Wirtinger calculus-based expectation propagation in latent variable models applied to grant-free NOMA

Fakher Sagheer, Frédéric Lehmann, Antoine Berthet

Wirtinger calculus-based expectation propagation in latent variable models applied to grant-free NOMA

Fakher Sagheer, Frédéric Lehmann and Antoine O. Berthet

***Abstract*—Grant-free non-orthogonal multiple access is an emerging communication paradigm, where devices transmit to an access point without explicit permission. However, unknown user activities add discrete latent variables on top of the usual hidden variables (accounting for channels and data), thus rendering exact inference - in the form a mixture of underlying distributions growing with time - intractable. Thus low-complexity user activity detection with reliable approximate distributed inference in the form of message-passing is relevant. We propose a new generic expectation propagation solution in the form of complex Gaussian distributions, that are consistently derived using a Wirtinger calculus-based second-order approximation. The proposed algorithm is then integrated with expectation propagation for other tasks (channel estimation, symbol detection and decoding) in a fully Bayesian inference setting. The main outcome is an excellent tradeoff between user missed detections and false alarms in a non-orthognal multiple access context, thus enabling further complexity reduction by gradually disabling message-passing for successfully identified inactive users. Numerical results demonstrate the superiority of the proposed method over similar approximations.**

***Index Terms*—Grant-free NOMA, user activity detection, message-passing, expectation propagation, Wirtinger calculus.**

I. Introduction

In grant-free non-orthogonal multiple access (NOMA) [1], accurate user activity detection (UAD) is mandatory at the access point (AP) in addition to classical tasks, such as channel estimation (CE), multi-user detection (MUD) and decoding (DEC). In this letter, we introduce a unified solution to perform all aforementioned tasks without leaving the expectation propagation (EP) framework [2]. While in the signal processing literature EP [3]–[5] has been applied previously for MUD [6], DEC [7] and dynamic CE [8], our main contribution is to introduce EP for discrete latent variables suitable also for UAD, using a Wirtinger calculus-based approximation [9]. Although a similar idea appeared in [10] in a different context (massive MIMO OFDM-based systems), our derivation is both more general (no restriction to scalar observables or to an implicit assumption of i.i.d. hidden variables) and straightforward (no dependency on two additional intermediate variables besides the latent variable of interest). Also to the best of our knowledge, the limited number of references in the literature aiming at the same goal (using the EP framework) handle UAD using context-dependent hybridization of several approximate inference frameworks instead [11]–[15].

F. Sagheer and F. Lehmann are with SAMOVAR, Télécom Sud-Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France (e-mail: {fakher.sagheer,frederic.lehmann}@telecom-sudparis.eu). F. Sagheer was supported by the LabEx DigiCosme under Project ANR-11-LABEX-0045-DIGICOSME.

A. O. Berthet is with L2S, Université Paris-Saclay, CNRS, CentraleSupélec, Plateau de Moulon, 3 rue Joliot Curie 91192 Gif-sur-Yvette, France (e-mail: antoine.berthet@centralesupelec.fr).

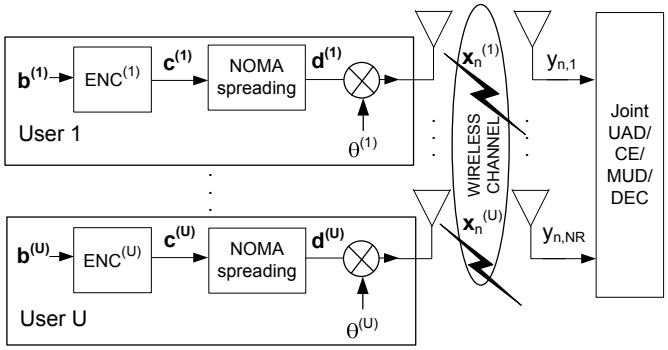

Fig. 1: Grant-free NOMA system model.

