\frac{\nu}{2} + 1)}{\Gamma(\frac{\nu}{2})} \frac{2\lambda}{\pi\nu} \\ &\times \left[1 + \frac{2\lambda}{\nu} |h_{nk} - \mu|^2 \right]^{-(\frac{\nu}{2} + 1)} \end{aligned} \quad (11)$$

with $\mu = 0$, $\lambda = \frac{a}{b}$ and $\nu = 2a$. In practice, a non-informative *priori* for a and b is assumed and we choose $a = 10^{-7}$ and $b = 10^{-7}$ in this paper. The Student's t-distribution exhibits heavy tails [11] and this property makes h_{nk} favour sparse solution such that most of h_{nk} in \mathbf{H} are near zero values. Note that the distribution of the active users in a network is sparse and an inactive user is equivalent to have zero CIRs.

Based on the factorization (5), the factor graph representation of the SCMA system is plotted in Fig. 1 where square nodes are used to denote functions (e.g., f_{tn}) and circular nodes are used to denote variables (e.g., h_{nk}). As can be observed in Fig. 1, due to the sparse structure of SCMA

codewords, each user chooses only part of subcarriers to transmit data and only part of users are collided in each subcarrier. We define F_n to be the set of collision users in subcarrier n while V_k to be the set of subcarriers for user k to transmit data. The factor graph is divided into three loops for data detection, channel estimation and active user detection, respectively. In the next section, we formulate a joint channel estimation and data/active user detection scheme based on this factor graph representation.

III. EP BASED MPA RECEIVER

A. Joint Detection Based on BP and EP

We will discuss the data detection loop in Fig. 1 first. Based on the BP rule [8], the message from function node f_{tn} to variable node \mathbf{x}_{tk} can be written as,

$$I_{f_{tn} \rightarrow \mathbf{x}_{tk}}(x_{tnk}) = \sum_{\mathbf{x}_i: i \in F_n \setminus k} I_{\mathbf{x}_{ti} \rightarrow f_{tn}}(x_{tni}) \cdot \int f_{tn}(\mathbf{X}_{tn}, \mathbf{H}_n) \prod_{i \in F_n} I_{h_{ni} \rightarrow f_{tn}}(h_{ni}) \prod_{i \in F_n} dh_{ni}, \quad (12)$$

where $I_{\mathbf{x}_{ti} \rightarrow f_{tn}}(x_{tni})$ and $I_{h_{ni} \rightarrow f_{tn}}(h_{ni})$ are the extrinsic message passed from nodes \mathbf{x}_{ti} and h_{ni} to f_{tn} , respectively. \mathbf{X}_{tn} and \mathbf{H}_n respective denote the symbols and CIRs collided in subcarrier n and $f_{tn}(\mathbf{X}_{tn}, \mathbf{H}_n)$ is given by,

$$f_{tn}(\mathbf{X}_{tn}, \mathbf{H}_n) \propto \exp \left\{ -\frac{1}{\sigma^2} |y_{tn} - \sum_{k \in F_n} h_{nk} x_{tnk}|^2 \right\}. \quad (13)$$

Since the function node f_{tn} involves a mixture of discrete variables x_{tnk} and continuous variables h_{nk} , the multiple integration as well as the marginalization make the direct computation of (12) prohibitively complex.

To reduce the computation complexity of (12), in [9], [10], a belief propagation mean field (BP-MF) message passing (or variational message passing) approach is proposed based on variational Bayesian (VB) inference [11]. Although BP-MF has a simple update rule, the interference cancellation structure (e.g., equation (14) in [10]) only involves mean value of the interferences while the covariance is not being considered. This makes BP-MF perform poor in estimating the LLR of x_{tnk} when the interferences exist. In [12], Gaussian approximation of the interferences based on central-limit theorem is proposed. While central-limit theorem is effective in large scale MIMO-OFDM system, it may result in a large performance degradation in SCMA since the number of collision users in each subcarrier is limited (d_c compared with the total K users) due to the sparse structure of SCMA codewords.

In this paper, instead of using central-limit theorem, the distribution of each interference item $h_{nu} x_{tnu}$ is projected into Gaussian families separately based on expectation propagation [7]. Expectation propagation belongs to a class of approximate inference that the intractable distributions are always approximated with some simpler distributions (e.g., Gaussian

families) by minimizing the Kullback-Leibler divergence,

$$D_{KL}(p(x) || q(x)) = \int p(x) \log \frac{p(x)}{q(x)} dx, \quad (14)$$

where $p(x)$ is the original distribution and $q(x)$ is the approximated distribution.

