Grant-Free Access via Bilinear Inference for Cell-Free MIMO With Low-Coherence Pilots

Hiroki Iimori, *Graduate Student Member, IEEE*, Takumi Takahashi, *Member, IEEE*, Koji Ishibashi, *Senior Member, IEEE*, Giuseppe Thadeu Freitas de Abreu, *Senior Member, IEEE*, and Wei Yu, *Fellow, IEEE*

***Abstract*— We propose a novel joint activity, channel and data estimation (JACDE) scheme for multiple-input multiple-output (MIMO) systems. The contribution aims to allow significant overhead reduction of MIMO systems by enabling grant-free access, while maintaining moderate throughput per user. To that end, we extend the conventional MIMO transmission framework so as to incorporate activity detection capability without resorting to spreading informative data symbols, in contrast with related work which typically relies on signal spreading. Our method leverages a Bayesian message passing scheme based on Gaussian approximation, which jointly performs active user detection (AUD), channel estimation (CE), and multi-user detection (MUD), incorporating also a well-structured low-coherence pilot design based on frame theory, which mitigates pilot contamination, and finally complemented with a detector empowered by bilinear message passing. The efficacy of the resulting JACDE-based grant-free access scheme in the cell-free MIMO system setup compliant with fifth generation (5G) new radio (NR) orthogonal frequency-division multiplexing (OFDM) signaling is demonstrated by simulation results. The results are shown to outperform the current state-of-the-art and approach the performance of an idealized (genie-aided) scheme in which user activity and channel coefficients are perfectly known.**

***Index Terms*— Grant-free access, bayesian inference, multiple-input multiple-output (MIMO), bilinear approximate message passing, frame theory.**

I. Introduction

MULTIPLE-ANTENNA architectures, in particular massive multiple-input multiple-output (MIMO) and its extensions, will continue to be one of essential technologies in fifth generation (5G) and future sixth generation (6G) networks, in order to satisfy the ever-growing demand for higher data rates and user capacities as well as the heterogeneous requirements raised by massive machine type communications (mMTC), enhanced mobile broadband (eMBB), ultra reliable low latency communications (URLLC) and their various combinations. What makes massive MIMO technology a simultaneous enabler of high throughput communications and massive connectivity is the significant amount of the spatial degrees of freedoms (DoFs) it provides, which can be exploited to solve inherent problems such as multi-user detection (MUD), channel estimation (CE), and active user detection (AUD) in uplink scenarios, among others [1], [2].

Compared to conventional coherent MIMO communication mechanism, where AUD and CE are sequentially performed based on predetermined reference signals ($e.g.$, pilot sequences) followed by MUD relying on the estimated channel state information (CSI), a main challenge of massive uplink access is the communication overhead required for CSI acquisition, which scales with the number of potential uplink users in the system due to the need of orthogonal pilot sequences so as to maintain accurate CSI knowledge. It is also worth mentioning that utilizing non-orthogonal pilot sequences for channel estimation, while contributing to reducing the overhead, leads to severe MUD performance deterioration due to the rank-deficient ($i.e.$, underdetermined) conditions typically faced, even under the assumption that perfect AUD is available at the receiver. In addition, in massive MIMO settings, excessive piloting might exceed channel coherence time, particular in the case of fast fading environments, which makes non-orthogonal pilots necessity.

A promising emerging approach to tackle this issue is joint channel and data estimation (JCDE) which takes advantage of estimated data symbols as soft pilot sequences, while exploiting their statistical quasi-orthogonality to improve system performance and efficiency. In particular, the bilinear generalized approximate message passing (BiGAMP) scheme proposed in [3] has been considered a key ingredient to solve such a detection problem in wireless systems. In that scheme, Onsager correction is employed to decouple the self-feedback of messages across iterations as is the case with the approximate message passing (AMP), leading to stable convergence behavior as shown, for instance, in [4], [5].

It has, however, been recently shown in [6] that the estimation performance of BiGAMP severely deteriorates when non-orthogonal pilot sequences are exploited, even if adaptive damping is employed, because the derivation of BiGAMP relies heavily on the assumption of very large

Manuscript received August 28, 2020; revised April 4, 2021 and May 29, 2021; accepted June 4, 2021. Date of publication June 18, 2021; date of current version November 11, 2021. This work was supported in part by the Japan Science and Technology Agency, Strategic International Collaborative Research Program (JST SICORP), Japan, under Grant JPMJSC20C1. The work of Wei Yu was supported by the Canada Research Chairs Program. The associate editor coordinating the review of this article and approving it for publication was S. Yang. *(Corresponding author: Hiroki Iimori.)*

Hiroki Iimori and Giuseppe Thadeu Freitas de Abreu are with the Focus Area Mobility, Department of Electrical and Computer Engineering, Jacobs University Bremen, 28759 Bremen, Germany (e-mail: h.iimori@ieee.org; g.abreu@jacobs-university.de).

Takumi Takahashi is with the Graduate School of Engineering, Osaka University, Suita 565-0871, Japan (e-mail: takahashi@comm.eng.osaka-u.ac.jp).

Koji Ishibashi is with the Advanced Wireless and Communication Research Center (AWCC), The University of Electro-Communications, Tokyo 182-8585, Japan (e-mail: koji@ieee.org).

Wei Yu is with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: weiyu@comm.utoronto.ca).

Digital Object Identifier 10.1109/TWC.2021.3088125

systems, although shortening the pilot sequence is the very aim of the method itself. In order to circumvent this issue, the authors in [7] proposed a novel bilinear message passing algorithm, referred to as bilinear Gaussian belief propagation (BiGaBP), with the aim of generalizing BiGAMP on the basis of belief propagation (BP) [8] for robust recovery subject to non-orthogonal piloting. Despite the aforementioned progresses, many existing works including the ones mentioned above, focus only on joint CE and MUD while assuming that perfect AUD is available at the receiver.

One of the solutions to the AUD problem is grant-free random access [9], [10], which has been intensively investigated in the last few years and can be categorized as a variation of joint activity and channel estimation (JACE) or (if symbol detection is also integrated) of joint activity, channel and data estimation (JACDE), with Bayesian receiver design components. In the context of Bayesian approaches, Bayesian JACE can be seen as a non-orthogonal pilot-based random access protocol in which active users simultaneously transmit their unique spreading signatures to their base stations (BSs), and the BS employs a massage passing algorithm – $e.g.$ multiple measurement vector approximate message passing (MMV-AMP) – as receiver, with the aim of detecting user activity patterns and their corresponding channel responses, while taking advantage of the time-sparsity resulting from the activity patterns. As for Bayesian JACDE schemes, most of the existing works on that approach, $e.g.$ [11]–[15], are extensions of the aforementioned Bayesian JACE in which spreading data sequences generated by multiplying data symbols with their unique spreading signatures are transmitted (please refer to the system model given in $e.g.$ [11]–[15] for more technical details), while leveraging a similar receiver design as that of Bayesian JACE methods.

There is also another approach to AUD that takes advantage of the sample covariance matrix constructed from the large number of antennas at the receiver. This covariance-based method has also attracted attention due to its applicability to unsourced random access (URA), where JACDE can be achieved by letting active users transmit a codeword sequence selected from a common predetermined codebook over a given time slot. To elaborate, it has been shown in [16] that the covariance-based approach is able to accommodate a larger number of active users, while using limited per-user wireless resources[1] due to the nature of index-type modulation based on spread codewords, which is therefore suited to super low-rate mMTC scenarios.

Compounding on the above, a fundamental challenge of grant-free approaches from a system viewpoint is the spatial correlation of the massive MIMO channel, which has been argued in [18] to be a limiting factor of centralized massive MIMO, although most of the existing work in the area, including $e.g.$ [6], [7], [11], [12], [19]–[21], make use of the assumption that channels are subjected to ideal (uncorrelated) Rayleigh fading. In order to iron out this issue, the cell-free massive MIMO concept – studied in [22], [23], which virtually configures a massive MIMO setup by spatially distributing access points (APs) connected through wired fronthaul links to a common central computing unit (CCU) – has recently emerged, offering an architecture capable of decorrelating the spatial dependence between APs, leading to an ideal independently distributed channel structure.

Within the context outlined above, this article aims to tackle the challenging task of non-coherent JACDE in a cell-free MIMO architecture. In order to achieve this objective without sacrificing spectral efficiency, we seek in particular a solution which, unlike most of existing works [11]–[13], does not require spreading the data symbols. To that end, a key component of the proposed approach is an appropriately designed bilinear message passing algorithm resulting in a novel grant-free JACDE scheme for non-orthogonal random access in massive MIMO systems with *non*-spread data streams, which, to the best of our knowledge, has not been presented.

A. Related Work

As described above, there is a variety of approaches to solve MUD, CE, and AUD in uplink MIMO systems, which for the sake of readability are categorized into two distinct approaches. First, there are grant-based approaches in which active users are assumed known ($i.e.$ having been granted access to the system), such that only MUD and CE are jointly carried out, including convex relaxation methods [24], non-convex successive over-relaxation methods [25], variational-Bayes methods [26], and AMP methods, among others. In [3], [4], a unified AMP-based approach to matrix completion, robust principle component analysis (PRCA), and dictionary learning was proposed, and the resulting BiGAMP was empirically shown in [4], [5] to be competitive in terms of phase transition and computation complexity. In the context of self-calibration and matrix compressed sensing, parametric BiGAMP (PBiGAMP) was proposed in [27] and found to yield improved phase transitions in comparison with the aforementioned counterparts.

In the attempt to overcome AMP's vulnerability against measurement correlation [28], [29], the vector AMP (VAMP) [30], orthogonal AMP (OAMP) [31] and other iterative detectors based on the expectation propagation (EP) framework [32], [33] have been proposed, which handle a class of unitary-invariant measurement matrices. These algorithms require, however, matrix inversion operations unlike the original AMP approach. A rigorous analysis of the convergence property of this approach was presented in [30], [33], and potential connection among different methods was investigated in [34]–[36], with the extension to the bilinear inference method proposed in [37], [38]. Also, it is worth-noting that the Gaussian belief propagation (GaBP) [8] approach can be interpreted as the origin of the aforementioned AMP-based message passing rules.

[1]In [16], [17], for instance, 96 bits per user are sent by exploiting 3200 symbol lengths, which correspond to approximately $11 \sim 22$ orthogonal frequency-division multiplexing (OFDM) frames in 5G new radio (NR) setups with a sub-GHz carrier frequency. In other words, as a 5G NR OFDM frame is designed to amount to 10 [ms], 96 bits are delivered with hundreds of milliseconds in this setup.

In turn, in the second approach known as grant-free massive random access, MUD, CE and AUD are all performed jointly. Aligned with this approach is the work in [39], where it was shown theoretically that non-orthogonal access schemes with random coding not only outperform classical multiple-access channel (MAC) protocols such as coded ALOHA, but also have the potential to nearly achieve theoretical limits in terms of user capacity. Motivated by the above, various authors [16], [17], [40]–[43] studied random coding schemes and/or its covariance-based receiver designs for URA in grant-free MIMO systems with massive numbers of potential users. In this line of work, the massive MIMO JACE problem was considered in [44], [45] with a similar receiver design based on the sample covariance approach. Finally, as for the Bayesian approach, the authors in [20], [46]–[48] have investigated JACE for massive random access with Bayesian compressed sensing receivers, which was also extended to JACDE in [11]–[13], [49], where informative bits are embedded into spreading codewords. And since many existing works, including the ones mentioned above, considered the conventional centralized massive MIMO architecture, which in practice might suffer from spatial correlation, a more recent contribution [50] has tackled this issue by studying a structured massive access scheme for JCDE in a cell-free massive MIMO setting, while assuming perfect activity detection and a grant-based architecture.

In summary, it can be said that the majority of contributions addressing JACDE in MIMO uplink channels can be categorized as either covariance-based receivers for the grant-free URA, or Bayesian receivers with informative data symbols embedded into spreading sequences.

B. Contributions

In light of all the above, the contributions of the article can be summarized as follows.

- **Feasibility of non-spread JACDE**: We demonstrate the feasibility of grant-free JACDE *without spreading data sequences* drawn from a predetermined constellation as is the case for the conventional coherent MIMO systems. This is in contrast to most of related literature addressing AUD, CE, and MUD jointly in a grant-free fashion [11]–[14], in which the spreading of data sequences by pilot sequences is required, sacrificing spectral efficiency, limiting data rate per user, and increasing sensitivity to fading.
- **Cell-free assisted grant-free access**: We extend previous works such as [19], [50], [51] such that the proposed algorithm is directly applicable to a *cell-free architecture*, without sacrificing suitability also to centralized MIMO systems. We remark that a potential advantage of the cell-free architecture is that it helps resolve spatial correlation problems of massive MIMO. To the best of our knowledge, no grant-free design for cell-free massive MIMO scheme without pilot-based data spreading has been proposed yet.
- **Suitability of short pilot sequence**: We employ a signal model incorporating a frame-theoretic non-orthogonal pilot design, whose mutual coherence approach Welch's theoretical coherence lower bound even for relatively *short pilot sequences*. This is in contrast to much of existing literature, which relies on Gaussian pilot designs, exhibiting poor mutual coherence properties, except in the asymptotic (large-system) regime.
- **Bilinear inference for JACDE**: In order to enable the above, a *novel JACDE algorithm, dubbed activity-aware BiGaBP*, is presented here, in which bilinear inference, message passing rules, Gaussian approximation, and a new belief scaling technique that forges resilience of the derived messages, are combined.

Simulation results adhering to 5G NR configurations are offered to support the claims above, demonstrating the concrete applicability of the proposed method in real life systems. We emphasize that, to the best of our knowledge, a JACDE mechanism for cell-free grant-free MIMO systems without spreading data symbols, which is the key contribution of this article, has not yet been proposed.