Notation. Bold letters indicate vectors and matrices while $\mathbf{0}_{m\times n}$ (resp. $\mathbf{I}_m$) is the $m \times n$ all-zero (resp. the $m \times m$ identity) matrix. $\mathcal{CN}(\mathbf{x}; \mathbf{m}, \mathbf{\Sigma})$ denotes a complex Gaussian distribution of the variable $\mathbf{x}$, with mean $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$. $\mathcal{B}(p)$ denotes a Bernoulli distribution with success probability p. The operator $\mathrm{tr}(.)$ denotes the trace of a matrix.

II. System model

We consider the generic grant-free NOMA system depicted in Fig. 1, where U denotes the maximum number of users.

A. Transceiver structure

First, let us consider the emitter for the u-th user. $ENC^{(u)}$ converts the vector of uniformly, identically and independently distributed (i.i.d.) information bits $\mathbf{b}^{(u)}$ to the vector of channel coded bits $\mathbf{c}^{(u)}$, which is in turn converted to the vector of complex Q-ary data symbols $\mathbf{d}^{(u)} = [d_1^{(u)}, d_2^{(u)}, \ldots, d_{N-1}^{(u)}]^T$ by one of the code-domain NOMA spreaders listed in [1]. Independent binary random variables $\theta^{(u)} \sim \mathcal{B}(p_a^{(u)})$ for $u = 1, \ldots, U$ model the users' activity. If the u-th user is active, $\mathbf{d}^{(u)}$ is transmitted over N consecutive resource elements (REs) via a single antenna to a wireless access point (AP) equipped with N_R receive antennas.

B. Dynamic channel model

The wireless channels are independent for all users and consider dynamic Rayleigh fading over N consecutive REs along with inter-antenna correlation [16]. A tractable model for the complex baseband equivalent length-N_R vector of channel coefficients between the u-th user and the AP over the n-th RE, $\mathbf{x}_n^{(u)}$, is obtained from the random walk model

$$\mathbf{x}_n^{(u)} = \mathbf{x}_{n-1}^{(u)} + \mathbf{\Delta}_n^{(u)}, \tag{1}$$

with i.i.d. process noise $\mathbf{\Delta}_n^{(u)} \sim \mathcal{CN}(\mathbf{\Delta}_n^{(u)}; \mathbf{0}_{N_R\times 1}, \zeta E_s^{(u)}\sigma^2\mathbf{\Gamma}^{(u)})$, where $E_s^{(u)}$, σ^2 and $\mathbf{\Gamma}^{(u)} = [\rho^{(u)^{|i-j|}}]_{1\leq i,j\leq N_R}$ account for the receive energy, correlation coefficient between successive REs and inter-antenna correlation matrix, respectively. ζ is an additional tuning parameter controlling the modeling error.

C. Observation model

The complex basedband equivalent signal at the AP $\mathbf{y}_n = [y_{n,1}, \ldots, y_{n,N_R}]^T$ over the n-th RE correponds to the superposition

$$\mathbf{y}_n = \sum_{u=1}^{U} \theta^{(u)} d_n^{(u)} \mathbf{x}_n^{(u)} + \mathbf{w}_n, \tag{2}$$

where $\mathbf{w}_n \sim \mathcal{CN}(\mathbf{0}_{N_R\times 1}, \mathbf{R})$ is a i.i.d. Gaussian noise with covariance matrix $\mathbf{R} = N_0\mathbf{I}_{N_R}$.

D. Factor graph representation

We use the factor graph formalism [7] to obtain a graphical model serving as the support of approximate distributed Bayesian inference in the form of a message-passing algorithm. Let $\mathbf{X}^{(u)} = [\mathbf{x}_0^{(u)}, \ldots, \mathbf{x}_{N-1}^{(u)}]$ denote the Gauss-Markov random process modeling the u-th user channel. Using the conditional independence hypotheses introduced in Sec. II-A-II-C, the joint posterior distribution of all hidden variables has the form