Define $u_{tnk} = h_{nk} x_{tnk}$. With the extrinsic messages $I_{\mathbf{x}_{tk} \rightarrow f_{tn}}(x_{tnk})$ and $I_{h_{nk} \rightarrow f_{tn}}(h_{nk})$, the distribution of variable u_{tnk} is given by (15) (see Appendix) where we have assumed that $I_{h_{nk} \rightarrow f_{tn}}(h_{nk}) \sim \mathcal{CN}(h_{nk}; \tau_{h_{nk} \rightarrow f_{tn}}, v_{h_{nk} \rightarrow f_{tn}})$ which will be shown later in (53). As can be seen from (15), the distribution of u_{tnk} is a mixture of Gaussian and is discontinuous at $u_{tnk} = 0$. Consequently, we resort to EP method to project $I_{u_{tnk} \rightarrow f_{tn}}(u_{tnk})$ into Gaussian distribution.

With $I_{f_{tn} \rightarrow u_{tnk}}(u_{tnk}) \sim \mathcal{CN}(u_{tnk}; \tau_{f_{tn} \rightarrow u_{tnk}}, v_{f_{tn} \rightarrow u_{tnk}})$ as an output information from f_{tn} shown later in (26), the belief [8] of variable u_{tnk} is given by (16) with $\beta(x_{tnk})$, $\tau_{u_{tnk}}$ and $v_{u_{tnk}}$ computed in (17)-(19) and C being a normalization constant.

$$v_{u_{tnk}} = \frac{v_{f_{tn} \rightarrow u_{tnk}} v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2}{v_{f_{tn} \rightarrow u_{tnk}} + v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2}. \quad (19)$$

By EP principle, $b(u_{tnk})$ is projected into a Gaussian distribution $\hat{b}(u_{tnk})$ so that $D_{KL}[b(u_{tnk}) || \hat{b}(u_{tnk})]$ is minimized. The result is reduced to moment matching such that,

$$\hat{b}(u_{tnk}) = \mathcal{CN}(u_{tnk}; \hat{\tau}_{u_{tnk}}, \hat{v}_{u_{tnk}}), \quad (20)$$

$$\hat{\tau}_{u_{tnk}} = \sum_{x_{tnk}} \beta(x_{tnk}) \tau_{u_{tnk}}, \quad (21)$$

$$\hat{v}_{u_{tnk}} = \sum_{x_{tnk}} \beta(x_{tnk}) (|\tau_{u_{tnk}}|^2 + v_{u_{tnk}}) - |\hat{\tau}_{u_{tnk}}|^2, \quad (22)$$

Since $I_{f_{tn} \rightarrow u_{tnk}}(u_{tnk}) \sim \mathcal{CN}(u_{tnk}; \tau_{f_{tn} \rightarrow u_{tnk}}, v_{f_{tn} \rightarrow u_{tnk}})$, we have,

$$\begin{aligned} I_{u_{tnk} \rightarrow f_{tn}}(u_{tnk}) &= \frac{\hat{b}(u_{tnk})}{I_{f_{tn} \rightarrow u_{tnk}}(u_{tnk})} \\ &\propto \mathcal{CN}(u_{tnk}; \tau_{u_{tnk} \rightarrow f_{tn}}, v_{u_{tnk} \rightarrow f_{tn}}), \end{aligned} \quad (23)$$

where

$$\tau_{u_{tnk} \rightarrow f_{tn}} = v_{u_{tnk} \rightarrow f_{tn}} \left(\frac{\hat{\tau}_{u_{tnk}}}{\hat{v}_{u_{tnk}}} - \frac{\tau_{f_{tn} \rightarrow u_{tnk}}}{v_{f_{tn} \rightarrow u_{tnk}}} \right), \quad (24)$$

$$v_{u_{tnk} \rightarrow f_{tn}} = \left(\frac{1}{\hat{v}_{u_{tnk}}} - \frac{1}{v_{f_{tn} \rightarrow u_{tnk}}} \right)^{-1}. \quad (25)$$