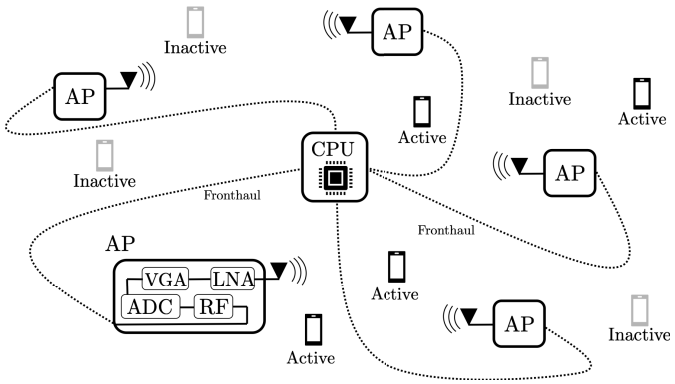
C. Notation

In the remainder of the article, the following notation will be employed. Sets of numbers in the real, complex, and Hilbert spaces will be denoted by $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{H}$, respectively. The transpose, transpose conjugate, and Hermitian adjoint operators, correspondingly in the real, complex, and Hilbert spaces will be respectively expressed as $\cdot^{\mathrm{T}}$, $\cdot^{\mathrm{H}}$, and $\cdot^{\mathrm{H}}$. The N-dimensional multivariate circular symmetric complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ will be denoted by $\mathcal{CN}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, while $\mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ will denote its real-domain counterpart. $\|\cdot\|_p$ will be used to denote the p-norm with $p \in \{2, \infty\}$, while $\langle \cdot, \cdot \rangle$ is the inner product operator. Finally, $\mathrm{s.t.}$, $\max$, and $\min$ stands for "subject to", "maximize", and "minimize", respectively, which are utilized to formulate optimization problems, while $\propto$ denotes "proportional to".

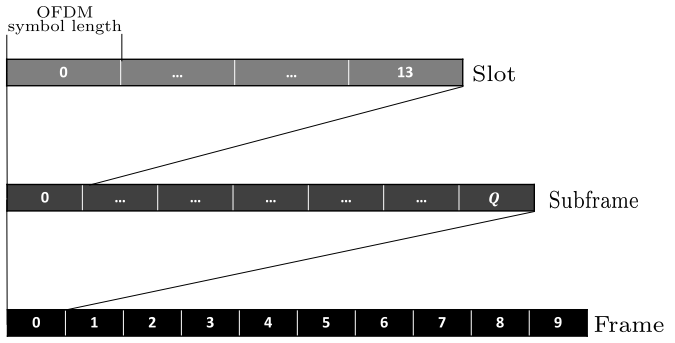
II. System Model

Consider a cell-free large MIMO system composed of N spatially distributed single-antenna APs connected by wired fronthaul links to a common high-performance CCU, serving M synchronized single-antenna users in a grant-free fashion, as depicted in Figure 1(a). Due to the dynamic nature of grant-free systems, it is assumed that a fraction of the total M single-antenna users become active within a given 5G NR OFDM symbol frame [52], whereas the rest of uplink users remain silence during that period. As shown in Figure 1(b), in the 5G NR signaling structure, each OFDM frame consists of 10 subframes and each subframe consists of Q slots. Furthermore, each slot is composed of 14 OFDM symbols. Taking advantage of this signaling design, one may utilize a part of the radio frame structure as reference signals and the rest as data streams. We also remark that the number of slots within a certain subframe ($i.e.$, Q) depends on the subcarrier spacing employed in a system.

Given the above, let K be the total number of discrete time indices within an OFDM frame and $\mathscr{C} \triangleq \{c_1, c_2, \ldots, c_{2^b}\}$



(a) A model illustration of cell-free MIMO systems with distributed single-antenna APs serving uplink users which access the system in a grant-free basis.



(b) Radio frame structure in 5G NR [52].

Fig. 1. System and Signal Model.

represent a given constellation of symbols with b denoting the number of bits per symbol. Introducing the user index set $\mathscr{M} \triangleq \{1, 2, \ldots, M\}$ and a set of active users $\mathscr{A}$, the received signal vector $\boldsymbol{y}_k \in \mathbb{C}^{N\times 1}$ at the k-th time index with $k \in \{1, 2, \ldots, K\}$ can be written as

$$\boldsymbol{y}_k = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{w}_k, \tag{1}$$

where $\boldsymbol{x}_k \in \mathbb{C}^{M\times 1}$ denotes a possibly sparse transmitted signal vector, such that only its elements of indices $a \in \mathscr{A}$ are non-zero; $\boldsymbol{w}_k \in \mathbb{C}^{N\times 1}$ is a circularly symmetric zero-mean i.i.d. additive white Gaussian noise (AWGN), *i.e.*, $\boldsymbol{w}_k \sim \mathcal{CN}_N(\boldsymbol{0}, N_0\mathbf{I}_N)$; and $\boldsymbol{H} \in \mathbb{C}^{N\times M}$ denotes a flat block fading communication channel matrix assumed to be constant during K successive transmissions.

Assuming that the user activity remains consistent within an OFDM frame, we can concatenate the K consecutive symbol transmissions into the transmitted signal matrix $\boldsymbol{X} \triangleq [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K]$. In turn, the row sparse nature of $\boldsymbol{X}$ can be translated to a column-sparsity in the channel matrix $\boldsymbol{H}$, such that the m-th column $\boldsymbol{h}_m$ of the channel matrix can be modeled as a multivariate Bernoulli-Gaussian random variable, that is

$$\boldsymbol{h}_m \sim (1-\lambda)\delta(\boldsymbol{h}_m) + \lambda\mathcal{CN}_N(\boldsymbol{0}, \boldsymbol{\Gamma}_m), \tag{2}$$

where λ is the activity factor, $\delta(\boldsymbol{h}_m)$ denotes the Dirac delta function that takes the value 0 everywhere except at $\boldsymbol{h}_m = \boldsymbol{0}$ where it takes the value 1, and $\boldsymbol{\Gamma}_m$ is the covariance matrix of the channel.

In light of the above, the received signal matrix $\boldsymbol{Y} \in \mathbb{C}^{N\times K}$ concatenating the received signal vectors given in equation (1) over K successive time indices, can be readily expressed as

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{W}, \tag{3}$$

where $\boldsymbol{Y} \triangleq [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K]$ and $\boldsymbol{W} \triangleq [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K]$.

Let us employ the subscripts $\cdot_p$ and $\cdot_d$ to indicate pilot and data signals, respectively, and in order to explicitly express the pilot and data sequences within the transmit and received signal, let us define, without loss of generality,

$$\boldsymbol{Y} \triangleq [\boldsymbol{Y}_p, \boldsymbol{Y}_d] \quad \text{and} \quad \boldsymbol{X} \triangleq [\boldsymbol{X}_p, \boldsymbol{X}_d], \tag{4}$$

where $\boldsymbol{Y}_p \in \mathbb{C}^{N\times K_p}, \boldsymbol{Y}_d \in \mathbb{C}^{N\times K_d}, \boldsymbol{X}_p \in \mathbb{C}^{M\times K_p}$ and $\boldsymbol{X}_d \in \mathbb{C}^{M\times K_d}$, with $K_p + K_d = K$ and $K_p \ll K_d < K$; and where each element of $\boldsymbol{X}_d$ is assumed to be drawn from the constellation set $\mathscr{C}$ in a similar way to conventional coherent MIMO systems.

At this point, we stress that although cell-free systems are our primary interest in this article, the signal model given in equation (3), and consequently the proposed JACDE method presented later, are not limited to cell-free architectures. For instance, a conventional MIMO system model follows directly by setting $\boldsymbol{\Gamma}_m \triangleq \operatorname{diag}([\gamma_{1m}, \ldots, \gamma_{Nm}])$, with $\gamma_m = \gamma_{1m} = \ldots = \gamma_{Nm}$, where γ_m denotes the power of the channel from the m-th user. In contrast, in order to model a cell-free MIMO system the receive channel powers γ_{nm} are simply allowed to be different for each link between the n-th AP and m-th user.

We remark that unlike most grant-free systems found in literature, *e.g.* [11]–[13], [47], [53], [54], where data symbols are transmitted after spreading by pilot sequences, our approach can be seen, in light of equation (3) as a cell-free-implementable activity-aware variant of conventional MIMO-OFDM systems, having the potential to enable both grant-free random access and pilot length reduction by jointly performing user activity, channel state, and data detection. These features make our approach better suited to meet the demand of higher spectrum efficiency per user than grant-free methods previously proposed, including those aforementioned.

A. Pilot Sequence Design

In this subsection we review a frame-theoretic approach to effectively design a structured pilot matrix for non-orthogonal transmission [55]–[58], aiming at efficiently reducing the pilot length K_p while preserving the linear independence between vectors in the pilot matrix $\boldsymbol{X}_p$ as much as possible. To this end, assuming a non-orthogonal representation of the pilot matrix (*i.e.*, $K_p < M$), let us first define the mutual coherence as a measure of the similarity between non-orthogonal signals.

Definition 1 (Mutual Coherence): Let $\boldsymbol{F} \triangleq [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_L] \in \mathbb{H}^{J\times L}$ be a *frame* matrix over a Hilbert space $\mathbb{H}^{J\times L}$, comprising of frame vectors $\boldsymbol{f}_\ell \in \mathbb{H}^{J\times 1}$, with $\ell \in \{1, \ldots, L\}$ and $J < L$. The mutual coherence of $\boldsymbol{F}$ is given by

$$\mu(\boldsymbol{F}) \triangleq \max_{\ell\neq\ell'} \frac{|\langle \boldsymbol{f}_\ell, \boldsymbol{f}_{\ell'}\rangle|}{\|\boldsymbol{f}_\ell\|_2 \|\boldsymbol{f}_{\ell'}\|_2}, \quad \forall\{\ell, \ell'\} \in \{1, 2\ldots, L\}, \tag{5}$$

which in the case of an equal-norm frame, reduces to

$$\mu(\boldsymbol{F}) \triangleq \max_{\ell\neq\ell'} |\langle \boldsymbol{f}_\ell, \boldsymbol{f}_{\ell'}\rangle|, \quad \forall\{\ell, \ell'\} \in \{1, 2\ldots, L\}. \tag{6}$$

One readily notices from the above that the mutual coherence of a frame matrix $\boldsymbol{F}$ is equivalent to the maximum absolute value of the non-diagonal elements of the corresponding gram matrix $\boldsymbol{G} \triangleq \boldsymbol{F}^{\mathrm{H}}\boldsymbol{F}$, which for $L \leq J^2$ is known to be lower-bounded by the Welch bound[2] [59]

$$\mu(\boldsymbol{F}) \geq \sqrt{\frac{L-J}{J(L-1)}}. \tag{7}$$

Besides low mutual coherence, for the sake of fairness – not in terms of throughputs but in terms of resource utilization [58] – that pilot sequences employed by different users have the same energy, which in the context of frame designs translates to the following desired property.

Definition 2 (Tightness): By means of the Rayleigh-Ritz Theorem, a frame matrix $\boldsymbol{F}$ possesses the following inequalities.

$$\alpha\|\boldsymbol{a}\|_2^2 \leq \|\boldsymbol{F}^{\mathrm{H}}\boldsymbol{a}\|_2^2 \leq \beta\|\boldsymbol{a}\|_2^2, \quad \forall \boldsymbol{a} \in \mathbb{H}^J, \tag{8}$$

where $0 < \alpha \leq \beta < \infty$ and $\boldsymbol{F}$ is called *tight* if and only if (iff) $\alpha = \beta$.

In light of the above, our goal is to design low-coherence tight frames as pilot sequences, which can be utilized even under severe non-orthogonal scenarios ($i.e.$, $K_p \ll M$). It has been recently shown in [61] that a group-theoretic tight frame construction approach achieves near-Welch-bound performance, but unfortunately such cyclic-group approach is applicable only to particular cases in which the number of frame vectors ($i.e.$, L) is a prime number, while in the context hereby frames with arbitrary L and J are required.

We therefore consider instead a convex optimization based construction method recently proposed in [55], [56] and further developed in [57], [58]. Such a low-coherence unit-norm tight frame with arbitrary dimensions can be obtained by taking advantage of an iterative method, referred to as sequential iterative decorrelation via convex optimization (SIDCO) [55], whose extension to the complex space – dubbed as complex SIDCO (CSIDCO) [56] – has been studied, where the strategy to minimize the mutual coherence is to solve *iteratively* ($i.e.$, for different τ) the following convex optimization problem:

$$\min_{\substack{\boldsymbol{f}_\ell \in \mathbb{C}^{J\times 1} \\ \forall \ell \in \{1,2,\ldots,L\}}} \quad \|\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}}\boldsymbol{f}_\ell\|_\infty \tag{9a}$$

$$\text{s.t. } \|\boldsymbol{f}_\ell - \tilde{\boldsymbol{f}}_\ell\|_2^2 \leq T_\ell^{(\tau)}, \tag{9b}$$

where $\tilde{\boldsymbol{F}}_\ell \in \mathbb{C}^{J\times L-1}$ is obtained by prunning the ℓ-th column of $\tilde{\boldsymbol{F}}$, τ indicates the iteration index, and the search region is limited to a multidimensional Euclidean ball of radius

$$T_\ell^{(\tau)} = 1 - \max_{\ell' | \ell' \neq \ell} \frac{|\langle \boldsymbol{f}_\ell^{(\tau-1)}, \boldsymbol{f}_{\ell'}^{(\tau-1)} \rangle|^2}{\|\boldsymbol{f}_\ell^{(\tau-1)}\|_2^2 \|\boldsymbol{f}_{\ell'}^{(\tau-1)}\|_2^2}. \tag{10}$$

Although the CSIDCO reformulation given in equation (9) already follows the disciplined convex programming conventions, such that this problem can be easily solved via interior point methods through CVX available in high-level numerical computing programming languages such as MATLAB and Python, the abstraction penalty of such high-level programming languages are too high for real-world communication systems. Aiming at real-time processing of convex optimization problems, the authors in [62] have developed an automatic low-level code generator for conic programming problems, which solves convex problems with moderate size on the order of microseconds or milliseconds, although it is limited to quadratic program (QP)-representable convex problems. In light of the above, for the sake of completeness, equation (9) was transformed in [57], [58] into the following quadratic program.