$$\begin{aligned} &p\left(\{\theta^{(u)}, \mathbf{b}^{(u)}, \mathbf{c}^{(u)}, \mathbf{d}^{(u)}, \mathbf{X}^{(u)}\}_{u=1}^{U} | \{\mathbf{y}_n\}_{n=0}^{N-1}\right) \\ &\propto \prod_{n=0}^{N-1} p\left(\mathbf{y}_n | \{\theta^{(u)}\}_{u=1}^{U}, \{d_n^{(u)}\}_{u=1}^{U}, \{\mathbf{x}_n^{(u)}\}_{u=1}^{U}\right) \\ &\prod_{u=1}^{U}\left\{ P(\theta^{(u)}) p(\mathbf{c}^{(u)}|\mathbf{b}^{(u)}) p(\mathbf{d}^{(u)}|\mathbf{c}^{(u)}) p(\mathbf{x}_0^{(u)}) \prod_{n=1}^{N-1} p(\mathbf{x}_n^{(u)}|\mathbf{x}_{n-1}^{(u)}) \right\}. \end{aligned} \tag{3}$$

Introducing the shorthand notations

$$\begin{aligned} f_n^{(u)} &= p(\mathbf{x}_n^{(u)}|\mathbf{x}_{n-1}^{(u)}) \\ g_n &= p\left(\mathbf{y}_n | \{\theta^{(u)}\}_{u=1}^{U}, \{d_n^{(u)}\}_{u=1}^{U}, \{\mathbf{x}_n^{(u)}\}_{u=1}^{U}\right) \end{aligned} \tag{4}$$

leads to the portion of the factor graph associated with the u-th user depicted in Fig. 2.

III. Proposed EP-based iterative receiver

In probabilistic graphical models, EP performs approximate inference in the form of message-passing using projections over predefined families of distributions. This is particularly relevant for models with discrete latent variables, where exact inference has complexity growing exponentially over time. Therefore, a new generic EP-based message-passing in the form of Gaussian distributions is introduced in Sec. III-A, exemplified for UAD. For the sake of completeness, one round of the proposed EP receiver, as executed sequentially for each user, is described in Sec. III-B, where existing EP-based CE, MUD and DEC message-passing rules are briefly recalled.

In the sequel, we let $\mathcal{G}$ denote the family of circularly-symmetric Gaussian distributions. Accordingly, the EP message from node a to node b projected onto $\mathcal{G}$ in the factor graph will be denoted as $\mu_{a\to b}^{EP}(.) \propto \mathcal{CN}(.; m_{a\to b}, \sigma_{a\to b}^2)$ (resp. $\mu_{a\to b}^{EP}(.) \propto \mathcal{CN}(.; \mathbf{m}_{a\to b}, \mathbf{\Sigma}_{a\to b})$) when the hidden variable of interest is a scalar (resp. a vector).

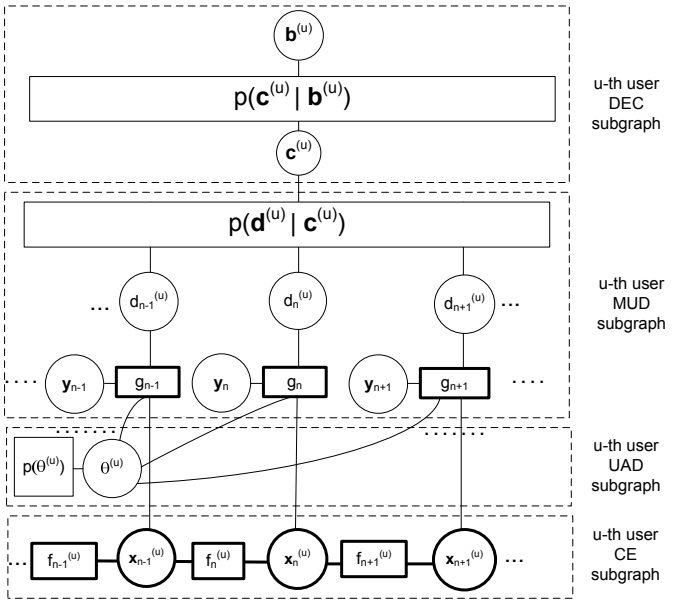

Fig. 2: Fraction of the factor graph corresponding to the u-th user in the proposed grant-free NOMA system model.