Given the extrinsic message of interference $u_{tni}, i \in F_n \setminus k$ follow independent Gaussian distributions, $u_{tnk} = y_{tn} - \sum_{i \in F_n \setminus k} u_{tni}$ is also a Gaussian variable such that,

$$I_{f_{tn} \rightarrow u_{tnk}}(u_{tnk}) \sim \mathcal{CN}(u_{tnk}; \tau_{f_{tn} \rightarrow u_{tnk}}, v_{f_{tn} \rightarrow u_{tnk}}), \quad (26)$$

where

$$\tau_{f_{tn} \rightarrow u_{tnk}} = y_{tn} - \sum_{i \in F_n \setminus k} \tau_{u_{tni} \rightarrow f_{tn}}, \quad (27)$$

$$v_{f_{tn} \rightarrow u_{tnk}} = \sigma^2 + \sum_{i \in F_n \setminus k} v_{u_{tni} \rightarrow f_{tn}}. \quad (28)$$

$$I_{u_{tnk} \rightarrow f_{tn}}(u_{tnk}) \propto \begin{cases} \sum_{x_{tnk}} I_{\mathbf{x}_{tk} \rightarrow f_{tn}}(x_{tnk}) |x_{tnk}| \mathcal{CN}(u_{tnk}; \tau_{h_{nk} \rightarrow f_{tn}} x_{tnk}, v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2), & x_{tnk} \neq 0; \\ I_{\mathbf{x}_{tk} \rightarrow f_{tn}}(x_{tnk}), & x_{tnk} = 0. \end{cases} \quad (15)$$

$$\begin{aligned} b(u_{tnk}) &= I_{u_{tnk} \rightarrow f_{tn}}(u_{tnk}) I_{f_{tn} \rightarrow u_{tnk}}(u_{tnk}) \\ &= \begin{cases} \sum_{x_{tnk}} \beta(x_{tnk}) \mathcal{CN}(u_{tnk}; \tau_{u_{tnk}}, v_{u_{tnk}}), & x_{tnk} \neq 0; \\ C^{-1} I_{\mathbf{x}_{tk} \rightarrow f_{tn}}(x_{tnk}) \mathcal{CN}(u_{tnk}; \tau_{f_{tn} \rightarrow u_{tnk}}, v_{f_{tn} \rightarrow u_{tnk}}), & x_{tnk} = 0. \end{cases} \end{aligned} \quad (16)$$

$$\beta(x_{tnk}) = C^{-1} I_{\mathbf{x}_{tk} \rightarrow f_{tn}}(x_{tnk}) |x_{tnk}| \mathcal{CN}(\tau_{f_{tn} \rightarrow u_{tnk}}; \tau_{h_{nk} \rightarrow f_{tn}} x_{tnk}, v_{f_{tn} \rightarrow u_{tnk}} + v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2), \quad (17)$$

$$\tau_{u_{tnk}} = \frac{\tau_{h_{nk} \rightarrow f_{tn}} v_{f_{tn} \rightarrow u_{tnk}} x_{tnk} + \tau_{f_{tn} \rightarrow u_{tnk}} v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2}{v_{f_{tn} \rightarrow u_{tnk}} + v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2}, \quad (18)$$

Now by BP rule, the message sent from function node f_{tn} to variable node \mathbf{x}_{tk} can be updated as,

$$\begin{aligned} I_{f_{tn} \rightarrow \mathbf{x}_{tk}}(x_{tnk}) &= \int I_{f_{tn} \rightarrow u_{tnk}}(u_{tnk}) I_{h_{nk} \rightarrow f_{tn}}(h_{nk}) dh_{nk} \\ &\propto \exp\{-\Delta_{f_{tn} \rightarrow x_{tnk}}(x_{tnk})\}, \end{aligned} \quad (29)$$

where

$$\begin{aligned} \Delta_{f_{tn} \rightarrow x_{tnk}}(x_{tnk}) &= \frac{|\tau_{f_{tn} \rightarrow u_{tnk}} - \tau_{h_{nk} \rightarrow f_{tn}} x_{tnk}|^2}{v_{f_{tn} \rightarrow u_{tnk}} + v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2} \\ &+ \ln(v_{f_{tn} \rightarrow u_{tnk}} + v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2), \end{aligned} \quad (30)$$

and we have assumed that the message $I_{h_{nk} \rightarrow f_{tn}}(h_{nk})$ follows $\mathcal{CN}(h_{nk}; \tau_{h_{nk} \rightarrow f_{tn}}, v_{h_{nk} \rightarrow f_{tn}})$ which will be given in (53). Notice that in computing (30), not only the mean value but also the covariance of interferences are involved.

