Energy-Efficient and Low-Complexity Uplink Transceiver for Massive Spatial Modulation MIMO

Shengchu Wang, Yunzhou Li, *Member, IEEE*, Ming Zhao, *Member, IEEE*, and Jing Wang, *Member, IEEE*

***Abstract*—In this paper, we design the uplink transceiver for a new multiuser (MU) massive spatial modulation (SM) multiple-input–multiple-output (SM-MIMO) system over frequency-selective fading channels, where the base station (BS) is equipped with massive antennas, and user equipment (UE) has multiple transmit antennas (TAs) but only one radio-frequency (RF) chain. UE transmit data bits to the BS simultaneously by cyclic-prefix single-carrier (CP-SC) SM. For the uplink MU detection (MUD), we construct a low-complexity generalized approximate message passing detector (GAMPD), which can exploit both the sparsity and prior probability distribution of the transmitted signal and is suitable for hardware implementation because its most complex operation is only matrix–vector multiplication. Its mean square error (MSE) and uncoded bit error rate (BER) performances are also analyzed based on the state evolution (SE). Compared with stagewised linear detectors, GAMPD shows orders-of-magnitude lower complexity. Moreover, simulation results indicated that GAMPD approaches to the performance of maximum-likelihood (ML) detection and outperforms minimum MSE (MMSE) significantly. Finally, to design energy-efficient massive SM-MIMO, we propose a practical algorithm to optimize the key system parameters (e.g., the transmission power, the numbers of the BS antennas or UE, or the TAs at the UE). Numerical results indicate that low BS circuit power consumption and long channel coherence time are the two key prerequisites for the success of massive SM-MIMO.**

***Index Terms*—Approximated message passing, energy efficient, massive multiple-input–multiple-output (MIMO), multiuser detection (MUD), spatial modulation (SM), state evolution (SE).**

I. Introduction

NEXT-GENERATION cellular communication systems could adopt massive antennas at the base station (BS) to serve much user equipment (UE) simultaneously [1]–[5]. Through harvesting extremely high multiple-input–multiple-output (MIMO) diversity and multiplexing gains [6], the rate, reliability, and power efficiency of the uplink transmission can be significantly improved [1], [2], [7]. In classical massive MIMO [1], [2], [8], [9], UE are assumed to have a single transmit antenna (TA) connecting to a single radio-frequency (RF) chain. Massive spatial modulation (SM) MIMO (SM-MIMO) was introduced in [5], [12] where different UE still had one RF chain but multiple TAs [12]–[14]. At every transmission instance, one TA is activated to connect to the exclusive RF chain. Until now, not only the quadrature amplitude modulation (QAM) symbols but also the index of the active TA (spatial constellation [12]–[14]) help transmit information to the BS. Consequently, the uplink spectral efficiency is improved without sacrificing the advantages of single RF [5], [10], [11],

However, the possibility but not the implementation details of massive SM-MIMO was only pointed out in [5]. SM-MIMO was researched in [12] and [47] over flat-fading channels. Unfortunately, based on the channel measurements [15], [16], the channel delay spread is usually larger than the baseband sampling time; hence, the frequency-selective case should be researched. In [17], cyclic-prefix single-carrier (CP-SC) SM has been proposed for single-use SM-MIMO over frequency-selective fading channels. In this paper, it is extended into the multiuser (MU) case. Finally, we construct a new MU SC massive SM-MIMO system over frequency-selective fading channels (from then on, massive SM-MIMO indicates this new system specifically).

Until now, there exist two key problems on designing the uplink transceiver for massive SM-MIMO. The first problem is how to design MU detectors. Unfortunately, almost all existing detectors in [5], [12]–[14], [19] lose their effectiveness considering that their computational complexities soar due to the curse of dimensionality of massive SM-MIMO (the numbers of both the BS antennas and UE are huge). For example, maximum likelihood (ML) [14], [47] and sphere decoding [19] are hopeless to be implemented just as their counterparts are hard to be used in massive MIMO [1], [8]. Message passing detector (MPD) [12] is specially designed for the flat-fading case and becomes very cumbersome under the frequency-selective scenario due to its large number of messages. Although the linear detectors have been viewed as the simplest and most effective detectors for conventional SM-MIMO [5], [13], [14], they are still impractical for massive SM-MIMO because they have to construct and then solve large-scale linear equations for every input SC symbol. Therefore, new low-complexity detectors have to be designed for massive SM-MIMO.

Manuscript received May 2, 2014; revised September 17, 2014; accepted November 4, 2014. Date of publication November 24, 2014; date of current version October 13, 2015. This work will be presented in part at GlobeCom 2014 Workshop-Emerging Technologies for 5G Wireless Cellular Networks (GC14 WS-Wi5G), Austin, TX, USA, December 8–12, 2014. This work was supported in part by the National Basic Research Program of China under Grant 2013CB329002, by China's 863 Project under Grant 2012AA011402, by the National Major Project under Grant 2014ZX03003002-002), and by the Program for New Century Excellent Talents in University under Grant NCET-13-0321. The review of this paper was coordinated by Dr. S. Sun.

S. Wang is with the Department of Electrical Engineering and the Wireless and Mobile Communications R&D Center, Tsinghua University, Beijing 100084, China (e-mail: wsc11@mails.tsinghua.edu.cn).

Y. Li, M. Zhao, and J. Wang are with the Wireless and Mobile Communications R&D Center, Tsinghua University, Beijing 100084, China (e-mail: liyunzhou@tsinghua.edu.cn; zhaoming@tsinghua.edu.cn; wangj@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2014.2373364

In this paper, we solve the MU detection (MUD) problem in massive SM-MIMO by the following two successive stages. The first stage is channel estimation (CE). There, different UE alternatively activate their TAs to transmit their own training sequences, which are generated by cyclic-shifting a common sequence with different fixed steps. Then the channel responses between the BS and UE are separated out based on the orthogonality in the time domain. The second stage is signal detection. After absorbing the effects of the CE errors by Gaussian noise, we solve the MUD problem under the framework of generalized approximate message passing (GAMP) [18] and construct a new GAMP detector (GAMPD). Note that GAMP is a powerful algorithm for generalized linear mixing problems [18] and has been applied to massive MIMO recently [33], [48]. In this paper, it is applied to massive SM-MIMO. Based on [18], [21], [22], and [33], we construct a state evolution (SE) equation to analyze and predict the mean square error (MSE) and uncoded bit error rate (BER) performances of GAMPD. Compared with the existing detectors [5], [12]–[14], [19], GAMPD has the following two advantages. First, it calls for parallelized matrix–vector multiplication as its most complex operation. Therefore, it has low computational complexity and is suitable for hardware implementation. Second, it exploits both the sparsity (first pointed out in [20]) and the prior probability distribution function (pdf) of the transmitted signal. Simulation results show that GAMPD outperforms the linear detectors significantly, and its performances are analyzed by SE successfully.

The second problem is how to design energy-efficient massive SM-MIMO with GAMPD. It is key in designing the system parameters (e.g., transmission power, the numbers of the BS antennas or UE, or the TAs at the UE). A new metric is proposed to characterize the energy efficiency (EE) of massive SM-MIMO. Then, it is utilized to construct an optimization problem that involves different system parameters as the optimization variables and some specified uncoded BER as the constraint. Finally, the optimization problem is solved by a practical algorithm that is developed based on the SE for GAMPD. Experiments' results indicate that the success of massive SM-MIMO relies on both the long-enough duration of the channel coherence time and the low-enough circuit power consumption at the BS.

The rest of this paper is organized as follows. Section II introduces the system model of massive SM-MIMO. Section III describes the MUD problem whose preparation work, i.e., CE, is also introduced. Then the MUD problem is solved by stagewised linear detectors and the new proposed GAMPD in Sections IV and V, respectively. Section VI discusses how to design energy-efficient massive SM-MIMO through optimizing the key system parameters. Finally, simulation results are presented in Section VII, and this paper is concluded in Section VIII.