A. New Wirtinger calculus-based EP rule for UAD

The EP rule applied at the function node g_n has the form [2]

$$\mu_{g_n\to\theta^{(u)}}^{EP}(\theta^{(u)}) \propto \frac{proj_{\mathcal{G}}\left(\frac{1}{Z}\mu_{\theta^{(u)}\to g_n}^{EP}(\theta^{(u)}) l(\mathbf{y}_n|\theta^{(u)})\right)}{\mu_{\mu_{\theta^{(u)}\to g_n}^{EP}(\theta^{(u)})}}, \tag{5}$$

where $l(\mathbf{y}_n|\theta^{(u)})$ can be viewed as the likelihood function of $\theta^{(u)}$ averaged over all other hidden variables at fixed $\mathbf{y}_n$ - see (14) for a tractable Gaussian approximation. The core idea is to consider the logarithm of the argument of the projection operator in (5) as a function of $\theta^{(u)}$ alone. Then, we leverage a Wirtinger calculus-based second-order expansion [9, Eq. (99)] - around some $\theta^{(u)} = \theta_0$ to be optimized - and neglecting higher-order terms not corresponding to the desired standard circularly-symmetric Gaussian approximation leads to

$$\mu_{g_n\to\theta^{(u)}}^{EP}(\theta^{(u)}) \propto \mathcal{CN}(\theta^{(u)}; m_{g_n\to\theta^{(u)}}, \sigma_{g_n\to\theta^{(u)}}^2), \tag{6}$$

whose mean and variance are obtained as (16), after tedious algebra whose details are omitted due to lack of space. In the relatively few instances where $\sigma_{g_n\to\theta^{(u)}}^2 < 0$, the message is made uninformative (i.e. zero-mean and high-variance) [10].

The message in the reverse direction is obtained by applying the EP rule at the hidden variable $\theta^{(u)}$:

$$\mu_{\theta^{(u)}\to g_n}^{EP}(\theta^{(u)}) \propto \frac{proj_{\mathcal{G}}\left(\frac{1}{Z}P(\theta^{(u)}) \prod_{n'=0}^{N-1} \mu_{g_{n'}\to\theta^{(u)}}^{EP}(\theta^{(u)})\right)}{\mu_{g_n\to\theta^{(u)}}^{EP}(\theta^{(u)})}. \tag{7}$$

Using moment-matching, this can be rewritten as

$$\mu_{\theta^{(u)}\to g_n}^{EP}(\theta^{(u)}) \propto \mathcal{CN}(\theta^{(u)}; m_{\theta^{(u)}\to g_n}, \sigma_{\theta^{(u)}\to g_n}^2), \tag{8}$$

where

$$\sigma_{\theta^{(u)}\to g_n}^2 = m_{\theta^{(u)}\to g_n}(1 - m_{\theta^{(u)}\to g_n}). \tag{9}$$

$$m_{\theta^{(u)}\to g_n} = \frac{p_a^{(u)} \prod\limits_{\substack{n'=0:\\ n'\neq n}}^{N-1} \frac{\mu^{EP}_{g_{n'}\to\theta^{(u)}}(\theta^{(u)}=1)}{\mu^{EP}_{g_{n'}\to\theta^{(u)}}(\theta^{(u)}=0)}}{1-p_a^{(u)}+p_a^{(u)} \prod\limits_{\substack{n'=0:\\ n'\neq n}}^{N-1} \frac{\mu^{EP}_{g_{n'}\to\theta^{(u)}}(\theta^{(u)}=1)}{\mu^{EP}_{g_{n'}\to\theta^{(u)}}(\theta^{(u)}=0)}}. \tag{10}$$

The belief of $\theta^{(u)}$ being the argument of the projection operator in (7)

$$P(\theta^{(u)}|\mathbf{y}_{0:N-1}) \propto P(\theta^{(u)}) \prod_{n=0}^{N-1} \mu^{EP}_{g_n\to\theta^{(u)}}(\theta^{(u)}) \tag{11}$$

it is used to perform u-th user maximum *a posteriori* (MAP) activity hard detection as

$$\hat{\theta}^{(u)} = \begin{cases} 1 & \text{if} \quad P(\theta^{(u)}=1|\mathbf{y}_{0:N-1}) > P(\theta^{(u)}=0|\mathbf{y}_{0:N-1}) \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