With $I_{f_{tm} \rightarrow \mathbf{x}_{tk}}(x_{tnk})$ and the factor graph in Fig. 1, the message $I_{\mathbf{x}_{tk} \rightarrow f_{tn}}(x_{tnk})$ can be calculated according to the BP rule,

$$\begin{aligned} I_{\mathbf{x}_{tk} \rightarrow f_{tn}}(x_{tnk}) &= \prod_{m \neq n} I_{f_{tm} \rightarrow \mathbf{x}_{tk}}(x_{tmk}) \\ &\propto \exp\left\{-\sum_{m \in V_k \setminus n} \Delta_{f_{tm} \rightarrow x_{tmk}}(x_{tmk})\right\}. \end{aligned} \quad (31)$$

Now we discuss the channel estimation loop. As in (29), the message passed from function node f_{tn} to variable node h_{nk} can be updated as,

$$\begin{aligned} I_{f_{tn} \rightarrow h_{nk}}(h_{nk}) &= \sum_{x_{tnk}} I_{\mathbf{x}_{tk} \rightarrow f_{tn}}(x_{tnk}) I_{f_{tn} \rightarrow u_{tnk}}(u_{tnk}) \\ &= \sum_{x_{tnk}} I_{\mathbf{x}_{tk} \rightarrow f_{tn}}(x_{tnk}) \\ &\cdot \mathcal{CN}(h_{nk} x_{tnk}; \tau_{f_{tn} \rightarrow u_{tnk}}, v_{f_{tn} \rightarrow u_{tnk}}), \end{aligned} \quad (32)$$

which is a mixture of Gaussian distribution. To get a simpler form for $I_{f_{tn} \rightarrow h_{nk}}(h_{nk})$, again we project $I_{f_{tn} \rightarrow h_{nk}}(h_{nk})$ into Gaussian distribution by EP. The belief of h_{nk} is given by,

$$\begin{aligned} b(h_{nk}) &= I_{f_{tn} \rightarrow h_{nk}}(h_{nk}) I_{h_{nk} \rightarrow f_{tn}}(h_{nk}) \\ &= \sum_{x_{tnk}} \beta(x_{tnk}) \mathcal{CN}(h_{nk}; \tilde{\tau}_{h_{nk}}, \tilde{v}_{h_{nk}}), \end{aligned} \quad (33)$$

where we have assumed $I_{h_{nk} \rightarrow f_{tn}}(h_{nk})$ follows Gaussian distribution $\mathcal{CN}(h_{nk}; \tau_{h_{nk} \rightarrow f_{tn}}, v_{h_{nk} \rightarrow f_{tn}})$ which will be shown later in (53). $\beta(x_{tnk})$, $\tilde{\tau}_{h_{nk}}$ and $\tilde{v}_{h_{nk}}$ are given by (34)-(36) and C is a normalization constant.

$$\tilde{v}_{h_{nk}} = \frac{v_{f_{tn} \rightarrow u_{tnk}} v_{h_{nk} \rightarrow f_{tn}}}{v_{f_{tn} \rightarrow u_{tnk}} + v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2}. \quad (36)$$

By EP principle, after projecting $b(h_{nk})$ into Gaussian distribution, we have,

$$\hat{b}(h_{nk}) = \mathcal{CN}(h_{nk}; \hat{\tau}_{h_{nk}}, \hat{v}_{h_{nk}}), \quad (37)$$

where by moment matching,

$$\hat{\tau}_{h_{nk}} = \sum_{x_{tnk}} \beta(x_{tnk}) \tilde{\tau}_{h_{nk}}, \quad (38)$$

$$\hat{v}_{h_{nk}} = \sum_{x_{tnk}} \beta(x_{tnk}) (|\tilde{\tau}_{h_{nk}}|^2 + \tilde{v}_{h_{nk}}) - |\hat{\tau}_{h_{nk}}|^2. \quad (39)$$