Notation: $[N]=0,1,\ldots,N-1$ for any integer N. $i=\sqrt{-1}$. For a vector $\mathbf{a}$, a_k is the kth element, $\|\mathbf{a}\|_p=(\sum_{k=0}^{N-1}|a_k|^p)^{1/p}$ is the ℓ_p-norm, and $\mathbf{a}(\Omega)$ or $\mathbf{a}_\Omega$ is a subvector composed by elements $\{a_k\}_{k\in\Omega}$, where Ω is an integer set. For a matrix $\mathbf{A}$, A_{mn} is the (m,n)th element, $\mathbf{A}_k$ is the kth column, and $\mathbf{A}_\Omega$ is a submatrix comprising the columns $\{\mathbf{A}_k\}_{k\in\Omega}$. $m:n$ indicate the integers increasing from m to n. Without special notice, a vector is always columnwised. $\circledast$ indicates the circular convolution, and $\otimes$ means the Kronecker product. $\mathbf{I}_N$ is the identity matrix with dimension $N\times N$. $\mathbf{0}_N$ is the all-zero vector with dimension N. For a matrix or vector, $()^*$ indicates the conjugate operation, $(\cdot)^T$ means the transpose, and $(\cdot)^H$ or $(\cdot)^\dagger$ is the complex conjugate transpose.

II. System Model

Here, Part A shows the uplink configuration of massive SM-MIMO over frequency-selective fading channels. Part B introduces the channel training sequence design for different UE. Finally, Part C mathematically formulates the procedures of the data transmission at UE and signal reception at the BS.

A. Configuration of the Massive SM-MIMO

Fig. 1 shows the uplink architecture of massive SM-MIMO. At the transmitter side, there exist U UE. Every UE has N_t TAs but only one RF chain. At the receiver side, the BS is equipped with N_r antennas connecting to independent RF chains (e.g., ARGO design [8]). $\sqrt{\rho}\mathbf{h}^u_{n_rn_t}\in\mathbb{C}^L$ is the multipath channel response between the n_rth BS antenna and the n_tth TA of the uth user. ρ is the channel fading coefficient, which is the same for all the UE. L is the number of channel taps. Similar to [23] and [24], we consider the Rayleigh channel model where $\mathbf{h}^u_{n_rn_t}\sim\mathcal{CN}(\mathbf{0}_L,1/L\mathbf{I}_L)$ with $u\in[U]$, $n_r\in[N_r]$, and $n_t\in[N_t]$.

B. Channel Training Sequences for Different UE

As [1], [9], and [25], the coherence time interval is divided in two parts, where the first part is reserved for channel training, and the second part is used for data transmission. Fig. 2 shows an example of the frame structure for the uth UE (all UE share similar frame structures).

Next, we discuss the channel training sequence design for the U UE. Every UE has one RF chain; hence, it has to activate the N_t TAs alternatively [26], [27]. Consequently, the time for channel training is further divided as N_t training periods (see Fig. 2). During the n_tth training period, all the UE activate their own n_tth TAs to transmit their own training sequences shown in Fig. 3.

Without loss of generality, the uth UE is focused on. Its training sequence is composed of the data part and the CP part. The data part is generated by cyclically right-shifting the common sequence

$$\mathbf{t}=[t_0,t_1,\ldots,t_{L-1},t_L,\ldots,t_{2L-1},\ldots,t_{K-L},\ldots,t_{K-1}]$$

with uL steps. The new shifted sequence is denoted as $\mathbf{t}\circledast\delta(n-uL)$. Its last L symbols are repeated as the CP part. The length of $\mathbf{t}$ satisfies $K\geq UL$. Similarly, we design the training sequences for all the other UE.

This sequence design is based on the previous work on the MU CE in the uplink of Long-Term Evolution [28]. How it helps the BS accomplish the CE will be introduced in Section III.

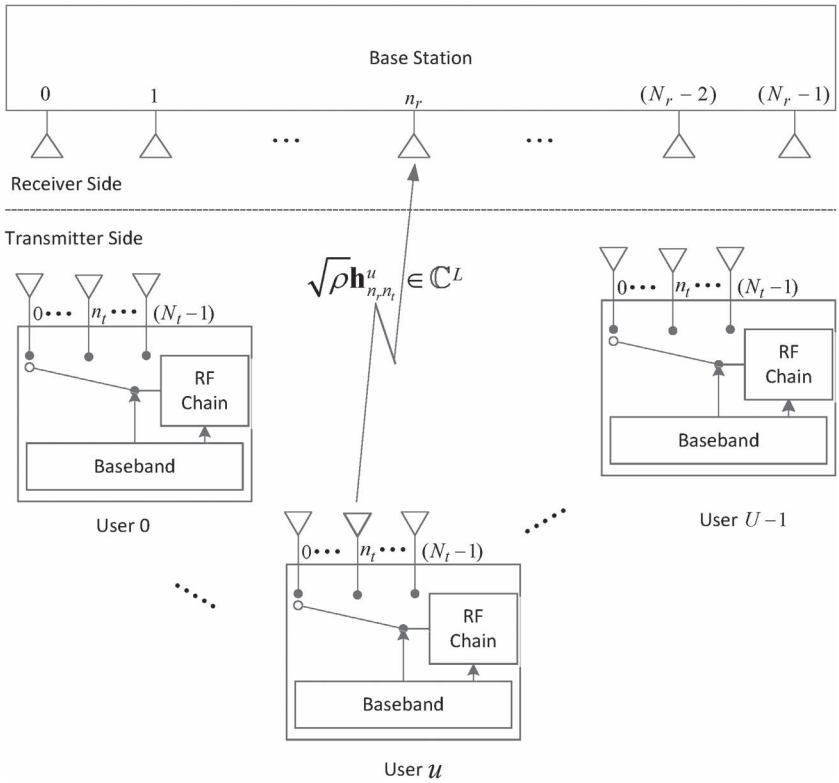


Fig. 1. Uplink system configuration of massive SM-MIMO over the frequency-selective fading channels.

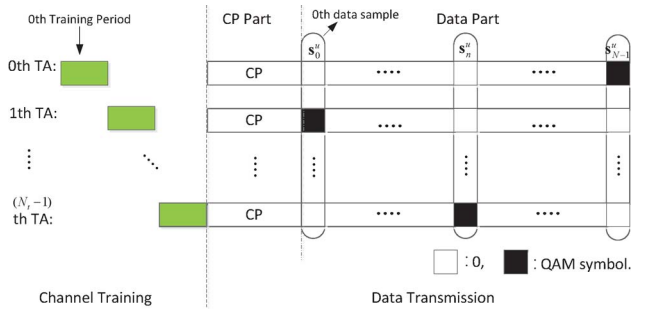


Fig. 2. Data frames for the uth UE; at the stage of channel training, N_t TAs are alternatively activated to transmit the training sequence specified for the uth UE; at the stage of data transmission, TAs transmit their own SC symbols (only one symbol is shown here for clarity).

C. Data Transceiving in Massive SM-MIMO

Once the channel training is completed, the data transmission will be started (see Fig. 2). Here, we mathematically formulate the procedures of the data transmission at the UE and data reception at the BS.

In the uplink, UE transmit data to the BS by the CP-SC SM technique [17]. In this paper, we assume that all UE adopt the same uplink transmission power p_s and the same QAM type M-QAM. The M-QAM alphabet is defined as

$$\mathbb{A} = A \times \left\{ \pm(2k_R+1) \pm i(2k_I+1), k_R, k_I \in \left[\sqrt{M}/2 \right] \right\} \quad (1)$$

where M is the QAM order, and A is the power normalization factor (e.g., $1/\sqrt{10}$ for 16-QAM).