B. One round of the proposed EP receiver processing user u

$\mu_{\theta^{(u)}\to g_n}(\theta^{(u)})$ is initialized to $\mathcal{NC}(\theta^{(u)};1,0)$ for $n = 0,\ldots,N-1$, since we found that enforcing user existence at the start works best. Then, at each round, EP messages over the u-th user subgraph are updated in the following order:

a) Dynamic CE subgraph: The dynamic CE subgraph in Fig. 2 is a chain, therefore EP is remniscent of smoothing in hidden Markov models and the associated messages can be found in [8], after a straightforward generalization to the vector case.

b) MUD subgraph: EP over the MUD subgraph in Fig. 2 is remniscent of soft demodulation and the associated EP messages are readily available in [6].

c) DEC subgraph: EP over the DEC subgraph in Fig. 2 using projections over the family of Bernoulli distributions boils down to ordinary binary belief propagation (BP), whose details can be found in [7] for practical codes.

d) UAD subgraph: EP over the UAD subgraph is the method introduced in Sec. III-A, where a sensible choice for θ_0 (at which second-order expansion is performed) is the mean of the belief (11) at the previous round.

Complexity orders of the all steps [8] are listed in Tab. I.

TABLE I: Per-round and per-user complexity order.

Task	CE	MUD	UAD
Proposed EP receiver	$\mathcal{O}(N_R^3)$	$\mathcal{O}(N_R^3 Q)$	$\mathcal{O}(N_R^3)$

The likelihood of $\theta^{(u)}$ has the form of a continuous Gaussian mixture

$$l(\mathbf{y}_n|\theta^{(u)}) = \int g_n \prod_{u'\neq u} \mu^{EP}_{\theta^{(u')}\to g_n}(\theta^{(u')}) \prod_{u'=1}^{U} \mu^{EP}_{d_n^{(u')}\to g_n}(d_n^{(u')}) \prod_{u'=1}^{U} \mu^{EP}_{\mathbf{x}_n^{(u')}\to g_n}(\mathbf{x}_n^{(u')}) \prod_{u'\neq u} d\theta^{(u')} \prod_{u'=1}^{U} dd_n^{(u')} \prod_{u'=1}^{U} d\mathbf{x}_n^{(u')}. \tag{13}$$

Using ordinary moment-matching [17, p. 106-108], we obtain a Gaussian approximation:

$$l(\mathbf{y}_n|\theta^{(u)}) \approx \mathcal{CN}(\mathbf{y}_n, \mathbf{h}_{\theta^{(u)}\to g_n}\theta^{(u)} + \mathbf{I}_{d_n^{(u)}\to g_n}, |\theta^{(u)}|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)}), \tag{14}$$

whose parameters are given below:

$$\begin{aligned}
\mathbf{h}_{\theta^{(u)}\to g_n} &= m_{d_n^{(u)}\to g_n}\mathbf{m}_{\mathbf{x}_n^{(u)}\to g_n},\ \mathbf{I}_{d_n^{(u)}\to g_n} = \sum_{u'\neq u} m_{\theta^{(u')}\to g_n} m_{d_n^{(u')}\to g_n}\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n} \\
\mathbf{A}_n^{(u)} &= |m_{d_n^{(u)}\to g_n}|^2\mathbf{\Sigma}_{\mathbf{x}_n^{(u)}\to g_n} + \sigma^2_{d_n^{(u)}\to g_n}(\mathbf{m}_{\mathbf{x}_n^{(u)}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u)}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u)}\to g_n}) \\
\mathbf{B}_n^{(u)} &= \sum_{u'\neq u} \sigma^2_{\theta^{(u')}\to g_n}(|m_{d_n^{(u')}\to g_n}|^2 + \sigma^2_{d_n^{(u')}\to g_n})(\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u')}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n}) \\
&\quad + \sum_{u'\neq u} |m_{\theta^{(u')}\to g_n}|^2 \left[|m_{d_n^{(u')}\to g_n}|^2\mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n} + \sigma^2_{d_n^{(u')}\to g_n}(\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u')}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n}) \right] + \mathbf{R}.
\end{aligned} \tag{15}$$