Theorem 1 (Quadratic CSIDCO): Introducing $\boldsymbol{x}_\ell \triangleq [\Re\{\boldsymbol{f}_\ell\}; \Im\{\boldsymbol{f}_\ell\}; t_{\ell,R}; t_{\ell,I}] \in \mathbb{R}^{2J+2\times 1}$ with slack variables $t_{\ell,R} \in \mathbb{R}_+$ and $t_{\ell,I} \in \mathbb{R}_+$ for all ℓ, the CSIDCO formulation given in equation (9) for a unit-norm low-coherence frame construction can be rewritten as

$$\min_{\substack{\boldsymbol{x}_\ell \\ \forall \ell \in \{1,2,\ldots,L\}}} \quad \boldsymbol{x}_\ell^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{x}_\ell \tag{11a}$$

$$\text{s.t. } \boldsymbol{A}_{\ell,R,1}\boldsymbol{x}_\ell \leq \boldsymbol{0}, \tag{11b}$$

$$\boldsymbol{A}_{\ell,R,2}\boldsymbol{x}_\ell \leq \boldsymbol{0}, \tag{11c}$$

$$\boldsymbol{A}_{\ell,I,1}\boldsymbol{x}_\ell \leq \boldsymbol{0}, \tag{11d}$$

$$\boldsymbol{A}_{\ell,I,2}\boldsymbol{x}_\ell \leq \boldsymbol{0} \tag{11e}$$

$$\boldsymbol{x}_\ell^{\mathrm{T}}\boldsymbol{\Xi}\boldsymbol{x}_\ell - 2\boldsymbol{b}_\ell^{\mathrm{T}}\boldsymbol{x}_\ell + 1 - T_\ell \leq 0, \tag{11f}$$

where $\boldsymbol{\Phi} \triangleq \left[\begin{smallmatrix} \boldsymbol{0}_{2J} & \boldsymbol{0}_{2J\times 2} \\ \boldsymbol{0}_{2\times 2J} & \boldsymbol{I}_2 \end{smallmatrix}\right]$, $\boldsymbol{\Xi} \triangleq \left[\begin{smallmatrix} \boldsymbol{I}_{2J} & \boldsymbol{0}_{2J\times 2} \\ \boldsymbol{0}_{2\times 2J} & \boldsymbol{0}_{2\times 2} \end{smallmatrix}\right]$, $\boldsymbol{b}_\ell \triangleq \left[\tilde{\boldsymbol{f}}_\ell^{\mathrm{T}} \quad 0 \quad 0\right]^{\mathrm{T}}$, the other coefficient matrices are listed in Appendix A, and the iteration index τ is omitted for brevity.

Proof: See Appendix A □

One may argue that the frame matrix constructed via the quadratic CSIDCO method described above is not strictly tight due to the fact that tightness is not enforced during the optimization process. However, the tightening approach proposed in [63] based on the polar decomposition can be applied to the output of the quadratic CSIDCO. Consequently, one can obtain an arbitrarily-sized low-coherent equal-norm tight frame sufficiently close to the ideal equiangular tight frames which are not suited to practice as they exist only for particular dimensions. Owing to this flexibility and sufficiently high-performance, this approach is therefore adopted as pilot sequences considered in this article.

It is also worth-noting that an alternative to design a low-coherence pilot matrix is to leverage Grassmanian packing techniques [60], [64]–[66]. For instance, a pilot matrix design based on a particular structure such as one proposed in [60] may bring beneficial aspects in practice due to its low-complexity nature. However, optimizing such a pilot design in terms of performance and/or complexity is beyond the scope of this article, since the focus of this section is not to propose a new pilot design but to bring to light that such alternative exists and is suitable for pilot matrix design in grant-free access schemes.

To illustrate the performance of the quadratic CSIDCO method described above especially in 5G NR setups, mutual coherence comparisons of the method against popular pilot

[2]One may consider an extremely severe scenario $L > J^2$, although this is beyond the scope of this article. In this case, the Welch bound is no longer a proper benchmark, implying that one needs another alternative that can be drawn from $e.g.$, [60].

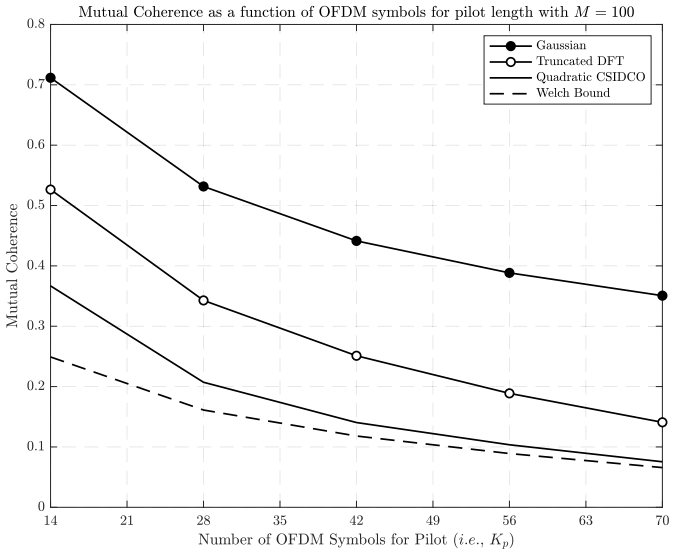


Fig. 2. Mutual coherence comparison as a function of OFDM symbols utilized for pilot lengths K_p with $M = 100$ uplink users. As benchmarks, we adopt the popular random Gaussian pilot sequence considered in, *e.g.*, [17], [20], and a randomly truncated discrete Fourier transform matrix.

construction approaches (*i.e.*, the random Gaussian and truncated discrete Fourier transform matrices) as a function of the number of OFDM symbols leveraged for pilot lengths are shown in Figure 2, which demonstrates that in fact the quadratic CSIDCO approaches the Welch bound while reducing the correlation between pilot sequences in comparison with the other two approaches, implying capability of sufficiently mitigating the inter-user interference even in highly non-orthogonal scenarios and efficiently decreasing the number of pilot lengths simultaneously.

III. Joint Detection via Bilinear Gaussian Belief Propagation

In this section, we describe a joint activity, channel, and data estimation mechanism via bilinear GaBP for large cell-free MIMO architectures, whose belief propagation can be modeled as the graph schematized in Figure 3.

In order to further clarify the process of the proposed method, a work flow chart of the proposed detection mechanism is also illustrated in Figure 4, where the detection procedure is split into two algorithms, namely, Algorithm 1) Belief consensus and Algorithm 2) Hard decision. As shown in the figure, the work flow starts with an initialization where a first guess of the channel estimate is obtained using only the pilot sequences, the result of which is however limited in accuracy due to the severely non-orthogonal structure of the pilot matrix as $K_p \ll M$. Aiming at improving the channel estimation accuracy by jointly detecting the data as well as the activity, the initial channel guess obtained by the initialization process is fed to Algorithm 1, which as illustrated in Figure 3, is composed of two different stages; 1) soft interference cancellation (SIC) and beliefs generation based on tentative soft estimates and 2) combining beliefs and soft estimates generation, described in the next two subsections, respectively.

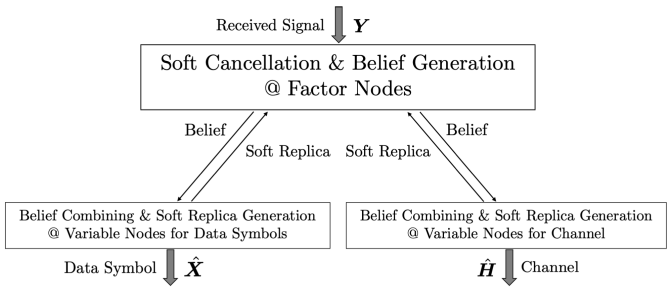


Fig. 3. Belief generation and combining model.

Followed by Algorithm 1, we proceed with Algorithm 2, where the final hard decision of the data and activity is carried out, which will be described in Section III-C.

A. Factor Nodes

Focusing on the received signal element y_{nk} at the n-th row and k-th column of $\boldsymbol{Y}$ with the aim of detecting x_{mk} at the m-th row and k-th column of $\boldsymbol{X}$, the received signal after SIC using soft estimates is given by

$$\tilde{y}_{m,nk} \triangleq y_{nk} - \overbrace{\sum_{i\neq m}^{M} \hat{h}_{k,ni}\hat{x}_{n,ik}}^{\text{Inter-user interference cancellation with soft-replicas}} \tag{12a}$$

$$= h_{nm}x_{mk} + \underbrace{\sum_{i\neq m}^{M}(h_{ni}x_{ik} - \hat{h}_{k,ni}\hat{x}_{n,ik})}_{\text{Residual interference}} + w_{nk}, \tag{12b}$$

where $\hat{x}_{n,ik}$ and $\hat{h}_{k,ni}$ with $i \in \{1, 2, \ldots, M\}$ are tentative estimates of x_{ik} and h_{ni}, respectively, generated at variable nodes in the previous iteration, and w_{nk} is the noise element at the n-th row and k-th column of $\boldsymbol{W}$.

Owing to the central limit theorem, the interference-plus-noise component can be approximated as a complex Gaussian random variable under large-system conditions, resulting in the fact that the conditional probability density function (PDF) of equation (12) for given x_{mk} and h_{nm} can be respectively expressed as

$$p_{\tilde{y}_{m,nk}|x_{mk}}(\tilde{y}_{m,nk}|x_{mk}) \propto e^{-\frac{|\tilde{y}_{m,nk} - \hat{h}_{k,nm}x_{mk}|^2}{v^x_{m,nk}}} \tag{13a}$$

$$p_{\tilde{y}_{m,nk}|h_{nm}}(\tilde{y}_{m,nk}|h_{nm}) \propto e^{-\frac{|\tilde{y}_{m,nk} - h_{nm}\hat{x}_{n,mk}|^2}{v^h_{m,nk}}}, \tag{13b}$$

with

$$v^x_{m,nk} \triangleq \sum_{i\neq m}^{M}\left\{|\hat{h}_{k,ni}|^2\psi^x_{n,ik} + (|\hat{x}_{n,ik}|^2 + \psi^x_{n,ik})\psi^h_{k,ni}\right\} + \psi^h_{k,nm} + N_0 \tag{14a}$$

$$v^h_{m,nk} \triangleq \sum_{i\neq m}^{M}\left\{|\hat{h}_{k,ni}|^2\psi^x_{n,ik} + (|\hat{x}_{n,ik}|^2 + \psi^x_{n,ik})\psi^h_{k,ni}\right\} + \gamma_{nm}\psi^x_{n,mk} + N_0, \tag{14b}$$

where $\psi^x_{n,ik}$ and $\psi^h_{k,ni}$ denote expected error variances corresponding to $\hat{x}_{n,ik}$ and $\hat{h}_{k,ni}$, respectively, which will be

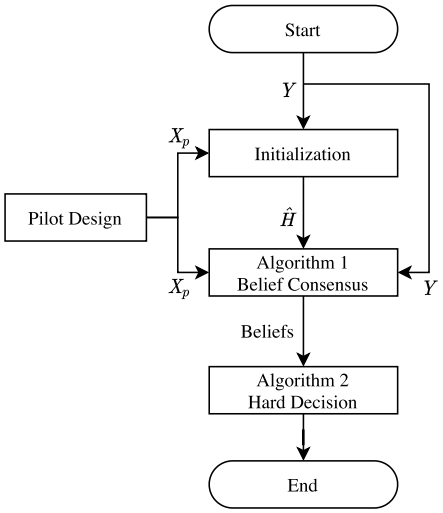


Fig. 4. Work flow of the proposed detection process.

derived in detail in the next subsection, and we remark that $\mathbb{E}\left[|x_{mk}|^2\right]=1$.

B. Variable Nodes

Given the SIC and belief generation above, one can combine the Gaussian beliefs given in equation (13a), yielding the PDF of an extrinsic belief $l^x_{n,mk}$ for x_{mk}

$$p_{l^x_{n,mk}|x_{mk}}(l^x_{n,mk}|x_{mk}) = \prod_{i\neq n}^{N} p_{\tilde{y}_{m,ik}|x_{mk}}(\tilde{y}_{m,ik}|x_{mk}) \quad (15a)$$

$$\propto e^{-\frac{|x_{mk}-\hat{r}_{n,mk}|^2}{\psi^r_{n,mk}}}, \quad (15b)$$

with

$$\psi^r_{n,mk} \triangleq \left(\sum_{i\neq n}^{N}\frac{|\hat{h}_{k,im}|^2}{v^x_{m,ik}}\right)^{-1} \quad (16a)$$

$$\hat{r}_{n,mk} \triangleq \psi^r_{n,mk}\sum_{i\neq n}^{N}\frac{\hat{h}^*_{k,im}\tilde{y}_{m,ik}}{v^x_{m,ik}}. \quad (16b)$$

In turn, since the activity can be expressed as column-sparsity in the channel matrix $\boldsymbol{H}$, combining beliefs of the m-th column of the channel matrix (*i.e.*, $\boldsymbol{h}_m$) needs to be jointly performed over $n\in\{1,2,\ldots,N\}$. Thus, the PDF of the extrinsic belief $\boldsymbol{l}^h_{k,m}$ for given $\boldsymbol{h}_m$ is given by

$$p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m) = \prod_{n=1}^{N}\prod_{i\neq n}^{K} p_{\tilde{y}_{m,ni}|h_{nm}}(\tilde{y}_{m,ni}|h_{nm}) \quad (17a)$$

$$\propto e^{-\left(\boldsymbol{h}_m-\boldsymbol{\mu}^h_{k,m}\right)^{\mathrm{H}}\boldsymbol{\Sigma}^{h^{-1}}_{k,m}\left(\boldsymbol{h}_m-\boldsymbol{\mu}^h_{k,m}\right)},$$

$$\propto \frac{e^{-\left(\boldsymbol{h}_m-\boldsymbol{\mu}^h_{k,m}\right)^{\mathrm{H}}\boldsymbol{\Sigma}^{h^{-1}}_{k,m}\left(\boldsymbol{h}_m-\boldsymbol{\mu}^h_{k,m}\right)}}{\pi^N|\boldsymbol{\Sigma}^h_{k,m}|}. \quad (17b)$$

Notice that the expression in equation (17b) is a complex multi-variate Gaussian PDF, that is

$$p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m) \propto \underbrace{\mathcal{CN}_N\left(\boldsymbol{\mu}^h_{k,m},\boldsymbol{\Sigma}^h_{k,m}\right)}_{N\text{-multivariate complex Gaussian distribution}}, \quad (18)$$

where

$$\boldsymbol{\mu}^h_{k,m} \triangleq [\hat{q}_{k,1m},\ldots,\hat{q}_{k,Nm}]^{\mathrm{T}} \quad (19a)$$

$$\boldsymbol{\Sigma}^h_{k,m} \triangleq \mathrm{diag}\left(\psi^q_{k,1m},\ldots,\psi^q_{k,Nm}\right), \quad (19b)$$

with

$$\psi^q_{k,nm} \triangleq \left(\sum_{i\neq k}^{K}\frac{|\hat{x}_{n,mi}|^2}{v^h_{m,ni}}\right)^{-1} \quad (20a)$$

$$\hat{q}_{k,nm} \triangleq \psi^q_{k,nm}\sum_{i\neq k}^{K}\frac{\hat{x}^*_{n,mi}\tilde{y}_{m,ni}}{v^h_{m,ni}}. \quad (20b)$$

Taking the expectation over the PDFs of the extrinsic beliefs given in equations (15) and (18), respectively, soft estimates of x_{mk} and $\boldsymbol{h}_m$ can be respectively obtained as

$$\hat{x}_{n,mk} = \sum_{x_q\in\mathscr{C}}\frac{x_q\cdot p_{l^x_{n,mk}|x_{mk}}(l^x_{n,mk}|x_q)p_{x_{mk}}(x_q)}{\sum_{x'_q\in\mathscr{C}}p_{l^x_{n,mk}|x_{mk}}(l^x_{n,mk}|x'_q)p_{x_{mk}}(x'_q)}, \quad (21a)$$

$$\hat{\boldsymbol{h}}_{k,m} = \int_{\boldsymbol{h}_m}\boldsymbol{h}_m\frac{p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m)p_{\boldsymbol{h}_m}(\boldsymbol{h}_m)}{\int_{\boldsymbol{h}'_m}p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}'_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}'_m)p_{\boldsymbol{h}_m}(\boldsymbol{h}'_m)}, \quad (21b)$$

where the denominators are introduced to normalize the integral of the posterior PDFs to 1.