Because additive noise is considered, we first focus on the noiseless transmission at the uth user during one SC symbol. Fig. 2 shows an example of the SC symbols transmitted by the N_t TAs of the uth UE. There, one SC symbol is composed of the CP part and the data part. The data part contains N data samples, whose last L symbols are repeated as CP. From then on, we only focus on how to formulate the data part.

Based on the principles of SC-SM [17], during one SC symbol, the uth UE can transmit $N \log_2(N_t M)$ bits, which are denoted as

$$\mathbf{b}^u = \left\{ \mathbf{b}^u_0, \mathbf{b}^u_1, \ldots, \mathbf{b}^u_{N-1} \right\} \quad (2)$$

where $\mathbf{b}^u_n \in \{0,1\}^{\log_2(N_t M)}$ with $n \in [N]$ as a binary subvector. During the nth data sample, $\mathbf{b}^u_n$ is extracted from $\mathbf{b}^u$. Then, based on the principles of SM [5], [13], [14], [17], its first $\log_2 N_t$ bits determine the index of the active TA (denoted as j^u_n), and its remaining $\log_2 M$ bits are mapped as a QAM symbol (denoted as $q^u_n \in \mathbb{A}$) transmitted by the active TA. In summary, $\mathbf{b}^u_n$ is mapped as

$$\mathbf{s}^u_n = \begin{bmatrix} 0 \cdots 0 \; \underset{\uparrow}{q^u_n} \; 0 \cdots 0 \\ {\scriptstyle j^u_n} \end{bmatrix}^T_{N_t \times 1} \quad (3)$$

which is the signal vector transmitted by the uth UE at the nth data sample (silent TAs are assumed to transmit 0 s).

After adding CP, the SC symbol transmitted by the n_tth TA is

$$\sqrt{p_s} \left[\underbrace{\mathbf{s}^u_{N-L}(n_t) \ldots \mathbf{s}^u_{N-1}(n_t)}_{\text{CP}} \vdots \underbrace{\mathbf{s}^u_0(n_t) \, \mathbf{s}^u_1(n_t) \ldots \mathbf{s}^u_{N-1}(n_t)}_{(\mathbf{x}^u_{n_t})^T} \right] \quad (4)$$

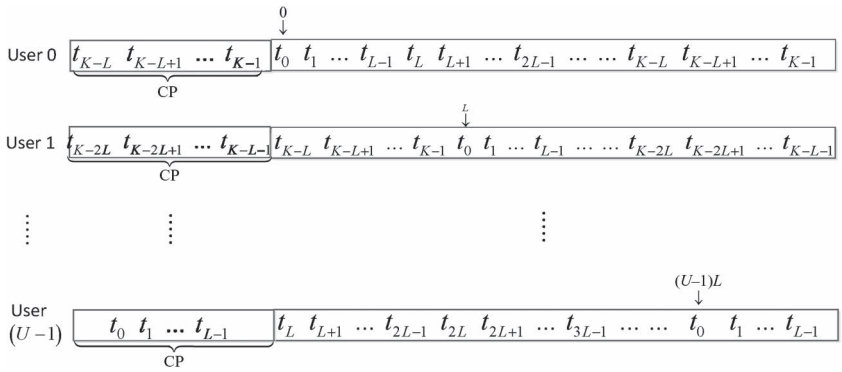


Fig. 3. Channel training sequences for the U UE.

whose data part is $\mathbf{x}_{n_t}^u \in \mathbb{C}^N$.

At the receiver side, after removing the CP, the received signal vector at the n_rth BS antenna is

$$\mathbf{y}_{n_r}^u = \sqrt{p_s}\sqrt{\rho}\mathbf{H}_{n_r}^u \mathbf{x}^u \tag{5}$$

where $\mathbf{x}^u = [(\mathbf{x}_0^u)^T \ (\mathbf{x}_1^u)^T \dots (\mathbf{x}_{(N_t-1)}^u)^T]^T$, and $\mathbf{H}_{n_r}^u = [\mathbf{H}_{n_r 0}^u \ \mathbf{H}_{n_r 1}^u \dots \mathbf{H}_{n_r(N_t-1)}^u]$. $\mathbf{H}_{n_r n_t}^u = \text{toep}(\mathbf{h}_{n_r n_t}^u, N)$ (with $n_t \in [N_t]$) is an $N \times N$-sized Toeplitz matrix. For a vector $\mathbf{c} \in \mathbb{C}^L$, $\text{toep}(\mathbf{c}, N)$ is defined as

$$\text{toep}(\mathbf{c}, N) = \begin{bmatrix} c_0 & 0 & \cdots & c_{L-1} & \cdots & \cdots & c_2 & c_1 \\ c_1 & c_0 & 0 & \cdots & \cdots & \cdots & c_3 & c_2 \\ c_2 & c_1 & c_0 & \cdots & \cdots & \cdots & 0 & c_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & c_{L-1} & \cdots & c_0 & 0 \\ 0 & 0 & \cdots & \cdots & 0 & c_{L-1} & \cdots & c_0 \end{bmatrix}. \tag{6}$$

The Toeplitz structure of (6) comes from the fact that the data transmission over the multipath channel is equivalent to a circulant convolution after adding CP at the transmitter and removing CP at the receiver [6].

Based on (3)–(5), $\mathbf{x}^u$ and $\mathbf{s}_n^u$ satisfy

$$\mathbf{x}^u(\Psi_n) = \mathbf{s}_n^u \tag{7}$$

where $\Psi_n = \{n_t N + n\}_{n_t \in [N_t], n \in [N]}$.

Next, we will discuss the MU case. Similar as the uth UE, all the UE will construct their own SC symbols, which are transmitted to the BS simultaneously. Then, the received signal at the n_rth BS antenna is a noisy combination of $\{\mathbf{y}_{n_r}^u\}_{u\in[U]}$. It can be expressed as

$$\mathbf{y}_{n_r} = \sum_{u\in[U]} \mathbf{y}_{n_r}^u + \mathbf{w}_{n_r} = \sqrt{p}_s\sqrt{\rho}\mathbf{H}_{n_r}\mathbf{x} + \mathbf{w}_{n_r} \tag{8}$$

where $\mathbf{H}_{n_r} = [\mathbf{H}_{n_r}^0 \ \mathbf{H}_{n_r}^1 \ \dots \ \mathbf{H}_{n_r}^{U-1}]$, $\mathbf{x} = [(\mathbf{x}^0)^T \ (\mathbf{x}^1)^T \ \dots \ (\mathbf{x}^{U-1})^T]$, and $\mathbf{w}_{n_r} \sim \mathcal{CN}(0, \sigma_{\text{circ}}^2 \mathbf{I}_N)$ is the circuit noise with variance σ_{circ}^2.

By stacking (8) over $n_r \in [N_r]$, the received signal at the BS is

$$\mathbf{y} = \sqrt{p_s}\mathbf{H}\mathbf{x} + \mathbf{w} \tag{9}$$

where $\mathbf{y} = (1/\sqrt{\rho})[\mathbf{y}_0^T \ \mathbf{y}_1^T \ \dots \ \mathbf{y}_{N_r-1}^T]^T$, $\mathbf{H} = [\mathbf{H}_0^T \ \mathbf{H}_1^T \ \dots \ \mathbf{H}_{N_r-1}^T]^T$, and $\mathbf{w} = (1/\sqrt{\rho})[\mathbf{w}_0^T \ \mathbf{w}_1^T \ \dots \ \mathbf{w}_{N_r-1}^T]^T$ is the scaled Gaussian noise with variance $\sigma^2 = \sigma_{\text{circ}}^2/\rho$. Equations (5), (8), and (9) show how $\mathbf{H}$ is constructed from $\{\mathbf{h}_{n_r n_t}^u\}_{n_r\in[N_r],\, n_t\in[N_t]}^{u\in[U]}$. This operation is denoted as $\mathbf{H} = \text{TOEP}(\{\mathbf{h}_{n_r n_t}^u\}_{n_r\in[N_r],\, n_t\in[N_t]}^{u\in[U]})$ where $\text{TOEP}(\cdot)$ is defined as (A.1) in Appendix A. Based on (3), (7), and (8), we know that $\Omega^u = \{j_n^u N + n\}_{n\in[N]}$ contains the positions of the N nonzeros in $\mathbf{x}^u$, and $\Omega = \{uN_t N + \Omega^u\}_{u\in[U]}$ contains the positions of the UN nonzero elements of $\mathbf{x}$.