Note that the first line accounts for the expected u-th user signal conditional on $\theta^{(u)}$ and inter-user interference. The second line accounts for the uncertainty on the hidden variables of the u-th user conditional on $\theta^{(u)}$. Finally, the third and and fourth line account for the uncertainty induced by inter-user interference and noise. The mean and variance of the Wirtinger calculus-based $\mu^{EP}_{g_n\to\theta^{(u)}}(\theta^{(u)})$ approximation are given by:

$$\begin{aligned}
&\frac{1}{\sigma^2_{g_n\to\theta^{(u)}}} = \mathrm{tr}\left\{(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}\mathbf{B}_n^{(u)}\right\} + \mathbf{h}^H_{\theta^{(u)}\to g_n}(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}\mathbf{B}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1} \\
&\times \mathbf{h}_{\theta^{(u)}\to g_n} + (\mathbf{y}_n - \mathbf{I}_{d_n^{(u)}\to g_n})^H(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n - \mathbf{I}_{d_n^{(u)}\to g_n}) - 2(\mathbf{y}_n - \mathbf{I}_{d_n^{(u)}\to g_n} - \mathbf{h}_{\theta^{(u)}\to g_n}\theta_0)^H \\
&\times (|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}\mathbf{B}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n - \mathbf{I}_{d_n^{(u)}\to g_n} - \mathbf{h}_{\theta^{(u)}\to g_n}\theta_0) \\
&\frac{m_{g_n\to\theta^{(u)}}}{\sigma^2_{g_n\to\theta^{(u)}}} = \mathbf{h}^H_{\theta^{(u)}\to g_n}(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n - \mathbf{I}_{d_n^{(u)}\to g_n} - \mathbf{h}_{\theta^{(u)}\to g_n}\theta_0) + \theta_0\Bigg(\frac{1}{\sigma^2_{g_n\to\theta^{(u)}}} - \mathrm{tr}\left\{\mathbf{A}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}\right\} + \ldots \\
&(\mathbf{y}_n - \mathbf{I}_{d_n^{(u)}\to g_n} - \mathbf{h}_{\theta^{(u)}\to g_n}\theta_0)^H(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n - \mathbf{I}_{d_n^{(u)}\to g_n} - \mathbf{h}_{\theta^{(u)}\to g_n}\theta_0)\Bigg).
\end{aligned} \tag{16}$$

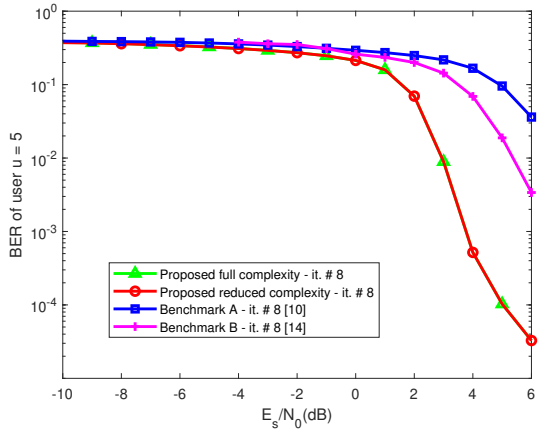

Fig. 3: BER at convergence: 12 (out of $U = 16$) active equal energy users with antenna correlation $\rho = 0.6$.

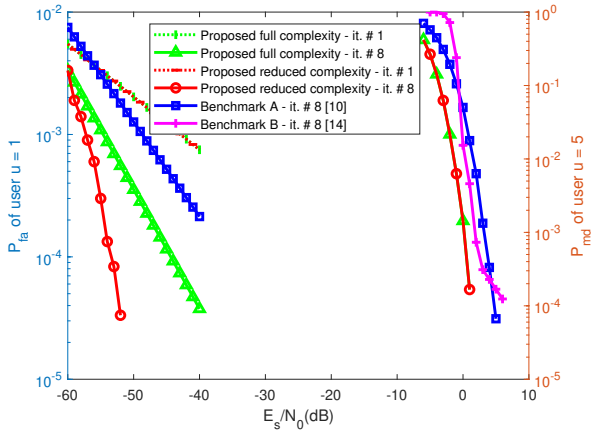

Fig. 4: P_{fa} and P_{md} at convergence: 12 (out of $U = 16$) active equal energy users with antenna correlation $\rho = 0.6$.