Although a closed-form expression of equation (21a) is not known for arbitrary discrete constellations, for Gray-coded quadrature phase-shift keying (QPSK)[3] it can be written as [67]

$$\hat{x}_{n,mk} = \frac{1}{\sqrt{2}}\Big(\tanh\Big(\frac{\sqrt{2}\Re(\hat{r}_{n,mk})}{\psi^r_{n,mk}}\Big)+j\tanh\Big(\frac{\sqrt{2}\Im(\hat{r}_{n,mk})}{\psi^r_{n,mk}}\Big)\Big), \quad (22)$$

with the corresponding error variance estimate given by

$$\psi^x_{n,mk} = 1-|\hat{x}_{n,mk}|^2. \quad (23)$$

A closed-form of equation (21b) can be obtained as follows. First, define the effective PDF

$$P^h_{k,m}(\boldsymbol{h}_m) \triangleq p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m)p_{\boldsymbol{h}_m}(\boldsymbol{h}_m) = \frac{1}{\pi^N}$$

$$\times\left[\frac{\lambda e^{-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}\boldsymbol{\mu}^h_{k,m}}\mathcal{CN}_N\left(\boldsymbol{\Gamma}_m\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}\boldsymbol{\Gamma}_m,\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}\boldsymbol{\Sigma}^h_{k,m}\right)}{|\boldsymbol{\Gamma}_m+\boldsymbol{\Sigma}^h_{k,m}|}\right.$$

$$\left.+\frac{(1-\lambda)\delta(\boldsymbol{h}_m)e^{-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\boldsymbol{\Sigma}^{h^{-1}}_{k,m}\boldsymbol{\mu}^h_{k,m}}}{|\boldsymbol{\Sigma}^h_{k,m}|}\right]. \quad (24)$$

Then, the normalizing factor in the denominator of equation (21b) can be calculated by integrating the latter over the entire N-dimensional complex field, which yields

$$C^h_{k,m} \triangleq \int_{\boldsymbol{h}'_m}P^h_{k,m}(\boldsymbol{h}'_m) \quad (25a)$$

$$= \frac{\lambda\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Gamma}_m+\boldsymbol{\Sigma}^h_{k,m}|}\tau_{k,m}, \quad (25b)$$

[3]For higher-order modulations, by running the algorithm on the equivalent PAM real-valued model [68], the computational cost required to evaluate equations (21a) and its MSE grows with $\mathcal{O}\left(\sqrt{Q}NMK\right)$, where Q denotes the modulation order.

with the activity detection factor

$$\tau_{k,m} \triangleq 1 + \frac{1-\lambda}{\lambda} \exp\big(- \big(\pi^h_{k,m} - \psi^h_{k,m}\big)\big), \tag{26a}$$

where

$$\pi^h_{k,m} \triangleq \boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\big(\boldsymbol{\Sigma}^{h\,-1}_{k,m} - \Big(\boldsymbol{\Sigma}^h_{k,m} + \boldsymbol{\Gamma}_m\Big)^{-1}\big)\boldsymbol{\mu}^h_{k,m} \tag{26b}$$

$$\psi^h_{k,m} \triangleq \log\Big(\big|\boldsymbol{\Sigma}^{h\,-1}_{k,m}\boldsymbol{\Gamma}_m + \mathbf{I}_N\big|\Big). \tag{26c}$$

With possession of equations (24) and (25), whose detailed derivations are given in Appendix B, the soft replica of $\boldsymbol{h}_m$ for a given effective distribution $P^h_{k,m}(\boldsymbol{h}_m)$ can be obtained as

$$\hat{\boldsymbol{h}}_{k,m} = \frac{\boldsymbol{\Gamma}_m\big(\boldsymbol{\Sigma}^h_{k,m} + \boldsymbol{\Gamma}_m\big)^{-1}}{\tau_{k,m}}\boldsymbol{\mu}^h_{k,m}, \tag{27}$$

and for the corresponding error variance, introducing $\boldsymbol{\Psi}^h_{k,m} \triangleq \mathrm{diag}\left(\psi^h_{k,1m}, \ldots, \psi^h_{k,Nm}\right)$ yields

$$\begin{aligned}\boldsymbol{\Psi}^h_{k,m} &= \mathrm{diag}\bigg(\int_{\boldsymbol{h}_m} \boldsymbol{h}_m\boldsymbol{h}^{\mathrm{H}}_m \frac{P^h_{k,m}(\boldsymbol{h}_m)}{C_{k,m}} - \hat{\boldsymbol{h}}_{k,m}\hat{\boldsymbol{h}}^{\mathrm{H}}_{k,m}\bigg)\\ &= (\tau_{k,m}-1)\mathrm{diag}(\hat{\boldsymbol{h}}_{k,m}\hat{\boldsymbol{h}}^{\mathrm{H}}_{k,m}) + \frac{\boldsymbol{\Sigma}^h_{k,m}\boldsymbol{\Gamma}_m(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m)^{-1}}{\tau_{k,m}}.\end{aligned} \tag{28}$$

C. Algorithm Description

In this subsection we summarize the belief propagation and consensus mechanisms described above and schematized in Figure 4, offering also detailed discussion on the algorithmic flow. For starters, the schemes are concisely described in the form of pseudo-codes in Algorithm 1 and Algorithm 2, respectively. For the sake of brevity, we define two sets of integers ($i.e.$, $\mathscr{K}_p$ and $\mathscr{K}_d$), which are respectively given by $\mathscr{K}_p \triangleq \{1, 2 \ldots, K_p\}$ and $\mathscr{K}_d \triangleq \{K_p+1, K_p+2 \ldots, K\}$.

As shown in the pseudo-codes, Algorithm 1 requires five different inputs: the received signal matrix $\boldsymbol{Y}$, the pilot sequence $\boldsymbol{X}_p$, an initial guess of the channel matrix $\hat{\boldsymbol{H}}$, an initial guess of the estimation error variance $\hat{\boldsymbol{\Psi}}^h$ corresponding to $\hat{\boldsymbol{H}}$, and the maximum number of iterations $t_{\max}$; while Algorithm 2 is fed with the beliefs obtained from Algorithm 1. We point out that rough estimates $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{\Psi}}^h$ can be obtained via state-of-the-art algorithms proposed for grant-free systems [17], [19], [20], although such mechanisms suffer from estimation inaccuracy in case of severely non-orthogonal pilot sequence ($K_p \ll M$). We emphasize again that reducing the overhead is however desired from a system-level perspective in terms of time resource efficiency. In Section III-D, the initialization process adopted in this article will be discussed in detail.

In turn, the outputs of Algorithm 2 are the following three quantities: an estimated symbol matrix $\hat{\boldsymbol{X}}$, an estimated channel matrix $\hat{\boldsymbol{H}}$, and estimates of active-user indexes $\hat{\mathscr{A}}$. As for $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{H}}$, they are obtained by combining all the beliefs ($i.e.$, consensus), while $\hat{\mathscr{A}}$ is determined by following a certain activity detection policy based on the estimated channel $\hat{\boldsymbol{H}}$. The activity detection scheme considered in this article will be described in detail in Section III-E.

Algorithm 1 Bilinear GaBP (Part 1: Belief Consensus)

Input: $\boldsymbol{Y}$, $\boldsymbol{X}_p$, $\hat{\boldsymbol{H}}$, $\hat{\boldsymbol{\Psi}}^h$, λ, $t_{\max}$
Output: $\forall k \in \mathscr{K}_d$, $\forall m$, $\forall n$: $\hat{x}_{n,mk}$, $\psi^x_{n,mk}$, $\forall k$, $\forall m$: $\hat{\boldsymbol{h}}_{k,m}$, $\boldsymbol{\Psi}^h_{k,m}$, $\forall m$: $\boldsymbol{\Gamma}_m$

1: For $k \in \mathscr{K}_p$, set $\hat{x}_{n,mk}(1) = \big[\boldsymbol{X}_p\big]_{mk}$ and $\psi^x_{n,mk}(1) = 0$, $\forall(m,n)$
2: For $k \in \mathscr{K}_d$, set $\hat{x}_{n,mk}(1) = 0$ and $\psi^x_{n,mk}(1) = 1$, $\forall(m,n)$
3: Set $\hat{h}_{k,nm}(1) = \big[\hat{\boldsymbol{H}}\big]_{nm}$ and $\psi^h_{k,nm}(1) = \big[\hat{\boldsymbol{\Psi}}^h\big]_{nm}$, $\forall(k,m,n)$
4: **for** $t = 1, \ldots, t_{max}$
For all (k,m,n):
5: Obtain from eq. (12) the received signals after SIC
6: Obtain from eq. (14a) the symbol estimate variances
7: Obtain from eq. (14b) the channel estimate variances
8: Obtain from eq. (16a) $\psi^r_{n,mk}(t)$ and $\hat{r}_{n,mk}(t)$
9: Obtain from eq. (20a) $\psi^q_{k,nm}(t)$ and $\hat{q}_{k,nm}(t)$
For all (k,m):
10: Obtain from eq. (19a) $\boldsymbol{\mu}^h_{k,m}(t)$ and $\boldsymbol{\Sigma}^h_{k,m}(t)$
11: Obtain from eq. (26a) $\tau_{k,m}(t)$, with $\pi^h_{k,m}(t)$ and $\psi^h_{k,m}(t)$
12: Obtain from eq. (27) $\hat{\boldsymbol{h}}'_{k,m}(t+1)$
13: Obtain $\hat{\boldsymbol{h}}_{k,m}(t+1) = \eta\,\hat{\boldsymbol{h}}'_{k,m}(t+1) + (1-\eta)\hat{\boldsymbol{h}}_{k,m}(t)$
14: Obtain from eq. (28) $\boldsymbol{\Psi}^{h'}_{k,m}(t+1)$
15: Obtain $\boldsymbol{\Psi}^h_{k,m}(t+1) = \eta\,\boldsymbol{\Psi}^{h'}_{k,m}(t+1) + (1-\eta)\boldsymbol{\Psi}^h_{k,m}(t)$
For all (m,n) and $k \in \mathscr{K}_d$:
16: Obtain from eq. (22) the symbol estimates $x'_{n,mk}(t+1)$
17: Obtain $\hat{x}_{n,mk}(t+1) = \eta\,x'_{n,mk}(t+1) + (1-\eta)\hat{x}_{n,mk}(t)$
18: Obtain $\psi^x_{n,mk}(t+1) = \frac{\eta(1-|\hat{x}_{n,mk}(t)|^2)}{\tau_{k,m}} + (1-\eta)\psi^x_{n,mk}(t)$
19: **end for**

The flow of Algorithms 1 and 2 generally follows the belief exchange steps described in Section III-A and III-B, but in Algorithm 1, two belief manipulation techniques, namely, damping [68] and scaling [67] were introduced, with the objective of further improving the estimation accuracy and escape from local minima. This is due to the fact that when the Gaussian approximation assumed in equation (12) does not sufficiently describe the actual stochastic behavior of the effective noise, the accuracy of soft-replicas is degraded by resultant belief outliers caused by the aforementioned approximation gap, leading to non-negligible estimation performance deterioration. Such approximation gap results from the increase in uncertainty that follows especially when the length of pilot sequences decreases, which is the very aim of this article.

The damping steps [3], [7] introduced with damping factor $\eta \in [0,1]$ in lines 13, 15, 17 and 18 aims to prevent the algorithm from converging to local minima by forcing a slow update of soft-replicas, whereas belief scaling with parameter $\gamma(t)$ is adopted in line 16, in order to adjust the reliability of beliefs ($i.e.$, harnessing harmful outliers). The scaling

parameter is designed to be a linear function of the number of iterations, that is,

$$\gamma(t) = \frac{t}{t_{\max}}. \tag{29}$$

In addition to the above, as shown in Section III-B, soft estimates of x_{mk} are obtained without considering the user activity ($i.e.$, row-sparsity) by imposing user activity detection upon the channel estimation process as described in equation (21b), indicating that such row-sparsity of $\boldsymbol{X}$ needs to be incorporated so as to avoid inconsistency with column-sparsity of $\boldsymbol{H}$. In order to resolve this issue, the activity detection factor $\tau_{k,m}$ given in equation (26) is introduced in lines 16 and 18 of Algorithm 1. Notice that $\tau_{k,m} \to 1$ when a given user is active and to ∞ otherwise, such that the rows of $\boldsymbol{X}$ corresponding to non-active columns of $\boldsymbol{H}$ are suppressed.

D. Initialization

Due to the fact that bilinear inference problems are strongly affected by the initial values of the solution variables, a reasonable initialization method is required so that the algorithm accurately estimates the channel, the informative data and the activity pattern simultaneously. However, one may also notice that due to the severe non-orthogonality of the pilot sequence ($i.e.$, $K_p \ll |\mathscr{A}| \ll M$) for overhead reduction, the accuracy of such an initial guess is not reliable enough. In light of the above, although several approaches developed for grant-free access such as covariance-based methods [17], [44] can be considered to produce initial $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{\Psi}}^h$, in this article we have leveraged MMV-AMP [20] from a computational complexity perspective,[4] which can be simply applied to equation (3) by regarding the first K_p columns of $\boldsymbol{Y}$ and the pilot matrix as the effective received signal matrix and its measurement matrix, respectively.

E. Activity Detection Policy

In this section, we describe how to identify active users based on the belief consensus performed in Algorithm 2. Due to the fact that miss-detections (MDs) and false alarms (FAs) are in a trade-off relationship as shown in the grant-free literature, such an activity detection policy is affected by system and user requirements, indicating that one needs to adopt a suitable criterion depending on the situation in practical implementations.