Since $I_{h_{nk} \rightarrow f_{tn}}(h_{nk}) \sim \mathcal{CN}(h_{nk}; \tau_{h_{nk} \rightarrow f_{tn}}, v_{h_{nk} \rightarrow f_{tn}})$, we have that,

$$\begin{aligned} I_{f_{tn} \rightarrow h_{nk}}(h_{nk}) &= \frac{\hat{b}(h_{nk})}{I_{h_{nk} \rightarrow f_{tn}}(h_{nk})} \\ &\propto \mathcal{CN}(h_{nk}; \tau_{f_{tn} \rightarrow h_{nk}}, v_{f_{tn} \rightarrow h_{nk}}), \end{aligned} \quad (40)$$

where

$$\tau_{f_{tn} \rightarrow h_{nk}} = v_{f_{tn} \rightarrow h_{nk}} \left(\frac{\hat{\tau}_{h_{nk}}}{\hat{v}_{h_{nk}}} - \frac{\tau_{h_{nk} \rightarrow f_{tn}}}{v_{h_{nk} \rightarrow f_{tn}}} \right), \quad (41)$$

$$v_{f_{tn} \rightarrow h_{nk}} = \left(\frac{1}{\hat{v}_{h_{nk}}} - \frac{1}{v_{h_{nk} \rightarrow f_{tn}}} \right)^{-1}. \quad (42)$$

As $I_{f_{tn} \rightarrow h_{nk}}(h_{nk})$ follows Gaussian distribution, the message passed from variable node h_{nk} to function node q_{nk} can be calculated according to the BP rule,

$$\begin{aligned} I_{h_{nk} \rightarrow q_{nk}}(h_{nk}) &= \prod_t I_{f_{tn} \rightarrow h_{nk}}(h_{nk}) \\ &\propto \mathcal{CN}(h_{nk}; \tau_{h_{nk} \rightarrow q_{nk}}, v_{h_{nk} \rightarrow q_{nk}}), \end{aligned} \quad (43)$$

where

$$\tau_{h_{nk} \rightarrow q_{nk}} = v_{h_{nk} \rightarrow q_{nk}} \sum_t \frac{\tau_{f_{tn} \rightarrow h_{nk}}}{v_{f_{tn} \rightarrow h_{nk}}}, \quad (44)$$

$$v_{h_{nk} \rightarrow q_{nk}} = \left(\sum_t \frac{1}{v_{f_{tn} \rightarrow h_{nk}}} \right)^{-1}, \quad (45)$$

$$\beta(x_{tnk}) = C^{-1} I_{\mathbf{x}_{tk} \rightarrow f_{tn}}(x_{tnk}) \mathcal{CN}(\tau_{h_{nk} \rightarrow f_{tn}} x_{tnk}; \tau_{f_{tn} \rightarrow u_{tnk}}, v_{f_{tn} \rightarrow u_{tnk}} + v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2), \quad (34)$$

$$\tilde{\tau}_{h_{nk}} = \frac{\tau_{h_{nk} \rightarrow f_{tn}} v_{f_{tn} \rightarrow u_{tnk}} + \tau_{f_{tn} \rightarrow u_{tnk}} v_{h_{nk} \rightarrow f_{tn}} x_{tnk}^*}{v_{f_{tn} \rightarrow u_{tnk}} + v_{h_{nk} \rightarrow f_{tn}} |x_{tnk}|^2}, \quad (35)$$

and the production in (43) is through all time slots that h_{nk} is maintained unchanged.

By BP-MF rule, the message passed from function node q_{nk} to variable α_{nk} is given by,

$$I_{q_{nk} \rightarrow \alpha_{nk}}(\alpha_{nk}) = \exp \{ \langle \ln q(h_{nk}|\alpha_{nk}) \rangle_{b(h_{nk})} \} \propto \text{Gama}(\alpha_{nk}; 2, |\tau_{h_{nk}}|^2 + v_{h_{nk}}), \quad (46)$$

where $q(h_{nk}|\alpha_{nk})$ is given by (8) and the belief $b(h_{nk})$ will be shown later in (50). The belief of α_{nk} is calculated as

$$b(\alpha_{nk}) = p(\alpha_{nk}|a, b) I_{q_{nk} \rightarrow \alpha_{nk}}(\alpha_{nk}) \propto \text{Gama}(\alpha_{nk}; \hat{a}, \hat{b}), \quad (47)$$

where $\hat{a} = a + 1$ and $\hat{b} = b + |\tau_{h_{nk}}|^2 + v_{h_{nk}}$.