TABLE II: Average number of tentative users processed by the proposed reduced-complexity (RC) algorithm in Fig. 3-4.

	E_s/N_0 (dB)			
iteration #	-10	-5	0	5
1	16	16	16	16
2	5	9.8	12	12
8	4.5	9.4	12	12

IV. Numerical results

We consider a grant-free NOMA system where users employ OFDM-IDMA-based modulation and spreading over REs in the frequency domain [1]. Channel encoding consists of the serial concatenation of a rate-$1/2$ recursive systematic convolutional encoder with generators $(5/7)_8$, a rate-$1/4$ repetition encoder and a user-specific pseudo-random interleaver. 16-QAM modulation with Gray labeling converts vectors of coded bits to vectors of complex symbols sent over $N = 1024$ consecutive OFDM subcarriers (after inserting known scattered orthogonal pilot sequences with spacing of 24 subcarriers). In our simulations, 12 equal-energy users (i.e. $E_s = E_s^{(u)}$ for $u = 1, \ldots, U$) out of $U = 16$ are active (here, $\theta^{(u)} = 0$ for $u = 1, \ldots, 4$ and $\theta^{(u)} = 1$ for $u = 5, \ldots, 16$). Assuming long-enough cyclic prefixes to compensate the combined effect of channel impulse response lengths and user asynchronism, sub-carrier orthogonality is preserved at the AP so that $\mathbf{y}_n$ (resp. $\mathbf{x}_n^{(u)}$) in the observation model (2) corresponds to the received signal (resp. the channel frequency response vector for the u-th user) over the n-th subcarrier. For each OFDM block, i.i.d. complex Gaussian SIMO channel impulse responses with $N_R = 4$ are drawn with power profile decreasing exponentially with a decay constant of three taps (thus determining the frequency correlation coefficient $\sigma^2 = 5.76 \times 10^{-4}$) and spatial correlation coefficient set to $\rho = \rho^{(u)} = 0.6$ for $u = 1, \ldots, U$. At the receiver side, the channel modeling error parameter is tuned to $\zeta = 15$ and in the absence of prior knowledge of user activity, we set $p_a^{(u)} = 1/2$ for $u = 1, \ldots, U$. We compare the performance of the following EP-based receivers:

1) **Proposed**: the proposed iterative receiver in Sec. III
2) **Proposed with reduced-complexity (RC)**: a modified version of the proposed receiver in Sec. III, that freezes all messages on the u-th user subgraph as soon as $\hat{\theta}^{(u)} = 0$. The rationale behind this algorithm is that since the proposed receiver reliably detects inactive users (as we shall see), disregarding them during subsequent iterations is advantageous since the complexity is proportional to the number of processed users.
3) **Benchmark A**: for the sake of fair comparison, we modify our receiver in Sec. III using for UAD the competing Wirtinger calculus-based EP rule derived in [10]. Note that the main difference when evaluating (5) is that the Wirtinger calculus approximation in [10] is restricted to scalar observations and hidden variables, which is tatamount to ignoring antenna correlation in our setting.
4) **Benchmark B**: we also consider a recent competing hybrid EP/BP method for grant-free access [14] based on scalar auxiliary variables.

Fig. 3 (resp. Fig. 4) compares the bit-error rate (BER) (resp. the false alarm probability, P_{fa} and miss detection probability, P_{md}) of all receivers after convergence, i.e. at iteration 8. At $E_s/N_0 = 6$ (dB), the proposed algorithm outperforms benchmark B (resp. A) by 2 (resp. 3) orders of magnitude in terms of BER. In terms of P_{md}, the advantage over the benchmarks is at least 1.5 dB over the entire E_s/N_0 range. While benchmark B has negligible P_{fa} over the entire E_s/N_0 range, it also exhibits an undesirable error floor on the P_{md} at $E_s/N_0 \geq 3$ dB. We attribute the superiority of the proposed method to its ability to explicitly take receive antenna correlation into account, while Benchmark A and B ignore it. Also as announced earlier, the validity of the proposed RC mechanism is justified by the vanishing P_{fa} at negative E_s/N_0 (dB) observed as early as at iteration 1 for the proposed receiver. This is further confirmed in Tab. II, where the average number of tentative users processed in the proposed RC setting quickly converges to the correct number of active users after iteration 1 for positive E_s/N_0 (dB). Fig. 3-4 also reveal that the complexity advantage of the proposed RC method vs. the full-complexity version induces no BER or P_{md} penalty.