With that in mind, we consider the log likelihood ratio method based on estimated channel quantities, which thanks to uncorrelated Gaussianity of the channel and residual estimation error, can be written as

$$\mathrm{LLR}_m \triangleq \ln \frac{\prod_{n=1}^{N} \mathcal{CN}\left(0, \gamma_{nm} + \psi_{nm}^h | \hat{h}_{nm}\right)}{\prod_{n=1}^{N} \mathcal{CN}\left(0, \psi_{nm}^h | \hat{h}_{nm}\right)}, \tag{30}$$

[4]Please refer to [16] for complexity analyses between existing grant-free schemes for further details.

Algorithm 2 Bilinear GaBP (Part 2: Hard Decision)

Input: $\boldsymbol{Y}$, $\forall k$, $\forall m$, $\forall n$: $\hat{x}_{n,mk}$, $\psi_{n,mk}^x$, $\forall k$, $\forall m$: $\hat{\boldsymbol{h}}_{k,m}$, $\boldsymbol{\Psi}_{k,m}^h$, $\forall m$: $\boldsymbol{\Gamma}_m$
Output: $\hat{\boldsymbol{X}}$, $\hat{\boldsymbol{H}}$, $\hat{\mathscr{A}}$

For all (k, m, n):
1: Obtain from eq. (12) the received signals after SIC
2: Obtain from eq. (14b) the channel estimate variances
For all (m, n):
3: Obtain $\psi_{nm}^q = \left(\sum_{i=1}^{K} \frac{|\hat{x}_{n,mi}|^2}{v_{m,ni}^h}\right)^{-1}$
and $\hat{q}_{nm} = \psi_{nm}^q \sum_{i=1}^{K} \frac{\hat{x}_{n,mi}^* \tilde{y}_{m,ni}}{v_{m,ni}^h}$
For all (m, n) and $k \in \mathscr{K}_d$:
4: Obtain from eq. (14a) the symbol estimate variances
5: Obtain $\psi_{mk}^r = \left(\sum_{i=1}^{N} \frac{|\hat{h}_{k,im}|^2}{v_{m,ik}^x}\right)^{-1}$
and $\hat{r}_{mk} = \psi_{mk}^r \sum_{i=1}^{N} \frac{\hat{h}_{k,im}^* \tilde{y}_{m,ik}}{v_{m,ik}^x}$
6: Obtain the soft symbol estimates
$x'_{mk} = \frac{1}{\sqrt{2}}\left[\tanh\left(\frac{\sqrt{2}\Re(\hat{r}_{mk})}{\psi_{mk}^r}\right) + j \cdot \tanh\left(\frac{\sqrt{2}\Im(\hat{r}_{mk})}{\psi_{mk}^r}\right)\right]$
7: Obtain the hard symbol estimates
$\hat{x}_{mk} = \underset{x_q \in \mathscr{C}}{\operatorname{argmin}} \ |x_q - x'_{mk}|$
For all m:
8: Obtain $\boldsymbol{\mu}_m^h = [\hat{q}_{1m}, \ldots, \hat{q}_{Nm}]^{\mathrm{T}}$
and $\boldsymbol{\Sigma}_m^h = \mathrm{diag}(\psi_{1m}^q, \ldots, \psi_{Nm}^q)$
9: Obtain $\tau_m = 1 + \frac{1-\lambda}{\lambda} \exp(-(\pi_m^h - \psi_m^h))$
with $\pi_m^h = \boldsymbol{\mu}_m^{h\,\mathrm{H}} (\boldsymbol{\Sigma}_m^{h^{-1}} - (\boldsymbol{\Sigma}_m^h + \boldsymbol{\Gamma}_m)^{-1}) \boldsymbol{\mu}_m^h$
and $\psi_m^h = \log(|\boldsymbol{\Sigma}_m^{h^{-1}} \boldsymbol{\Gamma}_m + \mathbf{I}_N|)$
10: Obtain $\hat{\boldsymbol{h}}_m = \frac{\boldsymbol{\Gamma}_m (\boldsymbol{\Sigma}_m^h + \boldsymbol{\Gamma}_m)^{-1}}{\tau_m} \boldsymbol{\mu}_m^h$
11: Update the set of active users $\mathscr{A}$ via eq. (33)

where LLR_m is the log likelihood ratio corresponding to the m-th column and $\prod_{n=1}^{N}$ is introduced to perform consensus over the receive antenna dimension.

After basic manipulations, the log likelihood criterion can then be simplified to

$$\mathrm{LLR}_m = p_{\mathrm{active}}(m) - p_{\mathrm{inactive}}(m), \tag{31}$$

where

$$p_{\mathrm{active}}(m) \triangleq \sum_{n=1}^{N} \frac{-|\hat{h}_{nm}|^2}{\gamma_{nm} + \psi_{nm}^h} + \ln\left(\frac{1}{\pi(\gamma_{nm} + \psi_{nm}^h)}\right) \tag{32a}$$

$$p_{\mathrm{inactive}}(m) \triangleq \sum_{n=1}^{N} \frac{-|\hat{h}_{nm}|^2}{\psi_{nm}^h} + \ln\left(\frac{1}{\pi \psi_{nm}^h}\right). \tag{32b}$$

With basis on the above, the set of active users is then determined as follows

$$\mathscr{A} = \left\{ m \,\middle|\, p_{\mathrm{active}}(m) > p_{\mathrm{inactive}}(m) \right\}. \tag{33}$$

IV. Performance Assessment

In this section, we evaluate via software simulation the proposed method in terms of bit error rate (BER),

effective throughput, normalized mean square error (NMSE), and AUD performance.

A. Simulation Setup

Throughout this performance assessment section, we consider the following simulation setup unless otherwise specified. The number of receive antennas is set to $N = 100$, which are distributed over a square of side of 1000 [m] in a square mesh fashion, where $M = 100$ potential users are accommodated in each subcarrier. It is assumed that 50% of the M users become active during each OFDM frame, $i.e.$, $|\mathscr{A}| = M/2 = 50$, while K and K_p are considered to be $K \in \{140, 280\}$ and $K_p = 14$ depending on the employed subcarrier spacing.[5] It is further worth-noting that since we accommodate $M = 100$ users with overhead length $K_p = 14$, a significant amount of overhead reduction can be achieved. Although one might concern about performance degradation due to the resultant severe non-orthogonality, we dispel such concerns throughout this section by demonstrating that the bilinear inference method employed here is able to handle the non-orthogonality.

The covariance matrix employed in the multivariate Bernoulli-Gaussian model of equation (2) is constructed following the 3GPP urban microcell model [69], $i.e.$, $\mathbf{\Gamma}_m = \mathrm{diag}\left([\gamma_{1m}, \ldots, \gamma_{Nm}]\right)$, with each diagonal element obeying the relation $\gamma_{nm} \triangleq 10^{-\frac{\beta_{nm}}{10}}$ and β_{nm} [dB] $= 30.5 + 36.7 \log_{10}(d_{nm}) + \mathcal{N}\left(0, 4^2\right)$ where d_{nm} denotes the distance between the n-th AP and m-th user, given by $d_{nm} = \sqrt{(\rho_{AP} - \rho_U)^2 + \rho_R^2}$, with $\rho_{AP} = 10$ [m] and $\rho_U = 1.65$ [m] corresponding to the heights of the APs and users, respectively, while ρ_R is a random quantity.

The transmit power range at each uplink user is determined based on the experimental study presented in [70], where the transmit power of each uplink user is limited by 16 [dBm]. Furthermore, Gray-coded QPSK modulation is assumed to be employed at each user, whereas the noise floor N_0 at each AP is assumed to be modeled as

$$\sigma_u^2 = 10 \log_{10}(1000 \kappa T) + \mathrm{NF} + 10 \log_{10}(W) \,[\mathrm{dBm}], \quad (34)$$

where κ is the Boltzmann's constant, $T = 293.15$ denotes the physical temperature at each AP in kelvins, the noise figure NF is assumed to be 5 [dB] and W expresses the subcarrier bandwidth.

As described in Section II-A, the pilot structure is designed via quadratic CSIDCO in order to mitigate pilot contamination effects as much as possible even in case of severely non-orthogonal scenarios such as one considered in this section. Regarding the effective throughput performance, we adopt the definition proposed in [71, Def. 1], which is given by

$$\mathrm{R}_{\mathrm{eff}} \triangleq (1 - P_e) \cdot K_d \cdot b, \quad (35)$$

where P_e denotes the block (packet) error rate and b is the number of bits per OFDM symbol.

Finally, the maximum number of iterations in Algorithm 1 is set to $t_{\max} = 32$ and the damping factor η is 0.5, while the belief scaling factor $\gamma(t)$ is defined in equation (29).

B. Computational Complexity

Before proceeding to the performance evaluation via software simulations, we describe the computational complexity per iteration of the proposed JACDE algorithm, comparing it against those of different reference methods considered.

As can be seen from Algorithms 1 and 2, all the calculations required by the proposed method can be carried out in an element-wise manner. Taking into account the fact that the matrices $\mathbf{\Sigma}_{k,m}^h$ and $\mathbf{\Gamma}_m$ are both diagonal, it follows that Algorithm 1 has complexity of order $\mathcal{O}(NMK)$ both for multiplication/division and for addition/subtraction operators, resulting also in the complexity order per iteration of $\mathcal{O}(NMK)$ in total, including all steps to jointly perform CE, AUD and MUD.

For the sake of comparison, as described in Section IV-C, we consider the MMV-AMP algorithm [11]–[14] as state-of-the-art for JACE, and either the GaBP algorithm or the conventional zero-forcing (ZF) method as state-of-the-art for MUD. The complexity order per iteration for JACE via MMV-AMP is of $\mathcal{O}(NMK_p)$, whereas the complexity of $\mathcal{O}(NK_d|\hat{\mathscr{A}}_{\mathrm{MMV}}|)$ and $\mathcal{O}(|\hat{\mathscr{A}}_{\mathrm{MMV}}|^3 + |\hat{\mathscr{A}}_{\mathrm{MMV}}|^2 K_d)$ is imposed for MUD by the GaBP algorithm and the ZF method [67], respectively, where $|\hat{\mathscr{A}}_{\mathrm{MMV}}| \leq M$ denotes the number of active users estimated by MMV-AMP.

Assuming that MMV-AMP is capable of estimating active patterns with reasonable accuracy in the considered setup ($i.e.$, $|\hat{\mathscr{A}}_{\mathrm{MMV}}| \approx M/2$), the complexity of the MMV-AMP algorithm followed by the GaBP method can be approximated written as $\mathcal{O}(NM(K_p + K_d/2))$. Therefore, it can be concluded that the proposed JACDE method is approximately in the same order of complexity of state-of-the-art alternatives.

C. Multi-user Detection

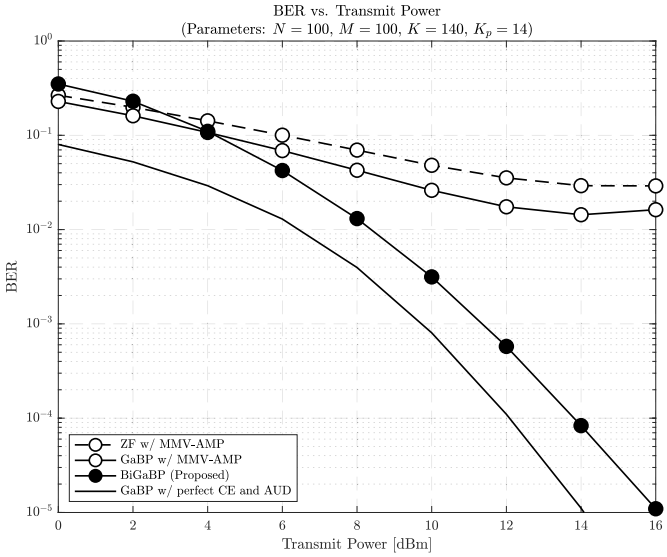
The MUD performance of the proposed algorithm is studied in terms of uncoded BER as a function of transmit power. In order to take into account MD effects on data detection, we count not only bits received in error but also the number of lost bits due to missing user activity, $i.e.$,

$$\mathrm{BER} = \frac{P_e^1 + P_e^2}{\text{Total number of bits}} \quad (36)$$

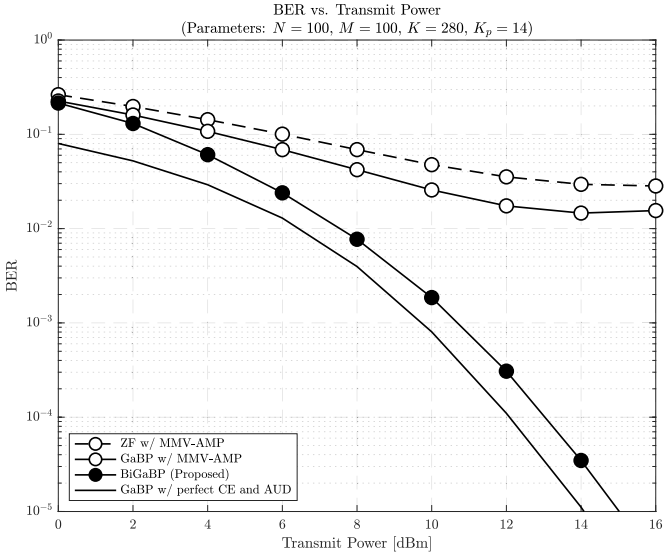
where P_e^1 denotes the number of errors due to failure of symbol detection and P_e^2 is the number of bits that have been lost due to failure of user detection.

Since there are no existing cell-free scheme with grant-free access which do not rely on spreading data sequences as described in equation (3), we consider the MMV-AMP algorithm as a state-of-the-art method to carry out JACE with basis of non-orthogonal pilot sequences $\boldsymbol{X}_p$, remarking that this receiver is widely employed in related literature [11]–[14]. For the same reason (of lack of a direct equivalent competitor),

[5] A scenario with $K = 140$ and $K = 280$ corresponds to OFDM subcarrier spacing of 15 [kHz] and 30 [kHz], respectively. As shown in Figure 1, $K_p = 14$ indicates that only one OFDM slot is utilized as pilot and the rest as data transmission, indicating that this situation imposes the most severe scenario as the pilot length is minimum.



(a) 15 kHz subcarrier spacing ($N = M = 100$, $K_p = 14$).



(b) 30 kHz subcarrier spacing ($N = M = 100$, $K_p = 14$).

Fig. 5. BER comparisons as a function of transmit power.

we also compare the performance of our method against an idealized system in which perfect CE and AUD is assumed, with signal detection performed by the GaBP algorithm.

Our assessment starts with Figure 5, where the BER performance of the proposed method is compared not only to the state-of-the-art but also to the ideal performance, for different subcarrier spacing scenarios ($i.e.$, $K \in \{140, 280\}$). The state-of-the-art methods compared are the linear ZF MIMO detector and the GaBP message passing MIMO detector, followed by MMV-AMP-based CE with the aid of the non-orthogonal pilot sequence. Results obtained from the GaBP MIMO detector with perfect CE and AUD provide lower bounds on the evaluated methods.