By BP-MF rule, the message sent from function node q_{nk} to variable h_{nk} is given by

$$I_{q_{nk} \rightarrow h_{nk}}(h_{nk}) = \exp \{ \langle \ln q(h_{nk}|\alpha_{nk}) \rangle_{b(\alpha_{nk})} \} \propto \mathcal{CN}(h_{nk}; 0, \tau_{\alpha_{nk}}^{-1}), \quad (48)$$

where

$$\tau_{\alpha_{nk}} = \frac{a+1}{b + |\tau_{h_{nk}}|^2 + v_{h_{nk}}}, \quad (49)$$

is the mean value of α_{nk} .

To compute the belief of h_{nk} , with BP rule we have,

$$b(h_{nk}) = I_{h_{nk} \rightarrow q_{nk}}(h_{nk}) I_{q_{nk} \rightarrow h_{nk}}(h_{nk}) \propto \mathcal{CN}(h_{nk}; \tau_{h_{nk}}, v_{h_{nk}}), \quad (50)$$

where

$$\tau_{h_{nk}} = \frac{\tau_{h_{nk} \rightarrow q_{nk}}}{1 + \tau_{\alpha_{nk}} v_{h_{nk} \rightarrow q_{nk}}}, \quad (51)$$

$$v_{h_{nk}} = \frac{v_{h_{nk} \rightarrow q_{nk}}}{1 + \tau_{\alpha_{nk}} v_{h_{nk} \rightarrow q_{nk}}}. \quad (52)$$

Finally, the message $I_{h_{nk} \rightarrow f_{tn}}(h_{nk})$ can be calculated as,

$$I_{h_{nk} \rightarrow f_{tn}}(h_{nk}) = \frac{b(h_{nk})}{I_{f_{tn} \rightarrow h_{nk}}(h_{nk})} \propto \mathcal{CN}(h_{nk}; \tau_{h_{nk} \rightarrow f_{tn}}, v_{h_{nk} \rightarrow f_{tn}}), \quad (53)$$

where

$$\tau_{h_{nk} \rightarrow f_{tn}} = v_{h_{nk} \rightarrow f_{tn}} \left(\frac{\tau_{h_{nk}}}{v_{h_{nk}}} - \frac{\tau_{f_{tn} \rightarrow h_{nk}}}{v_{f_{tn} \rightarrow h_{nk}}} \right), \quad (54)$$

$$v_{h_{nk} \rightarrow f_{tn}} = \left(\frac{1}{v_{h_{nk}}} - \frac{1}{v_{f_{tn} \rightarrow h_{nk}}} \right)^{-1}. \quad (55)$$

In (46)-(49) we used variational message passing to update the messages of h_{nk} and α_{nk} since VB [11] has been shown to be effective in sparse signals learning. In (48), $\tau_{\alpha_{nk}}$ serves as the inverse of the channel power. When $\tau_{\alpha_{nk}}^{-1} \rightarrow 0$, the CIR $h_{nk} \rightarrow 0$ as well. Since the inactive users are equivalent to have zero CIRs, an inactive user k is judged when $\sum_{n \in V_k} \alpha_{nk}^{-1} < \delta$ and vice versa, where δ is a small enough number.

B. Complexity Analysis

The computational analysis of the proposed algorithm is discussed in this subsection. In data detection loop, the computation consumption is dominated by (17)-(22) where the complexity in calculating $\beta(x_{tnk})$, $\tau_{u_{tnk}}$ and $v_{u_{tnk}}$ is in the order of $N(M+1)d_c$ for each time slot. Similarly, the computation consumption for channel estimation loop is dominated by calculating $\beta(x_{tnk})$, $\tilde{\tau}_{h_{nk}}$ and $\tilde{v}_{h_{nk}}$ in (34)-(36) where the complexity is also in the order of $N(M+1)d_c$. From the above discussion, the proposed algorithm has a complexity that grows linearly with the number of collision users in each subcarrier d_c and the alphabet size M .