Finally, the signal-to-noise ratio (SNR) during MUD is defined as $\text{SNR}_{\text{sig}} = 10\lg(Up_s/\sigma^2)$.

III. Channel Estimation and Multiuser Detection in Massive Spatial-Modulation Multiple-Input–Multiple-Output

In this paper, we focus on coherent detection. Therefore, to solve the MUD problem, the BS has to acquire the channel responses between itself and different UE. Without loss of generality, we focus on the n_rth BS antenna and the n_tth training period (see Fig. 2). Based on Section II-B, U users activate their n_tth TAs to transmit their own training sequences shown in Fig. 3. Here, the length of the training sequence is set as $K = UL$, and the transmission power is p_t.

After removing the CP, the received signal at the n_rth BS antenna is

$$\begin{aligned} \mathbf{z}_{n_r} &= \sum_{u\in[U]} \sqrt{p_t}\,(\mathbf{t} \circledast \delta(n - uL)) \circledast \bar{\mathbf{h}}_{n_r n_t}^u + \mathbf{w}_{n_r}^c \\ &= \sqrt{p_t}\mathbf{t} \circledast \underbrace{\left(\sum_{u\in[U]} \bar{\mathbf{h}}_{n_r n_t}^u \circledast \delta(n - uL)\right)}_{\bar{\mathbf{h}}_{n_r n_t}} + \mathbf{w}_{n_r}^c \\ &= \sqrt{p_t}\mathbf{T}\bar{\mathbf{h}}_{n_r n_t} + \mathbf{w}_{n_r}^c \end{aligned} \tag{10}$$

where $\mathbf{t} \circledast \delta(n - uL)$ is the data part of the training sequence for the uth user, n is the discrete time index, $\bar{\mathbf{h}}_{n_r n_t}^u = [(\mathbf{h}_{n_r n_t}^u)^T \ \mathbf{0}_{(U-1)L}^T]^T$, $\bar{\mathbf{h}}_{n_r n_t} = [(\mathbf{h}_{n_r n_t}^0)^T \ (\mathbf{h}_{n_r n_t}^1)^T \ \dots \ (\mathbf{h}_{n_r n_t}^{U-1})^T]^T$, $\mathbf{T} = \text{toep}(\mathbf{t}, K)$, and $\mathbf{w}_{n_r}^c \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_K)$ is the Gaussian noise with variance σ^2.

Based on (10), the least square (LS) estimation of $\bar{\mathbf{h}}_{n_rn_t}$ is

$$\hat{\bar{\mathbf{h}}}_{n_rn_t} = \bar{\mathbf{h}}_{n_rn_t} + \underbrace{\frac{1}{\sqrt{p_t}}(\mathbf{T}^H\mathbf{T})^{-1}\mathbf{T}^H\mathbf{w}^c_{n_r}}_{\bar{\mathbf{e}}_{n_rn_t}} \tag{11}$$

where $\bar{\mathbf{e}}_{n_rn_t} \in \mathbb{C}^{UL}$ is the CE error vector. To minimize $\|\bar{\mathbf{e}}_{n_rn_t}\|_2^2$, $\mathbf{T}$ should satisfy the orthogonality condition $\mathbf{T}^H\mathbf{T} = UL\mathbf{I}_{UL}$ [31], [32]. In this paper, $\mathbf{t}$ is designed as $\mathbf{Wd}$, where $\mathbf{W}$ is the Fourier matrix with $1/\sqrt{UL}e^{i2\pi mn/UL}$ as its (m,n)th element $(m,n \in [N])$, and $\mathbf{d} \in \mathbb{C}^{UL}$ contains UL random quadrature phase-shift keying (QPSK) symbols [defined by (1)]. Consequently, $\bar{\mathbf{e}}_{n_rn_t}$ is distributed as $\mathcal{CN}(\mathbf{0}_{UL}, \sigma^2/(p_tUL)\mathbf{I}_{UL})$.

Based on the structure of $\bar{\mathbf{h}}_{n_rn_t}$, (11) can be decomposed as

$$\hat{\mathbf{h}}^u_{n_rn_t} = \mathbf{h}^u_{n_rn_t} + \mathbf{e}^u_{n_rn_t}, \text{ with } u \in [U] \tag{12}$$

where $\hat{\mathbf{h}}^u_{n_rn_t} = \hat{\bar{\mathbf{h}}}_{n_rn_t}(uL : uL+L-1)$ and $\mathbf{e}^u_{n_rn_t} = \bar{\mathbf{e}}_{n_rn_t}(uL : uL+L-1)$ are the CE and CE errors of $\mathbf{h}^u_{n_rn_t}$, respectively.

Finally, based on (12), $\mathbf{H}$ in (9) is estimated as

$$\hat{\mathbf{H}} = \mathbf{H} + \mathbf{E} \tag{13}$$

where $\hat{\mathbf{H}} = \text{TOEP}(\{\hat{\mathbf{h}}^u_{n_rn_t}\}^{u\in[U]}_{n_r\in[N_r],\ n_t\in[N_t]})$, and $\mathbf{E} = \text{TOEP} \times (\{\mathbf{e}^u_{n_rn_t}\}^{u\in[U]}_{n_r\in[N_r],\ n_t\in[N_t]})$ is the CE error matrix. The SNR during CE is defined as $\text{SNR}_{\text{train}} = 10\lg(Up_t/\sigma^2)$.

Once $\hat{\mathbf{H}}$ is obtained, it will be utilized into MUD as follows. By substituting (13) into (9), we have

$$\mathbf{y} = \sqrt{p_s}\hat{\mathbf{H}}\mathbf{x} \underbrace{-\sqrt{p_s}\mathbf{E}\mathbf{x}}_{\mathbf{w}_{\text{I}}} + \mathbf{w} = \sqrt{p_s}\hat{\mathbf{H}}\mathbf{x} + \mathbf{n} \tag{14}$$

where $\mathbf{w}_{\text{I}} = -\sqrt{p_s}\mathbf{E}\mathbf{x}$ is the interference introduced by the CE error matrix $\mathbf{E}$, and $\mathbf{n} = \mathbf{w}_{\text{I}} + \mathbf{w}$ is the combination of the noise and interference.