As can be seen from both figures, state-of-the-art methods suffer from high error floors, which stem from the poor CE and AUD performances by MMV-AMP, caused by the severely overloaded condition. In fact the aspect ratio (number of users over number of pilot symbols) of the pilot matrix is $\frac{M}{K_p} = \frac{100}{14} \approx 7.1428$, indicating a highly non-orthogonal condition. Although only $m = 50$ users out of $M = 100$ are assumed to be active in each coherent frame, the overloading ratio is still sufficiently high to hinder CE and AUD.

In contrast, the proposed method enjoys a water-falling curve in terms of BER for the both situations, which can be achieved by taking advantage of the pseudo-orthogonality of the data structure. This advantage can be confirmed from the fact that increasing the total symbol length from $K = 140$ to $K = 280$ while fixing the pilot length to be $K_p = 14$ can indeed enhance the detection performance as shown in Figure 5(a) and 5(b). One may readily notice from the above that the corresponding CE performance can be also improved due to the same logic, which is offered in Section IV-E below.

D. Effective Throughput

In light of the definition given in equation (35), we next investigate the effective throughput performance per each OFDM frame of the proposed method. Please note that since the OFDM frame corresponds to 10 [ms], the effective throughput implies the number of successfully delivered bits within a resource block of 10 [ms] times an OFDM subcarrier.

Simulation results showing the effective throughput achieved with the proposed scheme and compared alternatives are offered in Figure 6 for different subcarrier spacing setups. Note that the unit of the vertical axis is set to *kilobits* per frame for the sake of readability. Furthermore, we also implicitly measure the packet (block) error performance of the methods as shown in equation (35), which is often used for practical performance assessment. Finally, in addition to the three counterparts considered in the previous section, we also offer in both figures the system-level achievable maximum data rate as reference, which is determined by

$$\text{Maximum Capacity} \triangleq K_d \cdot |\mathscr{A}| \cdot |\mathscr{C}| \quad \text{[bits/frame]}, \quad (37)$$

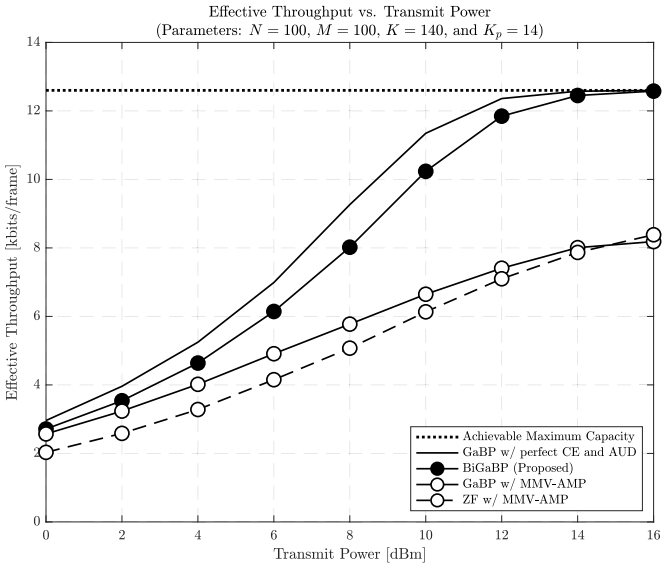
where $|\mathscr{C}|$ denotes the number of bits per symbol.

As expected from the discussion of the previous section, it is found that the two state-of-the-art alternatives are incapable of successfully delivering bits transmitted by the active users even with sufficiently high transmit power.

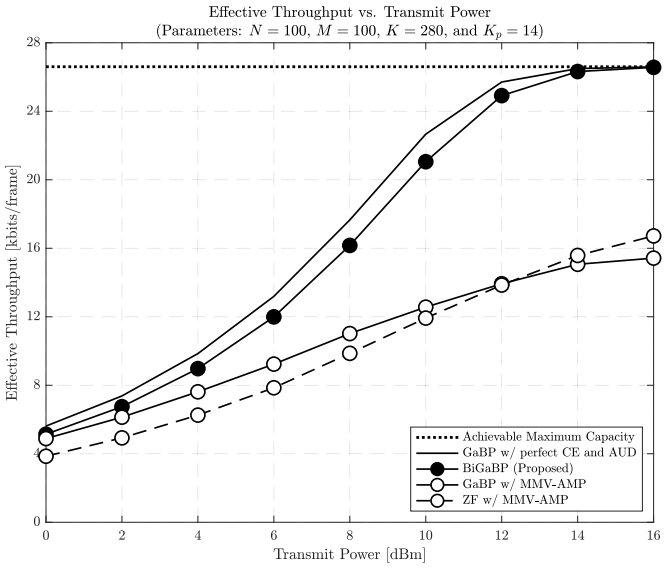
In contrast, the proposed method dynamically follows the same improvement in performance with transmit power as the idealized receiver, approaching the achievable capacity as the transmit power increases. It is furthermore seen that, as inferred in Section IV-C, the throughput gap from the idealized scheme narrows as the subcarrier spacing increases, which is again due to the pseudo-orthogonality of the data sequence.

E. Channel Estimation

In addition to the above, the CE performance of the proposed method is assessed in this section as a function of transmit power for different data lengths, so that one may observe that the performance improvement described in the above sections is, at least in part, induced by the resultant CE.

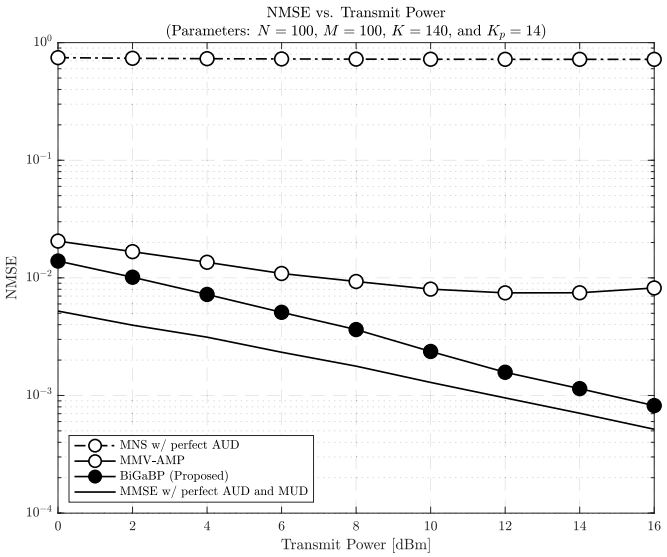


(a) 15 kHz subcarrier spacing ($N = M = 100$, $K_p = 14$).

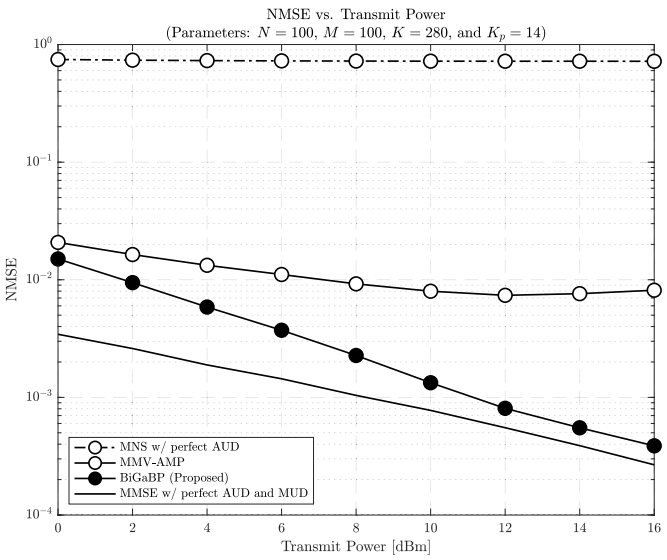


(b) 30 kHz subcarrier spacing ($N = M = 100$, $K_p = 14$).

Fig. 6. Effective throughput comparisons as a function of transmit power.



(a) 15 kHz subcarrier spacing ($N = M = 100$, $K_p = 14$).



(b) 30 kHz subcarrier spacing ($N = M = 100$, $K_p = 14$).

Fig. 7. NMSE comparisons as a function of transmit power.

To this end, the NMSE performance of the proposed method for $K = 140$ and $K = 280$ is offered in Figure 7(a) and 7(b), respectively, where the NMSE is defined as

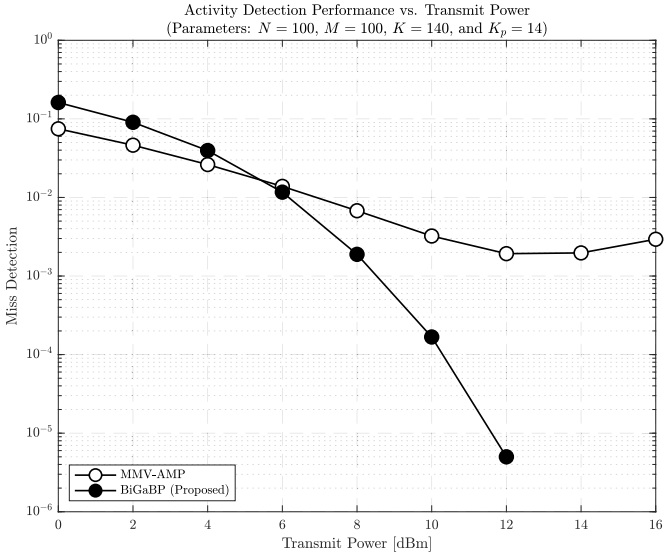
$$\text{NMSE} \triangleq \frac{\|\boldsymbol{H} - \hat{\boldsymbol{H}}\|_{\text{F}}^2}{\|\boldsymbol{H}\|_{\text{F}}^2}, \tag{38}$$

assuming a fixed pilot length $K_p = 14$.

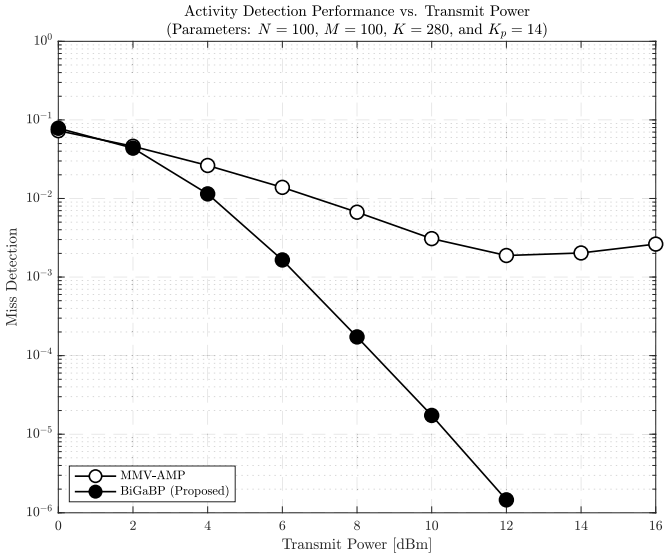
As for methods to compare, we have adopted not only MMV-AMP but also minimum norm solution (MNS) that is known to be a method to seek a closed-form unique CE solution in case of a non-orthogonal pilot sequence [7], while employing the minimum mean square error (MMSE) performance with perfect knowledge of AUD and MUD at the receiver as reference. Please note that since the non-Bayesian approach, which takes advantage of the sample covariance of the received signals in order to detect user activity patterns, aims at only AUD, the resultant performance in terms of CE can be lower-bounded by MNS with perfect AUD.

With that in mind, it can be observed from Figure 7(a) and 7(b) that the proposed method can indeed improve the CE performance and approach the unachievable MMSE performance with perfect AUD and MUD, maintaining a similar gradient with that of the MMSE, whereas MMV-AMP and MNS suffer from a relatively high error floor due to the non-orthogonality of the pilot, although MMV-AMP appears to offer moderate performance in comparison with MNS.

Thanks to the pseudo-orthogonality of the data structure, the proposed method with 30 kHZ subcarrier spacing again outperforms its own NMSE with 15 kHZ subcarrier spacing. Furthermore, it can be mentioned that due to the sufficiently high CE accuracy of the proposed method ($i.e.$, NMSE $\in [10^{-3}, 10^{-4}]$), the considered non-coherent transmission

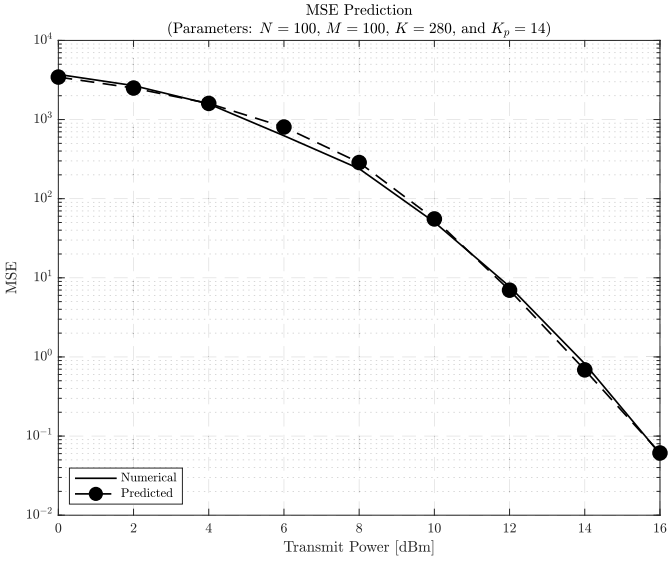


(a) 15 kHz subcarrier spacing ($N = M = 100$, $K_P = 14$).

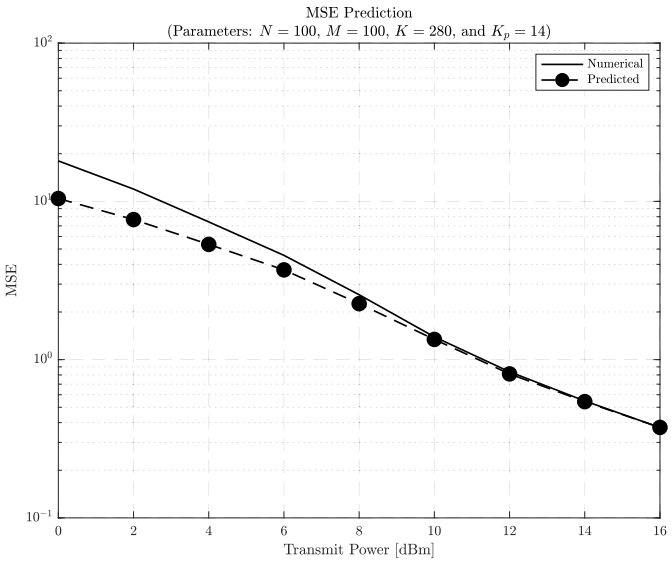


(b) 30 kHz subcarrier spacing ($N = M = 100$ K $= 14$).