IV. SIMULATION RESULTS

In this section Monte Carlo simulation is provided for the comparison of different schemes in terms of normalized minimum mean square (NMSE) for channel estimation and bit error rate (BER). The frequency-selective block-fading channel coefficients h_{nk} are generated randomly according to a zero mean unit variance Gaussian distribution. We assume that h_{nk} remains constant in $N_c = 64$ SCMA codewords and $N_p = 9$ pilot symbols are inserted within one fading block for channel estimation. The 200% overloaded SCMA system with $M = 4$, $N = 20$, $K = 40$, $d_v = 2$ and $d_c = 4$ is considered in this paper. 9 out of 40 active users are generated randomly and the active users transmit 2048 half rate turbo coded bits in each transmission. Monte Carlo with 1000 random transmissions are simulated to get the final results.

The proposed scheme is referred as BP-GA-EP (belief propagation based Gaussian approximation with expectation propagation) and is compared with BP-MF [9] and BP-GA [12], respectively. For BP-GA-EP and BP-GA, 5 iterations with the pilot symbols are ran to get an initial estimation of h_{nk} and 10 iterations are followed for joint channel estimation and data detection. The VB can only find a local optimum solution and is sensitive to the initialization. Therefore, for BP-MF, an MMSE estimation of h_{nk} is computed first before the pilot aided iteration. With the pilot aided channel estimation, MPA with 3 iterations are computed to get an initial estimation of the mean of x_{tnk} . After that, 10 iterations joint channel estimation and data detection are implemented.

In Fig. 2, the NMSE which is defined by $E\left\{ \frac{|h_{nk} - \hat{h}_{nk}|^2}{|h_{nk}|^2} \right\}$ is compared. As can be seen from the figure, the proposed BP-GA-EP performs best among the three methods. It has 3 dB gain compared with BP-GA at 10^{-2} and more than 2 dB gain compared with BP-MF at 10^{-3} . Further, the gap becomes larger as the SNR increases. In Fig. 3, BER performance is evaluated. From the figure, we can observe that the BP-GA has a poor performance due to the limited number of collision

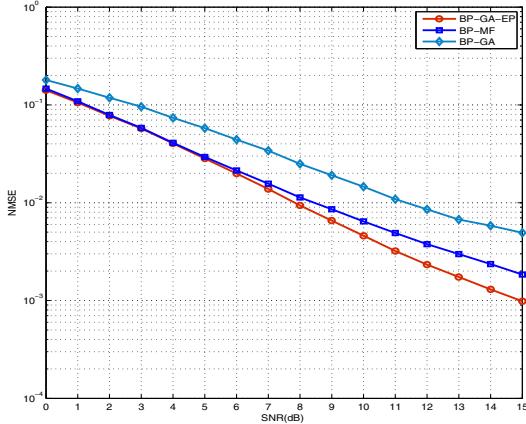


Fig. 2: Normalized MSE comparison

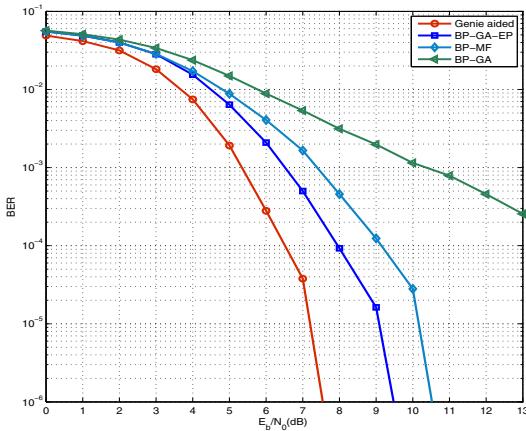


Fig. 3: Coded BER comparison

users in each subcarrier for SCMA. At this time, the central-limit theorem becomes less effective. Meanwhile, compared with BP-MF the BP-GA-EP has about 1 dB gain since it explores not only the mean value but also the covariance in computing (30), the probability of the data symbols. The genie aided case when the receiver has the perfect CSI and knows exactly the users' activity is also plotted and the performance degradation is about 2 dB due to the imperfect CSI estimation.

V. CONCLUSION

In this paper, a joint channel estimation and data detection receiver for uplink grant-free SCMA is developed. In estimating of the CSI and user activity, not only the pilot symbols but also the SCMA codewords are explored. However, the estimation of multi-variables makes the direct use of BP infeasible due to the high complexity. The approximate inference based on EP is proposed in order to reduce the computational complexity. Simulation results show that compared with the existing methods, the proposed scheme has a better performance for both channel estimation and data detection.