The MUD problem is about how to reconstruct $\mathbf{x}$ based on (14). Unfortunately, it is hard to be solved because the unknown interference $\mathbf{w}_{\text{I}}$ is correlated with the uncontaminated measurement vector $\sqrt{p_s}\hat{\mathbf{H}}\mathbf{x}$. Based on the Gaussian pdf of the CE errors $\{\bar{\mathbf{e}}^u_{n_rn_t}\}^u_{n_r,n_t}$ and the Bernoulli discrete (BD) pdf of $\mathbf{x}$ [see (20)], we have $\text{E}_{\mathbf{E},\mathbf{x}}(\mathbf{w}_{\text{I}}\mathbf{w}_I^H) = \sigma^2 p_s/p_t\mathbf{I}_{N_rN}$. Inciting by this fact, we approximate $\mathbf{w}_{\text{I}}$ as Gaussian noise $\sim\mathcal{CN}(\mathbf{0}_{N_rN}, \sigma^2 p_s/p_t\mathbf{I}_{N_rN})$. Then $\mathbf{n}$ is approximated as Gaussian noise $\sim\mathcal{CN}(\mathbf{0}_{N_rN}, \tilde{\sigma}^2\mathbf{I}_{N_rN})$ with $\tilde{\sigma}^2 = \sigma^2(p_s/p_t + 1)$ and is assumed to be independent to $\sqrt{p_s}\hat{\mathbf{H}}\mathbf{x}$. These Gaussian assumptions are relatively accurate, particularly when CE errors are small or stating that $\text{SNR}_{\text{train}}$ is high (e.g., they become precise when $\text{SNR}_{\text{train}} \to \infty$ dB). Based on the numerical results in Section VII-B, they bring about minor performance degradation even when $\text{SNR}_{\text{train}}$ is moderately high (e.g., 15 dB). Moreover, relative high $\text{SNR}_{\text{train}}$ could be satisfied in massive SM-MIMO because the BS is expected to acquire high-precision channel state information for uplink MUD and downlink precoding [23], [24]. Therefore, "Gaussian approximation" is effective enough to handle the CE errors. Next, after treating $\mathbf{n}$ as independent and identically distributed (i.i.d.) Gaussian noise, two MU detectors are developed in Sections IV and V, respectively.

IV. Stagewised Linear Detectors for Massive Spatial-Modulation Multiple-Input–Multiple-Output

In Part A, stagewised linear detectors are proposed for massive SM-MIMO. Then, their computational complexities are analyzed in Part B.

A. Stagewised Linear Detectors

Here, the stagewised linear detectors for SM-MIMO over flat-fading channels [5], [13], [14] are applied to solve the MUD problem (14). In general, the linear detectors contain three stages that intend to estimate Ω, $\mathbf{x}_\Omega$, and $\{\mathbf{b}^u\}_{u\in[U]}$, respectively. They are described as follows.

1) Support Detection: To estimate the positions of the nonzero elements of $\mathbf{x}$, we need to acquire a crude estimation of $\mathbf{x}$ first. $\mathbf{x}$ can be estimated by the matching filter (MF) as

$$\hat{\mathbf{x}} = \frac{1}{\sqrt{p_s}N_r}\hat{\mathbf{H}}^H\mathbf{y} \tag{15}$$

or by the minimum MSE (MMSE) estimator as

$$\hat{\mathbf{x}} = \frac{1}{\sqrt{p_s}}\left(\hat{\mathbf{H}}^H\hat{\mathbf{H}} + \sigma^2\left(\frac{1}{p_t} + \frac{1}{p_s}\right)\mathbf{I}_{UN_tN}\right)^{-1}\hat{\mathbf{H}}^H\mathbf{y}. \tag{16}$$

In (16), direct matrix inversion is cumbersome and unnecessary. Based on the Toeplitz-like structure of $\hat{\mathbf{H}}$, it can be decomposed as N $UN_t \times UN_t$-sized matrix inversions (see Appendix A).

Based on the structure of $\mathbf{x}$ in (8), $\hat{\mathbf{x}}$ can be rewritten as

$$\hat{\mathbf{x}} = \left[\ (\hat{\mathbf{x}}^0)^T\ (\hat{\mathbf{x}}^1)^T\ \ldots\ (\hat{\mathbf{x}}^{U-1})^T\ \right]^T \tag{17}$$

where $\hat{\mathbf{x}}^u = \hat{\mathbf{x}}(uN : uN+N-1)$ is the estimation of $\mathbf{x}^u$. Based on (7), $\hat{\mathbf{s}}^u_n = \mathbf{x}^u(\Psi_n)$ is the estimation of $\mathbf{s}^u_n$. Based on [13] and [24], j^u_n in (3) is estimated as $\hat{j}^u_n = \arg\max_{j\in N_t}|\hat{\mathbf{s}}^u_n(j)|$. Finally, Ω^u is approximated as $\hat{\Omega}^u = \{\hat{j}^u_nN + n\}_{n\in[N]}$, and then Ω is estimated as $\hat{\Omega} = \{uN_tN + \hat{\Omega}^u\}_{u\in[U]}$.

2) Signal Reconstruction: Based on $\hat{\Omega}$, (14) is shrunk as

$$\mathbf{y} = \sqrt{p_s}\hat{\mathbf{H}}_{\hat{\Omega}}\mathbf{x}_{\hat{\Omega}} + \mathbf{n}. \tag{18}$$

Then, the MMSE estimation of $\mathbf{x}_{\hat{\Omega}}$ is

$$\hat{\mathbf{x}}_{\hat{\Omega}} = \frac{1}{\sqrt{p_s}}\left(\hat{\mathbf{H}}^H_{\hat{\Omega}}\hat{\mathbf{H}}_{\hat{\Omega}} + \sigma^2\left(\frac{1}{p_t} + \frac{1}{p_s}\right)\mathbf{I}_{UN}\right)^{-1}\hat{\mathbf{H}}^H_{\hat{\Omega}}\mathbf{y} \tag{19}$$

which is treated as the final estimation of $\mathbf{x}_\Omega \in \mathbb{C}^{UN}$ containing the UN nonzero elements of $\mathbf{x}$.

3) Bit Recovery: Estimate q^u_n in (3) as the QAM hard decision of $\hat{\mathbf{x}}_{\hat{\Omega}}(uN_tN + \hat{j}^u_nN + n)$. Based on $\{\hat{j}^u_n\}^{u\in[U]}_{n\in[N]}$ and $\{\hat{q}^u_n\}^{u\in[U]}_{n\in[N]}$, recover the hard decisions of $\{\mathbf{b}^u\}_{u\in[U]}$ by reverting the SM processes shown by (2) and (3).

Unfortunately, the linear detectors have several drawbacks. First, at the stage of support detection, MF (15) has limited ability to remove the MU interference although it is featured on computational simplicity [1], [33]. MMSE (16) outperforms

MF significantly, but it requires to precompute N $UN_t \times UN_t$-sized matrix inversions (see Appendix A), which could be problematic in hardware implementation when UN_t is large [e.g., (16)] [8]. Second, (19) is equivalent to solving linear equations. However, the latter is complex due to the large-scale matrix–matrix multiplication and Cholesky decomposition. To make matters worse, this cumbersome operation has to be repeated for every input SC symbol because the support set Ω keeps changing (so does its estimation $\hat{\Omega}$). Third, the linear detectors lose their effectiveness once the total number of TAs at UE UN_t is larger than the number of BS antennas N_r. This is because, when $UN_t > N_r$, (14) is underdetermined, and both (15) and (16) are not reliable estimations of $\mathbf{x}$ for the subsequent support detection. In summary, the linear detectors are not suitable for massive MIMO. Therefore, new low-complexity detectors should be developed, which is the core topic of the next section.