Fig. 8. MD comparisons as a function of transmit power.



(a) MSE error of $\hat{\boldsymbol{X}}$.



(b) MSE error of $\hat{\boldsymbol{H}}$

Fig. 9. MSE performance prediction via the state evolution of Algorithm 1 in comparison with the actual numerical evaluation as a function of transmit power.

architecture is comparable to the CE performance of the conventional coherent MIMO-OFDM systems (please refer to, for instance, [72]) to verify this claim.

F. Active User Detection

In this section, we evaluate the AUD performance of the proposed BiGaBP method. Although the AUD performance may be examined in terms of either FA, MD, or both, FA can be removed at higher layers by leveraging cyclic redundancy check codes [12], which are widely employed in practice. In light of the above, in this article, we adopt the occurrence of MDs as an AUD performance index.

In Figure 8, the MD probabilities of the proposed BiGaBP and MMV-AMP algorithms are illustrated for different symbol lengths K as a function of transmit power at each uplink user, while assuming $K_p = 14$ for both scenarios. It is perceived from Figure 8 that the proposed method can exponentially reduce the occurrence of MDs as transmit power increases, the reason of which can be explained from the discussions given in the preceding sections as follows.

As observed in Figures 5-7, the proposed BiGaBP algorithm starts to gradually recover the data and the channel from the observations $\boldsymbol{Y}$ as transmit power increases, which stems from the fact that the residual noise variances given in equation (23) and (28) are also accordingly reduced. Consequently, the resultant LLR given in (30) intends to be positive when $|\hat{h}_{nm}|$ is not sufficiently close to 0 and negative when $\prod_{n=1}^{N} \mathcal{CN}\left(0, \gamma_{nm} + \psi_{nm}^{h} | \hat{h}_{nm}\right) \approx 0$ in comparison with $\prod_{n=1}^{N} \mathcal{CN}\left(0, \psi_{nm}^{h} | \hat{h}_{nm}\right)$ for a small ψ_{nm}^{h}. Furthermore, the reason why the MD performance of MMV-AMP dete-

riorates in high transmit power regions can be explained as follows.

Besides the insufficient observations due to a non-orthogonal pilot structure, MMV-AMP suffers from the fact that in such a high signal-to-noise ratio (SNR) region, its estimation error noise variance becomes indistinguishable from the AWGN noise level at the receiver, leading to a tendency to regard inactive users as active and vice versa. In contrast, the proposed method mitigates this bottleneck by taking advantage of DoFs in the time domain.

G. MSE Performance Prediction and Its Accuracy

Finally, in this subsection we evaluate the accuracy of mean square error (MSE) tracking via the state evolution of the proposed BiGaBP algorithm[6], where the predicted MSE performances of the data and channel are obtained by equation (23) and (28), respectively. In particular, the predicted MSE for $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{H}}$ is compared with the corresponding simulated counterpart in Figures 9(a) and 9(b), respectively, with $K = 280$ and $K_p = 14$ and where the solid line and the dashed line with markers correspond to the simulated and predicted performance, respectively.

It can be observed from the figures that the state evolution can track the error performance of the proposed BiGaBP for both data and channel estimates, since the predicted MSEs follow approximately the same trajectory of its simulated counterpart.

As a final remark, it has been observed throughout the experiments that the proposed algorithm shows its operating-stability; thus, no unstable numerical calculation such as negative variance calculations and inversion of a singular matrix, which is often the case with algorithms based on expectation propagation, is needed.

V. Conclusion

In this article, we proposed a novel JACDE mechanism based on bilinear inference for grant-free MIMO systems *without spreading data sequences*, while employing non-orthogonal pilot sequences designed via frame theory with the aim of an efficient overhead reduction.

In order to sustain moderate throughput per user, in contrast with most of the grant-free literature, the proposed method is developed based on the conventional MIMO transmission protocol, while employing activity detection capability without resorting to spreading informative data symbols. This is enabled by *bilinear inference and the pseudo-orthogonality of the independently-generated data symbols*. To this end, we derived new Bayesian message passing rules based on Gaussian approximation, which enables JACDE. The proposed scheme is employed in an OFDM cell-free MIMO system. The feasibility of JACDE grant-free access with *non-spread* data sequences is established via simulation-based performance assessment.

[6]Note that since the GaBP algorithm is a generalization of AMP, the resultant error level can be predicted in a similar fashion to the state evolution in AMP. For the sake of consistency, we call the corresponding error predicting quantities as BiGaBP's state evolution for the MSE performances.

Due to the nature of message passing algorithms, the proposed JACDE algorithm is adaptable to channel coding. For such an extension, one can compute the soft outputs based on the messages exchanged across the iterations of the algorithm, which include the information about the estimate and its variance, such as shown in [73], [74]. Although this is an interesting open problem, this extension is beyond the scope of this article and therefore left for a future work.

Appendix A
Proof of Theorem 1

Leveraging the slack variables $t_{\ell,R}$ and $t_{\ell,I}$, the real and imaginary parts of $\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}} \boldsymbol{f}_\ell$ can be respectively bounded as

$$\left|\Re\left\{\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}} \boldsymbol{f}_\ell\right\}\right| \le t_{\ell,R}\cdot \mathbf{1}_{L-1} \quad \text{and} \quad \left|\Im\left\{\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}} \boldsymbol{f}_\ell\right\}\right| \le t_{\ell,I}\cdot \mathbf{1}_{L-1}, \tag{39}$$

where the inequality is applied in an element-by-element manner.

From equation (39), one can readily obtain $\|\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}} \boldsymbol{f}_\ell\|_\infty \le \sqrt{t_{\ell,R}^2 + t_{\ell,I}^2}$. Furthermore, equation (39) can also be rewritten as

$$\underbrace{\begin{cases} \Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Re\{\boldsymbol{f}_\ell\} + \Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Im\{\boldsymbol{f}_\ell\} - t_{\ell,R}\cdot\mathbf{1}_{L-1} \le \mathbf{0} \\ -\left(\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Re\{\boldsymbol{f}_\ell\} + \Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Im\{\boldsymbol{f}_\ell\}\right) - t_{\ell,R}\cdot\mathbf{1}_{L-1} \le \mathbf{0} \end{cases}}_{\Leftrightarrow \left|\Re\left\{\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}} \boldsymbol{f}_\ell\right\}\right| - t_{\ell,R}\cdot\mathbf{1}_{L-1} \le \mathbf{0}} \tag{40a}$$

$$\underbrace{\begin{cases} \Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Im\{\boldsymbol{f}_\ell\} - \Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Re\{\boldsymbol{f}_\ell\} - t_{\ell,I}\cdot\mathbf{1}_{L-1} \le \mathbf{0} \\ -\left(\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Im\{\boldsymbol{f}_\ell\} - \Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}}\Re\{\boldsymbol{f}_\ell\}\right) - t_{\ell,I}\cdot\mathbf{1}_{L-1} \le \mathbf{0} \end{cases}}_{\Leftrightarrow \left|\Im\left\{\tilde{\boldsymbol{F}}_\ell^{\mathrm{H}} \boldsymbol{f}_\ell\right\}\right| - t_{\ell,I}\cdot\mathbf{1}_{L-1} \le \mathbf{0}} \tag{40b}$$

leading to

$$\underbrace{\begin{bmatrix} \Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}} & \Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}} & -\mathbf{1}_{(L-1)\times 1} & \mathbf{0}_{(L-1)\times 1} \end{bmatrix}}_{\triangleq \boldsymbol{A}_{\ell,R,1}} \boldsymbol{x}_\ell \le \mathbf{0} \tag{41a}$$

$$\underbrace{\begin{bmatrix} -\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}} & -\Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}} & -\mathbf{1}_{(L-1)\times 1} & \mathbf{0}_{(L-1)\times 1} \end{bmatrix}}_{\triangleq \boldsymbol{A}_{\ell,R,2}} \boldsymbol{x}_\ell \le \mathbf{0} \tag{41b}$$

$$\underbrace{\begin{bmatrix} -\Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}} & \Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}} & \mathbf{0}_{(L-1)\times 1} & -\mathbf{1}_{(L-1)\times 1} \end{bmatrix}}_{\triangleq \boldsymbol{A}_{\ell,I,1}} \boldsymbol{x}_\ell \le \mathbf{0} \tag{41c}$$

$$\underbrace{\begin{bmatrix} \Im\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}} & -\Re\{\tilde{\boldsymbol{F}}_\ell\}^{\mathrm{T}} & \mathbf{0}_{(L-1)\times 1} & -\mathbf{1}_{(L-1)\times 1} \end{bmatrix}}_{\triangleq \boldsymbol{A}_{\ell,I,2}} \boldsymbol{x}_\ell \le \mathbf{0} \tag{41d}$$

where $\boldsymbol{x}_\ell$ is defined in Theorem 1, equation (11f) can be readily obtained from equation (9b), and this completes the proof.

APPENDIX B
DERIVATION OF EQUATION (24) AND (25)

Given equation (18) and (21b), the effective PDF can be readily expressed as

$$\begin{aligned}
&p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}_m)p_{\boldsymbol{h}_m}(\boldsymbol{h}_m)\\
&= p_{\boldsymbol{h}_m}(\boldsymbol{h}_m)\mathcal{CN}_N\left(\boldsymbol{\mu}^h_{k,m},\boldsymbol{\Sigma}^h_{k,m}\right)\\
&= \left[\lambda\,\mathcal{CN}_N\left(0,\boldsymbol{\Gamma}_m\right)+(1-\lambda)\delta(\boldsymbol{h}_m)\right]\mathcal{CN}_N\left(\boldsymbol{\mu}^h_{k,m},\boldsymbol{\Sigma}^h_{k,m}\right)\\
&= \Bigg[\frac{\lambda\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Gamma}_m+\boldsymbol{\Sigma}^h_{k,m}|}\\
&\quad\cdot\mathcal{CN}_N\left(\left(\boldsymbol{\Sigma}^{h\,-1}_{k,m}+\boldsymbol{\Gamma}^{-1}_m\right)^{-1}\boldsymbol{\Sigma}^{h\,-1}_{k,m}\boldsymbol{\mu}^h_{k,m},\left(\boldsymbol{\Sigma}^{h\,-1}_{k,m}+\boldsymbol{\Gamma}^{-1}_m\right)^{-1}\right)\\
&\quad+\frac{(1-\lambda)\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\boldsymbol{\Sigma}^{h\,-1}_{k,m}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Sigma}^h_{k,m}|}\delta(\boldsymbol{h}_m)\Bigg] \qquad (42)
\end{aligned}$$

where by the Woodbury inverse lemma we obtained $\left(\boldsymbol{\Sigma}^{h\,-1}_{k,m}-\boldsymbol{\Sigma}^{h\,-1}_{k,m}\left(\boldsymbol{\Sigma}^{h\,-1}_{k,m}+\boldsymbol{\Gamma}^{-1}_m\right)^{-1}\boldsymbol{\Sigma}^{h\,-1}_{k,m}\right)=\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}$. Recalling that $\left(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right)^{-1}=\boldsymbol{B}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{A}$ for inversible $\boldsymbol{A}$ and $\boldsymbol{B}$, one may readily obtain equation (24) from (42). This completes the derivation of (24).

Similarly, the normalizing factor $C_{k,m}$ is given by

$$\begin{aligned}
C_{k,m} &\triangleq \int_{\boldsymbol{h}'_m} p_{\boldsymbol{l}^h_{k,m}|\boldsymbol{h}'_m}(\boldsymbol{l}^h_{k,m}|\boldsymbol{h}'_m)p_{\boldsymbol{h}_m}(\boldsymbol{h}'_m)\\
&= \frac{\lambda\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Gamma}_m+\boldsymbol{\Sigma}^h_{k,m}|}\\
&\quad+\frac{(1-\lambda)\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\boldsymbol{\Sigma}^{h\,-1}_{k,m}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Sigma}^h_{k,m}|}\delta(\boldsymbol{h}_m)\\
&= \frac{\lambda\exp\left(-\boldsymbol{\mu}^{h\,\mathrm{H}}_{k,m}\left(\boldsymbol{\Sigma}^h_{k,m}+\boldsymbol{\Gamma}_m\right)^{-1}\boldsymbol{\mu}^h_{k,m}\right)}{\pi^N|\boldsymbol{\Gamma}_m+\boldsymbol{\Sigma}^h_{k,m}|}\\
&\quad\cdot\left(1+\frac{1-\lambda}{\lambda}\left|\boldsymbol{\Sigma}^{h\,-1}_{k,m}\boldsymbol{\Gamma}_m+\mathbf{I}_N\right|\exp\left(-\pi^h_{k,m}\right)\right), \qquad (43)
\end{aligned}$$

which completes the derivation of equation (25).

REFERENCES

[1] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[2] A. Decurninge, L. G. Ordóñez, and M. Guillaud, "Covariance-aided CSI acquisition with non-orthogonal pilots in massive MIMO: A large-system performance analysis," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4489–4512, Jul. 2020.

[3] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing—Part I: Derivation," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5839–5853, Nov. 2014.

[4] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing—Part II: Applications," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5854–5867, Nov. 2014.

[5] Y. Kabashima, F. Krzakala, M. Mezard, A. Sakata, and L. Zdeborova, "Phase transitions and sample complexity in Bayes-optimal matrix factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4228–4265, Jul. 2016, doi: 10.1109/TIT.2016.2556702.

[6] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborova, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 2021–2025.

[7] K. Ito, T. Takahashi, S. Ibi, and S. Sampei, "Bilinear Gaussian belief propagation for large MIMO channel and data estimation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, Dec. 2020, pp. 1–6.

[8] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," *J. Phys. A, Math. Gen.*, vol. 36, no. 43, pp. 11111–11121, Oct. 2003.

[9] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-free radio access for short-packet communications over 5G networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–7.

[10] 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 Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[11] Y. Du *et al.*, "Joint channel estimation and multiuser detection for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 682–685, Aug. 2018.

[12] T. Hara and K. Ishibashi, "Grant-free non-orthogonal multiple access with multiple-antenna base station and its efficient receiver design," *IEEE Access*, vol. 7, pp. 175717–175726, 2019.

[13] S. Jiang, X. Yuan, X. Wang, C. Xu, and W. Yu, "Joint user identification, channel estimation, and signal detection for grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6960–6976, Oct. 2020.