APPENDIX

To get the distribution for message $I_{u_{tnk} \rightarrow f_{tnk}}(u_{tnk})$ in (15), we first compute the cumulative probability distribution (CDF) which is given by

$$\begin{aligned} p(h_{nk}x_{tnk} \leq u_{tnk}) &= \langle p(h_{nk}x_{tnk} \leq u_{tnk}|x_{tnk}) \rangle_{I_{x_{tnk} \rightarrow f_{tnk}}(x_{tnk})} \\ &= I_{x_{tnk} \rightarrow f_{tnk}}(x_{tnk} = 0)\delta(u_{tnk} \geq 0) \\ &\quad + \sum_{x_{tnk} \neq 0} I_{x_{tnk} \rightarrow f_{tnk}}(x_{tnk})p(h_{nk}x_{tnk} \leq u_{tnk}|x_{tnk}), \end{aligned} \quad (56)$$

where $\delta(u_{tnk} \geq 0) = p(u_{tnk} \geq h_{nk}x_{tnk}|x_{tnk} = 0)$ is an indicator function depending whether $u_{tnk} \geq 0$ or not. Due to this indicator function, the CDF is discontinuous at $u_{tnk} = 0$. Let $F(u_{tnk}) = p(h_{nk}x_{tnk} \leq u_{tnk})$, we have,

$$\begin{aligned} p(u_{tnk} = 0) &= \lim_{u_{tnk} \rightarrow 0^+} F(u_{tnk}) - \lim_{u_{tnk} \rightarrow 0^-} F(u_{tnk}) \\ &= I_{x_{tnk} \rightarrow f_{tnk}}(x_{tnk} = 0). \end{aligned} \quad (57)$$

For $x_{tnk} \neq 0$, the distribution of u_{tnk} can be get in a similar way as in [13], Chapter 6.

ACKNOWLEDGEMENT

This paper is supported by NSFC 61671294, by Shanghai S&T Key Project 16JC1402900, by Guangxi NSFC 2015GXNSFDA139037, by National Major Project 2017ZX03001002-005, and Huawei project YBN2016100106A3.

REFERENCES

- [1] H. Nikopour and H. Baligh, "Sparse code multiple access," in Proc. IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), pp. 332-336, 2013.
- [2] K. Au, L. Zhang, etc., "Uplink contention based SCMA for 5G radio access," 2014 IEEE Globecom Workshops, pp. 900-905, 2014.
- [3] Jinfang Zhang, Lei Lu, etc., "PoC of SCMA-based uplink grant-free transmission in UCNC for 5G," IEEE J. Sel. Areas Commun., vol. 35, pp. 1353-1362, June 2017.
- [4] G. Szabo, D. Orincsay, etc., "Traffic analysis of mobile broadband networks," in Proc. WICON, June 2007, pp. 1-8.
- [5] F. Wei and W. Chen, "Low complexity iterative receiver design for sparse code multiple access," IEEE Trans. Commun., vol. 65, no. 2, pp. 621-634, Feb. 2017.
- [6] X. Meng, Y. Wu, etc., "Low complexity receiver for uplink SCMA system via expectation propagation," in Proc. IEEE Wireless Communications and Networking Conference (WCNC), Mar. 2017, San Francisco, USA.
- [7] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in Uncertainty in Artificial Intelligence, 2001: 362-369.
- [8] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," IEEE Trans. Inf. Theory, vol. 47, no. 2, pp. 498-519, Feb. 2001.
- [9] E. Riegler, G. E. Kirkelund, etc., "Merging belief propagation and the mean field approximation: a free energy approach," IEEE Trans. Inf. Theory, vol. 59, no. 1, pp. 588-602, Jan. 2013.
- [10] C. N. Manchon, G. E. Kirkelund, E. Riegler, L. P. Christensen, and B. H. Fleury, "Receiver architectures for MIMO-OFDM based on a combined VMP-SP algorithm," arXiv preprint arXiv:1111.5848, Nov. 2011.
- [11] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," IEEE Signal Process. Mag., vol. 25, no. 6, pp. 131-146, Nov. 2008.
- [12] S. Wu, L. Kuang, etc., "Message passing receiver for joint channel estimation and decoding in 3D massive MIMO-OFDM systems," IEEE Trans. Wireless Commun., vol. 15, no. 12, pp. 8122-8138, Dec. 2016.
- [13] Athanasios Papoulis, S. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th Edition. McGraw-Hill, 2002.