B. Computational Complexity of the MUD via the Linear Detectors

We analyze the computational complexity from the MUD in one coherence time interval, which contains D SC symbols for data transmission. MMSE is focused. At the start of MUD, the estimator $(\hat{\mathbf{H}}^H\hat{\mathbf{H}} + \lambda\mathbf{I})^{-1}\hat{\mathbf{H}}^H$ ($\lambda > 0$) in (16) can be precalculated and stored. Based on Appendix A, its complexity is about $(16NN_rU^2{N_t}^2 + (16/3)NU^3{N_t}^3 + 5UN_tN_rN\log_2 N)$ flops, where the first term comes from the matrix–matrix multiplications, the second term from N $UN_t \times UN_t$-sized matrix inversions, and the third term from the UN_tN_r N-dimensional fast Fourier transforms (N-FFTs) used to obtain the channel frequency responses. Once this estimator is known, the *support detection* on the D SC symbols involves only N-FFTs and matrix–vector multiplications. Its total complexity is $5D(UN_t + N_r)N\log_2 N + 8DUN_tN_rN$ flops. During the *signal reconstruction*, the Hermitian matrix $\hat{\mathbf{H}}_{\hat{\Omega}}^H\hat{\mathbf{H}}_{\hat{\Omega}}$ in (19) has to be formulated explicitly. Then, (19) is equivalent to solve linear equations where the Cholesky decomposition can be exploited. Consequently, the total complexity of solving (19) is $(8N_rU^2N^3 + (4/3)U^3N^3)$. Finally, during one coherence time, the computational complexity from MUD is about $(16N_rU^2{N_t}^2 + (16/3)NU^3{N_t}^3 + 8DUN_tN_rN + 8DN_rU^2N^3 + (4/3)DU^3N^3)$ flops. Similarly, MF can be also analyzed (details are omitted here).

V. Generalized Approximate Message Passing Detector for Massive Spatial-Modulation Multiple-Input–Multiple-Output

In Part A, the GAMPD is constructed. Then, its uncoded BER performance and computational complexity are analyzed in Parts B and C, respectively.

A. GAMPD

In [33], we solved the MUD problem in classical massive MIMO [1], [2] by the GAMP algorithm in CS [18]. In this paper, this work is extended to massive SM-MIMO as follows.

First, the entries of $\mathbf{x}$ are treated as complex random variables. Based on the mapping processes from $\{\mathbf{b}^u\}_{u=0}^{U-1}$ to $\mathbf{x}$ in Section II-B, they are i.i.d. as

$$p_X(x) = \left(1 - \frac{1}{N_t}\right)\delta(x) + \frac{1}{N_tM}\sum_{k\in[M]}\delta(x - \mathbb{A}_k) \quad (20)$$

where $\delta(\cdot)$ is the Dirac function, and $\mathbb{A}_k$ is the kth element of the QAM alphabet $\mathbb{A}$ (1). In the BD pdf (20), $(1 - 1/N_t)$ and $1/N_t$ are the zero and nonzero probabilities for an arbitrary entry of $\mathbf{x}$. This is because, at every data sample, one UE randomly activates one TA among its N_t TAs with equal probability. The two $\delta(\cdot)$'s come from the fact that the inactive TAs transmit zeros and the active TA transmits QAM symbols chosen from $\mathbb{A}$ uniformly.

Second, the MMSE estimation of the elements of $\mathbf{x}$ is $\{\hat{x}_n = \sum_{x_n\in\mathbb{A}} x_n p(x_n|\mathbf{y})\}_{n\in[UN_tN]}$, where $p(x_n|\mathbf{y})$ is the posterior pdf of x_n. Based on [18] and [33]–[35], we know that loopy belief propagation (LBP) can be applied to calculate $p(x_n|\mathbf{y})$ approximately but has high computational complexity due to its high-dimensional integrations and large number of messages. Fortunately, LBP has been simplified as GAMP based on the central limit theorem and Taylor expansions in [18]. In this paper, GAMP is exploited to construct $\mathbf{x}$ in (14). Finally, a new GAMPD is constructed as follows.

1) Initialization: Set the iteration number as $t = 0$. With $m \in [N_tN]$ and $n \in [UN_tN]$, initialize the mean and variance of x_n as $\hat{x}_n(0) = 0$ and $\tau_n^x(0) = 1/N_t$, respectively. Set $\{u_m(-1)\}_{m\in[N_tN]} = 0$ and the termination parameter tol as a small positive number.

2) Iteration: Compute the following equation:

$$\tau^p(t) = \frac{p_s(1 + \sigma^2/p_tU)}{N}\sum_{n\in[UN_tN]}\tau_n^x(t). \quad (21)$$

For each $m \in [N_tN]$, calculate

$$\hat{p}_m(t) = \sqrt{p_s}\sum_{n\in[UN_tN]}\hat{H}_{mn}\hat{x}_n(t) - \tau^p(t)u_m(t-1) \quad (22)$$

$$u_m(t) = \left(y_m - \hat{p}_m(t)\right) / \left(\tau^p(t) + \tilde{\sigma}^2\right) \quad (23)$$

where $\tilde{\sigma}^2 = \sigma^2(p_s/p_t + 1)$ has been defined in Section III.

Compute the following equation:

$$\tau^r(t) = \frac{1}{N_rp_s(1 + \sigma^2/p_tU)}\left(\tau^p(t) + \tilde{\sigma}^2\right). \quad (24)$$

For each $n \in [UN_tN]$, calculate

$$\hat{r}_n(t) = \hat{x}_n(t) + \sqrt{p_s}\tau^r(t)\sum_{m\in[N_tN]}\hat{H}_{mn}^*u_m(t) \quad (25)$$

$$\hat{x}_n(t+1) = \mathrm{EXP}\left(x_n|\hat{r}_n(t), \tau^r(t)\right) \quad (26)$$

$$\tau_n^x(t+1) = \mathrm{VAR}\left(x_n|\hat{r}_n(t), \tau^r(t)\right). \quad (27)$$

Equations (26) and (27) are the expectation and variance of x_n at the $(t+1)$th iteration, which are evaluated with respect to $p(x_n|\hat{r}_n) \propto \mathcal{CN}(x_n; \hat{r}_n, \tau^r)p_X(x_n)$ where $p_X(x_n)$ is the BD pdf (22).

3) Termination Judgment: If the termination condition

$$\frac{\|\hat{\mathbf{x}}(t+1)-\hat{\mathbf{x}}(t)\|_2^2}{\|\hat{\mathbf{x}}(t+1)\|_2^2} < tol \tag{28}$$

is satisfied ($\hat{\mathbf{x}}(t)$ is a vector containing $\hat{x}_n(t)$ as the nth element), set $G=t+1$ and go to step 4); otherwise, set $t=t+1$ and go to step 2) for the next iteration.

4) Bit Recovery: Set $G=t$ to record the number of iterations before termination. Output $\hat{\mathbf{x}}(G)$ as the final estimation of $\mathbf{x}$. Based on $\hat{\mathbf{x}}(G)$, recover the hard decisions of $\{\mathbf{b}^u\}_{u\in[U]}$ similar to Section IV-A (details are omitted due to space constraints).

In GAMPD, the first three steps are obtained by adapting GAMP [18] to the MUD problem (14). They provide $\hat{\mathbf{x}}(G)$ as the final estimation of $\mathbf{x}$, which is used to generate the hard decisions of $\{\mathbf{b}^u\}_{u\in[U]}$ in the fourth step. Next, we present more details of GAMPD.