[14] 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 Trans. Commun.*, vol. 68, no. 5, pp. 2963–2977, May 2020.

[15] 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 Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8643–8656, Aug. 2020.

[16] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver," 2019, *arXiv:1910.11266*. [Online]. Available: http://arxiv.org/abs/1910.11266

[17] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Grant-free massive random access with a massive MIMO receiver," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput. (CSSC)*, Nov. 2019, pp. 23–30.

[18] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[19] Z. Chen, F. Sohrabi, and W. Yu, "Multi-cell sparse activity detection for massive random access: Massive MIMO versus cooperative MIMO," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4060–4074, Aug. 2019.

[20] 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, Jun. 2018.

[21] X. Shao, X. Chen, and R. Jia, "A dimension reduction-based joint activity detection and channel estimation algorithm for massive access," *IEEE Trans. Signal Process.*, vol. 68, pp. 420–435, 2020.

[22] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[23] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.

[24] O. Kuybeda, G. A. Frank, A. Bartesaghi, M. Borgnia, S. Subramaniam, and G. Sapiro, "A collaborative framework for 3D alignment and classification of heterogeneous subvolumes in cryo-electron tomography," *J. Struct. Biol.*, vol. 181, no. 2, pp. 116–127, Feb. 2013.

[25] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Math. Program. Comput.*, vol. 4, no. 4, pp. 333–361, Dec. 2012.

[26] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, Aug. 2012.

[27] J. T. Parker and P. Schniter, "Parametric bilinear generalized approximate message passing," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 795–808, Jun. 2016.

[28] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[29] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and message passing algorithms," *Ann. Appl. Probab.*, vol. 25, no. 2, pp. 753–822, Apr. 2015, doi: 10.1214/14-AAP1010.

[30] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 6664–6684, Oct. 2019.

[31] M. Opper and O. Winther, "Expectation consistent approximate inference," *J. Mach. Learn. Res.*, vol. 6, pp. 2177–2204, Dec. 2005. [Online]. Available: http://localhost/pubdb/p.php?3460

[32] T. P. Minka, "Expectation propagation for approximate Bayesian inference," *Comput. Res. Repository (CoRR)*, vol. abs/1301.2294, pp. 362–369, Jan. 2013. [Online]. Available: http://arxiv.org/abs/1301.2294

[33] K. Takeuchi and C.-K. Wen, "Rigorous dynamics of expectation-propagation signal detection via the conjugate gradient method," in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sapporo, Japan, Jul. 2017, pp. 1–5.

[34] X. Meng, S. Wu, and J. Zhu, "A unified Bayesian inference framework for generalized linear models," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 398–402, Mar. 2018.

[35] Q. Zou, H. Zhang, C.-K. Wen, S. Jin, and R. Yu, "Concise derivation for generalized approximate message passing using expectation propagation," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1835–1839, Dec. 2018.

[36] L. Liu, Y. Li, C. Huang, C. Yuen, and Y. L. Guan, "A new insight into GAMP and AMP," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8264–8269, Aug. 2019.

[37] A. K. Fletcher, P. Pandit, S. Rangan, S. Sarkar, and P. Schniter, "Plug in estimation in high dimensional linear inverse problems a rigorous analysis," *J. Stat. Mechanics: Theory Exp.*, vol. 2019, no. 12, Dec. 2019, Art. no. 124021, doi: 10.1088/1742-5468/ab321a.

[38] X. Meng and J. Zhu, "Bilinear adaptive generalized vector approximate message passing," *IEEE Access*, vol. 7, pp. 4807–4815, 2019.

[39] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2523–2527.

[40] V. K. Amalladinne, J.-F. Chamberland, and K. R. Narayanan, "A coded compressed sensing scheme for uncoordinated multiple access," 2018, *arXiv:1809.04745*. [Online]. Available: http://arxiv.org/abs/1809.04745

[41] J. Dong, J. Zhang, and Y. Shi, "Bandit sampling for faster activity and data detection in massive random access," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 8319–8323.

[42] Y. Wu, X. Gao, S. Zhou, W. Yang, Y. Polyanskiy, and G. Caire, "Massive access for future wireless communication systems," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 148–156, Aug. 2020.

[43] A. Munari, F. Clazzer, O. Simeone, and Z. Utkovski, "Grant-free access for IoT in beyond-5G systems: The potential of receiver diversity," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, Levi, Finland, Mar. 2020, pp. 1–5.

[44] Z. Chen, F. Sohrabi, Y.-F. Liu, and W. Yu, "Covariance based joint activity and data detection for massive random access with massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–6.

[45] Z. Chen and W. Yu, "Phase transition analysis for covariance based massive random access with massive MIMO," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019, pp. 36–40.

[46] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, Apr. 2018.

[47] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764–779, 2020.

[48] X. Shao, X. Chen, C. Zhong, J. Zhao, and Z. Zhang, "A unified design of massive access for cellular Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3927–3934, Apr. 2019.

[49] K. Senel and E. G. Larsson, "Device activity and embedded information bit detection using AMP in massive MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Singapore, Dec. 2017, pp. 1–6.

[50] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, "Structured massive access for scalable cell-free massive MIMO systems," 2020, *arXiv:2006.10275*. [Online]. Available: http://arxiv.org/abs/2006.10275

[51] J. Zhang, Y. Wei, E. Björnson, Y. Han, and S. Jin, "Performance analysis and power control of cell-free massive MIMO systems with hardware impairments," *IEEE Access*, vol. 6, pp. 55302–55314, 2018.

[52] *NR; User Equipment (UE) Radio Transmission and Reception*, document 3GPP, TS 38.101–1, V15.3.0, Sep. 2018.

[53] B. K. Jeong, B. Shim, and K. B. Lee, "MAP-based active user and data detection for massive machine-type communications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8481–8494, Sep. 2018.

[54] T. Ding, X. Yuan, and S. C. Liew, "Sparsity learning-based multiuser detection in grant-free massive-device multiple access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3569–3582, Jul. 2019.

[55] C. Rusu and N. González-Prelcic, "Designing incoherent frames through convex techniques for optimized compressed sensing," *IEEE Trans. Signal Process.*, vol. 64, no. 9, pp. 2334–2344, May 2016.

[56] C. Rusu, N. González-Prelcic, and R. W. Heath, Jr., "Algorithms for the construction of incoherent frames under various design constraints," *Signal Process.*, vol. 152, pp. 363–372, Nov. 2018.

[57] R.-A. Stoica, G. T. F. D. Abreu, and H. Iimori, "A frame-theoretic scheme for robust millimeter wave channel estimation," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Kansas City, MI, USA, Aug. 2018, pp. 1–6.

[58] R.-A. Stoica, G. T. F. de Abreu, T. Hara, and K. Ishibashi, "Massively concurrent non-orthogonal multiple access for 5G networks and beyond," *IEEE Access*, vol. 7, pp. 82080–82100, 2019.

[59] T. Strohmer and R. W. Heath, Jr., "Grassmannian frames with applications to coding and communication," *Appl. Comput. Harmon. Anal.*, vol. 14, no. 3, pp. 257–275, May 2003.

[60] K.-H. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "Cube-split: A structured Grassmannian constellation for non-coherent SIMO communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1948–1964, Mar. 2020.

[61] M. Thill and B. Hassibi, "Group frames with few distinct inner products and low coherence," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5222–5237, Oct. 2015.

[62] J. Mattingley and S. Boyd, "Real-time convex optimization in signal processing," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 50–61, May 2010.

[63] J. A. Tropp, I. S. Dhillon, R. W. Heath, Jr., and T. Strohmer, "Designing structured tight frames via an alternating projection method," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 188–209, Jan. 2005.

[64] I. Kammoun, A. M. Cipriano, and J.-C. Belfiore, "Non-coherent codes over the Grassmannian," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3657–3667, Oct. 2007.

[65] R. H. Gohary and T. N. Davidson, "Noncoherent MIMO communication: Grassmannian constellations and efficient detection," *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 1176–1205, Mar. 2009.

[66] I. S. Dhillon, R. W. Heath, Jr., T. Strohmer, and J. A. Tropp, "Constructing packings in Grassmannian manifolds via alternating projection," *Experim. Math.*, vol. 17, no. 1, pp. 9–35, Jan. 2008.

[67] T. Takahashi, S. Ibi, and S. Sampei, "Design of adaptively scaled belief in multi-dimensional signal detection for higher-order modulation," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 1986–2001, Mar. 2019.

[68] A. Chockalingam and B. S. Rajan, *Large MIMO Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[69] E. Bjornson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[70] P. Joshi, D. Colombi, B. Thors, L.-E. Larsson, and C. Törnevik, "Output power levels of 4G user equipment and implications on realistic RF EMF exposure assessments," *IEEE Access*, vol. 5, pp. 4545–4550, 2017.

[71] D. Shen, Z. Pan, K.-K. Wong, and V. O. K. Li, "Effective throughput: A unified benchmark for pilot-aided OFDM/SDMA wireless communication systems," in *Proc. 22nd Annu. Joint Conf. IEEE Comput. Commun. Societies (IEEE INFOCOM)*, San Francisco, CA, USA, Mar. 2003, pp. 1603–1613.

[72] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 264–273, Feb. 2013.

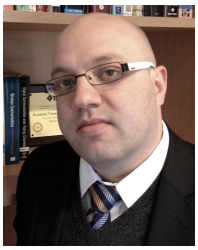
[73] J. Céspedes, P. M. Olmos, M. Sánchez-Fernández, and F. Perez-Cruz, "Probabilistic MIMO symbol detection with expectation consistency approximate inference," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3481–3494, Apr. 2018.

[74] K.-H. Ngo, M. Guillaud, A. Decurninge, S. Yang, and P. Schniter, "Multi-user detection based on expectation propagation for the non-coherent SIMO multiple access channel," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 6145–6161, Sep. 2020.

Hiroki Iimori (Graduate Student Member, IEEE) received the B.Eng. degree in electrical and electronic engineering and the M.Eng. degree (Hons.) in advanced electrical, electronic and computer systems from Ritsumeikan University, Kyoto, Japan, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Jacobs University Bremen, Germany. He was a Visiting Scholar with the Electrical and Computer Engineering Department, University of Toronto, Toronto, ON, Canada. In 2021, he was a Research Intern with the Ericsson Radio S&R Research Laboratory, Yokohama, Japan. His research interests include optimization theory, wireless communications, and signal processing. He was a recipient of the YKK Doctoral Fellowship awarded by the Yoshida Scholarship Foundation, Japan.

Takumi Takahashi (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in communication engineering from Osaka University, Osaka, Japan, in 2016, 2017, and 2019, respectively. From 2018 to 2019, he was a Visiting Researcher with the Centre for Wireless Communications, University of Oulu, Finland. In 2019, he joined the Graduate School of Engineering, Osaka University, as an Assistant Professor. His current research interests include belief propagation, compressed sensing, signal processing, and wireless communications.

Koji Ishibashi (Senior Member, IEEE) received the B.E. and M.E. degrees in engineering from The University of Electro-Communications, Tokyo, Japan, in 2002 and 2004, respectively, and the Ph.D. degree in engineering from Yokohama National University, Yokohama, Japan, in 2007. From 2007 to 2012, he was an Assistant Professor with the Department of Electrical and Electronic Engineering, Shizuoka University, Hamamatsu, Japan. From 2010 to 2012, he was a Visiting Scholar with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. Since April 2012, he has been with the Advanced Wireless and Communication Research Center (AWCC), The University of Electro-Communications, where he is currently a Professor. His current research interests include grant-free access, cell-free architecture, millimeter-wave communications, energy harvesting communications, wireless power transfer, channel codes, signal processing, and information theory.



Giuseppe Thadeu Freitas de Abreu (Senior Member, IEEE) received the B.Eng. degree in electrical engineering and the Latu Sensu degree in telecommunications engineering from the Universidade Federal da Bahia (UFBa), Salvador, Brazil, in 1996 and 1997, respectively, and the M.Eng. and D.Eng. degrees in physics, electrical and computer engineering from Yokohama National University, Japan, in March 2001 and March 2004, respectively. He was a Post-Doctoral Fellow and an later Adjunct Professor (Docent) on statistical signal processing and communications theory with the Department of Electrical and Information Engineering, University of Oulu, Finland, from 2004 to 2006 and from 2006 to 2011, respectively. Since 2011, he has been a Professor of electrical engineering with Jacobs University Bremen, Germany. From April 2015 to August 2018, he also held simultaneously a full professorship with the Department of Computer and Electrical Engineering, Ritsumeikan University, Japan. His research interests include communications and signal processing, including communications theory, estimation theory, statistical modeling, wireless localization, cognitive radio, wireless security, MIMO systems, ultra-wideband and millimeter wave communications, full-duplex and cognitive radio, compressive sensing, energy harvesting networks, random networks, connected vehicles networks, and many other topics. He received the Uenohara Award by Tokyo University in 2000 for his Master's Thesis work. He was a co-recipient of the best paper awards at several international conferences. He was also awarded the prestigious JSPS, Heiwa Nakajima, and NICT Fellowships in 2010, 2013, and 2015, respectively. He served as an Associate Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2009 to 2014, IEEE TRANSACTIONS ON COMMUNICATIONS from 2014 to 2017, and currently serves as an Executive Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Wei Yu (Fellow, IEEE) received the B.A.Sc. degree in computer engineering and mathematics from the University of Waterloo, Waterloo, ON, Canada, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1998 and 2002, respectively. Since 2002, he has been with the Electrical and Computer Engineering Department, University of Toronto, Toronto, ON, where he is currently a Professor and holds a Canada Research Chair (Tier 1) in Information Theory and Wireless Communications. He is a Fellow of the Canadian Academy of Engineering and a member of the College of New Scholars, Artists and Scientists of the Royal Society of Canada. He received the Steacie Memorial Fellowship in 2015, the IEEE Marconi Prize Paper Award in Wireless Communications in 2019, the IEEE Communications Society Award for Advances in Communication in 2019, the IEEE Signal Processing Society Best Paper Award in 2017 and 2008, the IEEE JOURNAL OF COMMUNICATIONS AND NETWORKS Best Paper Award in 2017, and the IEEE Communications Sociesty Best Tutorial Paper Award in 2015. Prof. Yu is the President of the IEEE Information Theory Society in 2021. He served as the Chair for the Signal Processing for Communications and Networking Technical Committee and IEEE Signal Processing Society from 2017 to 2018. He was an IEEE Communications Society Distinguished Lecturer from 2015 to 2016. He served as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION THEORY, IEEE TRANSACTIONS ON COMMUNICATIONS, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He is currently an Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.