REFERENCES

[1] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam and S. J. Johnson, "Grant-Free Non-Orthogonal Multiple Access for IoT: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1805-1838, Third Quarter 2020.

[2] T. Minka, "Divergence Measures and Message Passing," Microsoft Resaerch Cambridge, MSR-TR-2005-173, pp. 1-17, January 2005.

[3] K. Takeuchi, "Rigorous Dynamics of Expectation-Propagation-Based Signal Recovery from Unitarily Invariant Measurements", *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 368-386, Jan. 2020.

[4] X. Meng, S. Wu, L. Kuang, J. Lu, "Concise Derivation of Complex Bayesian Approximate Message Passing via Expectation Propagation," ArXiv abs/1509.08658 (2015).

[5] P. Sun, C. Zhang, Z. Wang, C. N. Manchon and B. H. Fleury, "Iterative Receiver Design for ISI Channels Using Combined Belief- and Expectation-Propagation", *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1733-1737, Oct. 2015.

[6] X. Meng, S. Wu, L. Kuang, Z. Ni and J. Lu, "Expectation Propagation Based Iterative Multi-User Detection for MIMO-IDMA Systems," 2014 IEEE 79th Vehicular Technology Conference (VTC Spring), 2014, pp. 1-5.

[7] F. R. Kschischang, B. J. Frey and H.-.A Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498-519, Feb 2001.

[8] F. Sagheer, F. Lehmann and A. O. Berthet, "Low-Complexity Dynamic Channel Estimation in Multi-Antenna Grant-Free NOMA," 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), Helsinki, Finland, 2022, pp. 1-7.

[9] K. Kreutz-Delgado, "The complex gradient operator and the CR-calculus." , 2009 [Online] Available: https://arxiv.org/abs/0906.4835.

[10] S. Wu, L. Kuang, Z. Ni, D. Huang, Q. Guo and J. Lu, "Message-Passing Receiver for Joint Channel Estimation and Decoding in 3D Massive MIMO-OFDM Systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8122-8138, Dec. 2016.

[11] F. Wei, W. Chen, Y. Wu, J. Ma and T. A. Tsiftsis, "Message-Passing Receiver Design for Joint Channel Estimation and Data Decoding in Uplink Grant-Free SCMA Systems," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 167-181, Jan. 2019.

[12] F. Lehmann, "Joint User Activity Detection, Channel Estimation, and Decoding for Multiuser/Multiantenna OFDM Systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8263-8275, Sept. 2018.

[13] W. Yuan, N. Wu, Q. Guo, D. W. K. Ng, J. Yuan and L. Hanzo, "Iterative Joint Channel Estimation, User Activity Tracking, and Data Detection for FTN-NOMA Systems Supporting Random Access," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 2963-2977, May 2020.

[14] Y. Zhang, Z. Yuan, Q. Guo, Z. Wang, J. Xi and Y. Li, "Bayesian Receiver Design for Grant-Free NOMA With Message Passing Based Structured Signal Estimation," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8643-8656, Aug. 2020.

[15] R. B. Di Renna and R. C. de Lamare, "Joint Channel Estimation, Activity Detection and Data Decoding Based on Dynamic Message-Scheduling Strategies for mMTC," *IEEE Transactions on Communications*, vol. 70, no. 4, pp. 2464-2479, April 2022.

[16] K.-H. Ngo, M. Guillaud, A. Decurninge, S. Yang and P. Schniter, "Multi-User Detection Based on Expectation Propagation for Non-Coherent SIMO Multiple Access Channel," *IEEE Transactions on Wireless Communications*, vol. 19, no. 9, pp. 6145-6161, September 2020.

[17] H. Tanizaki, *Nonlinear filters: estimation and applications*, Berlin, Germany: Springer, 1996.