First, compared with GAMP [18] and its applications [22], [33], [48], GAMPD has four important differences. First, in [18], [22], [33], and [48], the measurement matrix (MM) was usually assumed to be known precisely. However, in this paper, the MM $\mathbf{H}$ is unknown and estimated as $\hat{\mathbf{H}}$ accompanying with MM uncertainty (MMU) [37], [49] $\mathbf{E}$. Therefore, GAMPD has actually solved the signal reconstruction problem (14) under the MMU. Second, $\hat{\mathbf{H}}$ possesses a special Toeplitz-like structure that can be utilized to accelerate the matrix–vector multiplications in (22) and (25) (refer to Appendix B for more details). Third, GAMPD directly recovers the complex-valued sparse signal $\mathbf{x}$ without transforming (14) into the real domain through doubling the problem dimensions [33]. Therefore, it has exploited the important property that the real and imaginary parts of any entry of $\mathbf{x}$ are zeros or nonzeros simultaneously. Fourth, the signal prior distribution is known as a special BD pdf (20). By substituting (20) into (26) and (27), we have

$$\hat{x}_n(t+1) = \sum_{k\in[M]} \mathbb{A}_k p_k \tag{29}$$

$$\tau_n^x(t+1) = \sum_{k\in[M]} |\mathbb{A}_k|^2 p_k - |\hat{x}_n(t+1)|^2 . \tag{30}$$

In these equations, $p_k=\phi(\mathbb{A}_k;\hat{r}_n(t),\tau^r(t))/MN_tP$, where $\phi(s;\hat{r}_n(t),\tau^r(t))=\exp(-s-\hat{r}_n(t)|^2/\tau^r(t))$ is a complex Gaussian function, and $P=(1-1/N_t)\phi(0;\hat{r}_n(t),\tau^r(t))+1/MN_t\sum_{k\in[M]}\phi(\mathbb{A}_k;\hat{r}_n(t),\tau^r(t))$ is the probability normalization factor. In the derivation of (29) and (30), we have exploited the fact that $p(x_n|\hat{r}_n)\propto\mathcal{CN}(x_n;\hat{r}_n,\tau^r)p_X(x_n)$ in (26) and (27) is still a BD pdf after multiplying a Gaussian pdf $\mathcal{CN}(x_n;\hat{r}_n,\tau^r)$ with a BD pdf $p_X(x_n)$.

Second, in (21) and (24), we have utilized the approximation $|\hat{H}_{mn}|^2\approx(1+\sigma^2/p_tU)/N$ to remove two potential matrix–vector multiplications involving matrix $|\hat{\mathbf{H}}|^2$ whose (m,n)th element is $|\hat{H}_{mn}|^2$. We have this approximation because the expectation of the ℓ_2-norm of every row of $\hat{\mathbf{H}}$ is $UN_t(1+\sigma^2/p_tU)$. This simple trick has been widely used to simplify the approximate message passing (AMP)-like algorithms in [18] and [36].

Third, GAMPD exploits the signal prior BD pdf (20) in (26) and (27). Moreover, similar as the popular Bernoulli–Gaussian pdf in CS [29], [36], [38], the BD pdf is also a sparsity-inducing distribution. Consequently, the reconstructed signal would be sparse. In other words, GAMPD has exploited the sparsity of the transmitted signal $\mathbf{x}$ implicitly. This is why it could work well even when (9) and (14) are underdetermined (i.e., $UN_t>N_r$). This is exactly the most exciting story of CS [29]. In contrast, the linear detectors fail to exploit the prior pdf and sparsity of $\mathbf{x}$; hence, they cannot handle the underdetermined case and are inferior to GAMPD (see Figs. 4 and 5 in Section VII-A).

B. Predict the Uncoded BER Performance of GAMPD

Here, we introduce how to predict the uncoded BER performance of GAMPD.

At first, the MSE between the signal estimation at the tth iteration of GAMPD $\hat{\mathbf{x}}(t)$ and the true signal $\mathbf{x}$ is defined as

$$e_{\mathrm{GAMPD}}(t)=\frac{\|\hat{\mathbf{x}}(t)-\mathbf{x}\|_2^2}{UN_tN} \tag{31}$$

where t starts from 0, and $\hat{\mathbf{x}}(0)$ is initialized as $\mathbf{0}_{UN_tN}$. $\mathbf{x}\in\mathbb{C}^{UN_tN}$ is an all-zero vector except its UN M-QAM symbols with unit power. Consequently, $e_{\mathrm{GAMPD}}(0)\approx 1/N_t$.

Section V-A shows that the hard decisions of $\{\mathbf{b}^u\}_{u\in[U]}$ are derived from $\hat{\mathbf{x}}(G)$ by reverting the SM processes. Therefore, the uncoded BER (denoted as ber) strongly relies on the accuracy of $\hat{\mathbf{x}}(G)$ and parameters $\{N_t,M\}$. Under fixed $\{N_t,M\}$, the smaller $e_{\mathrm{GAMPD}}(G)$ is, the more accurate $\hat{\mathbf{x}}(G)$ is, and then the smaller ber is. Therefore, there should exist monotone relationship between ber and $e_{\mathrm{GAMPD}}(G)$. It is denoted as

$$ber=\mathrm{BER}_{N_t,M}\left(e_{\mathrm{GAMPD}}(G)\right). \tag{32}$$

This property is validated by the numerical results in Section VII-C. Deriving the expression of $\mathrm{BER}_{N_t,M}(\cdot)$ mathematically is difficult and left as our future research topic. Fortunately, it can be well approximated by a third-order polynomial (i.e., (41) in Section VII-C).

Until now, the key is how to predict G and $e_{\mathrm{GAMPD}}(G)$. This is achieved by resorting to the SE, which is widely used to analyze the MSE performance of the AMP-like algorithms [18], [21], [22]. Based on [18], and [22], [33], the SE iteration for GAMPD is

$$e_{\mathrm{SE}}(t+1)=\mathrm{E}\left(\mathrm{VAR}\left(x|r,\frac{1}{\beta}e_{\mathrm{SE}}(t)+\sigma_{SE}^2\right)\right) \tag{33}$$

where $e_{\mathrm{SE}}(0)=1/N_t$, $\mathrm{VAR}(\cdot)$ is defined in (27), $\beta=N_r/UN_t$, $\sigma_{\mathrm{SE}}^2=(\sigma^2/p_s+\sigma^2/p_t)/N_r(1+\sigma^2/Up_t)$, and the expectation $\mathrm{E}(\cdot)$ is taken over the scalar random variable $r=x+w$, where $x\sim p_X(x)$ [given by (20)] and $w\sim\mathcal{CN}(0,e_{\mathrm{SE}}(t)/\beta+\sigma_{SE}^2)$.

Based on [18], [22], and [33], if $\mathbf{n}$ in (14) is exactly i.i.d. Gaussian noise, the SE iteration (33) could precisely track the MSE performance of GAMPD, i.e., $\{e_{\mathrm{SE}}(t+1)\}_t$ are almost equal to $\{e_{\mathrm{GAMPD}}(t+1)\}_t$. Nevertheless, Section III has pointed out that $\mathbf{n}$ is actually correlated with the unknown signal $\mathbf{x}$ and has been approximated as Gaussian noise. This "Gaussian approximation" brings about performance loss. Therefore, the MSE performance of GAMPD is worse than

that predicted by SE, i.e., $\{e_{\rm SE}(t+1) \leq e_{\rm GAMPD}(t+1)\}_t$. However, when the CE errors are low or stating $\text{SNR}_{\rm train}$ that is relatively high (e.g., 15 dB), the Gaussian approximation could be accurate. Therefore, $\{e_{\rm SE}(t) \approx e_{\rm GAMPD}(t)\}_t$. These conclusions are further corroborated in Figs. 6 and 7 in Section VII. Finally, $e_{\rm GAMPD}(G)$ can be approximated as $e_{\rm SE}(G)$.

However, the key is how to predict G. This problem is solved as follows. At first, we introduce a termination condition for the SE iteration (33) as

$$|e_{\rm SE}(t+1) - e_{\rm SE}(t)| < TOL \tag{34}$$

where TOL is a small positive number. When (34) is satisfied (T records the value of $(t+1)$ at that moment), $e_{\rm SE}(t+1)$ becomes similar as $e_{\rm SE}(t)$. Due to the approximation $\{e_{\rm SE}(t+1) \approx e_{\rm GAMPD}(t+1)\}_t$, $e_{\rm GAMPD}(t+1)$ would be also similar to $e_{\rm GAMPD}(t)$. Based on (31), this indicates that $\hat{\mathbf{x}}(t+1)$ will be similar to $\hat{\mathbf{x}}(t)$. With high probability, (28) could be satisfied, and GAMPD would be terminated. Consequently, we have $G \approx T$ and $e_{\rm GAMPD}(G) \approx e_{\rm SE}(T)$. These two approximations are plugged into (32), and then ber is predicted as

$$ber \approx \text{BER}_{N_t,M}\left(e_{\rm SE}(T)\right). \tag{35}$$

In summary, the uncoded BER performance of GAMPD is predicted as follows. Under current parameter setting $\{U, N_t, M\text{-QAM}, \text{SNR}_{\rm sig}, N_r\}$, we obtain $e_{\rm SE}(T)$ by repeating the SE iteration (33) until (34) is satisfied. Then, $e_{\rm SE}(T)$ is substituted into (35) to obtain an estimation of ber. The caution is that, due to $e_{\rm SE}(T) \leqslant e_{\rm GAMPD}(G)$ and the monotonicity of $\text{BER}_{N_t,M}(\cdot)$, there exists $\text{BER}_{N_t,M}(e_{\rm SE}(T)) \leqslant ber$. Therefore, (35) actually provides an overoptimistic estimation of ber.

C. Computational Complexity of the MUD via GAMPD

Similar to Section IV-B, we focus on the MUD in one coherence time interval containing D SC symbols for data transmission. The two matrix–vector multiplications in (22) and (25) are the most complex operations of GAMPD. In Appendix B, they have been accelerated by exploiting the Toeplitz-like structure of $\mathbf{H}$. Their total complexity is $16UN_tN_rN + 10(UN_t + N_r)N\log_2 N$ flops, where the first term comes from the matrix–vector multiplication per subcarrier, and the second term comes from the N-FFTs before and after the matrix–vector multiplications. In addition, UN_tN_r N-FFTs are executed to obtain the frequency channel response, which will be stored and accessed directly from then on. Moreover, GAMPD needs G iterations. Finally, during the channel coherence time, the total complexity from MUD is $(5(UN_tN_r + 2DGUN_t + 2DGN_r)N\log_2 N + 16DGUN_tN_rN)$ flops, which is much less than its counterpart in Section IV-B.

Next, we analyze how G is changed under different system settings. After $\text{SNR}_{\rm train}$, $\text{SNR}_{\rm sig}$, M-QAM, and N_t are fixed, Section V-B showed that G is comparable to T. In Appendix C, we prove that T is irrelevant to the problem dimensions (values of N_r and U) but exclusively influenced by $\alpha = N_r/U$. In other words, G is almost unchanged if we fix α and scale up N_r and U simultaneously. Moreover, our numerical results (see Figs. 7 and 8) indicate that GAMPD can quickly converge after 5–15 iterations when α is 2–8. Therefore, the complexity from GAMPD is $o(DUN_tN_rN)$. In contrast, based on Section IV-B, the complexity from MMSE is $o(DN_rU^2N^3)$. Consequently, GAMPD enjoys orders-of-magnitude faster speed.

VI. Designing Energy-Efficient Massive Spatial-Modulation Multiple-Input–Multiple-Output

Once the number of BS antennas reaches tens even hundreds, the total circuit power consumption would be high [40]–[43], [45]. This could undermine or even overwhelm the benefits from decreasing the transmission power (power consumed at the power amplifiers) [1], [7]. Consequently, it was pointed out in [5] that more emphasis should be shifted from spectral efficiency to EE in the design of future communication systems. Therefore, we optimize the system parameters (e.g., the number of BS antennas and the number of TAs at UE) to design energy-efficient massive SM-MIMO, where GAMPD is considered as the MU detector.

In [5], [42], and [43], EE has been defined as the ratio of the channel capacity to the total energy consumption. Similar to [44], we obtain the capacity of MU SM-MIMO as

$$C = U\log_2(N_tM) - \text{E}_\theta\left[\log_2 \frac{\sum_{\mathbf{x}'\in\mathcal{X}} p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}';\hat{\mathbf{H}})}{p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x};\hat{\mathbf{H}})}\right] \tag{36}$$

where $\mathcal{X}$ contains the $(N_tM)^U$ possible values of $\mathbf{x}$, $\theta = (\mathbf{x}, \mathbf{y})$, and $p_{\mathbf{Y}|\mathbf{X}}(\cdot|\cdot)$ can be derived based on (14). Unfortunately, (36) has no closed-form expression [44] and has to be evaluated by Monte Carlo simulations that are cumbersome due to the high dimensionality of $\mathbf{x}$ and $\mathbf{H}$. Therefore, it is difficult to be exploited to optimize the system parameters.

In this paper, we design the system parameters from a new perspective. First, we only consider the possible combinations of $\{U, N_t, M, p_s, N_r\}$, which ensure that the recovered data bits can satisfy the BER constraint $ber \leq ber_{\rm target}$ ($ber_{\rm target}$ is the maximum tolerable BER). Second, under this constraint, we expect to maximize the transmission rate but minimize the energy consumption. However, the latter two usually contradict with each other and cannot be satisfied simultaneously. Based on the core idea of EE optimization [5], [42], [43], a natural strategy is to maximize the ratio of the number of the transmitted data bits to the total power consumption. Finally, a new optimization problem is mathematically formulated as follows.

We assume that the coherence time contains $T_{\rm coh}$ data samples where every data sample lasts T_s second (s). First, we calculate the total number of bits transmitted by the U UE during one coherence time. Based on Sections II-B and III, the time overhead for CE is $N_t(U+1)L$. Consequently, the time length for data transmission is $T_{\rm sig} = T_{\rm coh} - N_t(U+1)L$, and its corresponding number of SC symbols is $K_{\rm sc} = \lceil T_{\rm sig}/(L+N)\rceil$. Note that the data length of the K_{sc}th symbol is $N_{\rm last} = T_{\rm sig} - (K_{\rm sc}-1)(N+L) - L$, which could be less than N. Therefore, we have $T_{\rm eff} = (K_{\rm sc}-1)N + N_{\rm last}$ data samples for information transmission (CP has been removed).

The data rate of SM is $\log_2(N_t M)$ bits/symbol. Therefore, in one coherence time interval, U UE can transmit

$$B = UT_{\text{eff}}\log_2(N_t M) \text{ bits.}$$

Second, we analyze the power consumption at the BS and UE. The power consumption at the BS is

$$P_{\text{BS}} = p_{\text{circ}}^{\text{BS}} N_r T_{\text{coh}} T_s \text{ Joule (J)}$$

where $p_{\text{circ}}^{\text{BS}}$ (in watts) is the circuit power of every BS RF chain. The power consumed by the UE is

$$P_{\text{UE}} = p_t U N_t (U+1) L T_s + p_s U T_{\text{sig}} T_s + p_{\text{circ}}^{\text{UE}} U T_{\text{coh}} T_s \text{ J}$$

where the first and the second components indicate the transmission power during the periods of channel training and signal transmission, respectively, and the third component is the total circuit power consumption at the U UE. $p_{\text{circ}}^{\text{UE}}$ W is the UE circuit power.

Finally, we propose a new metric

$$EE = B/(P_{\text{UE}} + P_{\text{BS}}) \text{ bit/J, under } ber \leqslant ber_{\text{target}} \quad (37)$$

to characterize how energy-efficient massive SM-MIMO is.

Finally, system parameters are optimized to maximize (37). Based on (32), the BER constraint in (37) is converted as $e_{\text{GAMPD}}(G) \leqslant \text{BER}_{N_t,M}^{-1}(ber_{\text{target}})$. Then, we obtain an optimization problem

$$\begin{aligned} \underset{\{U,N_t,M,p_s,N_r\}}{\text{maximize}} \quad & EE \\ \text{subject to} \quad & e_{\text{GAMPD}}(G) \leqslant \text{BER}_{N_t,M}^{-1}(ber_{\text{target}}) \end{aligned} \quad (38)$$

where U, N_t, M, p_s, and N_r are the optimization variables. Note that p_t does not appear in the variable set of (38) because it can be determined by U as $10^{\text{SNR}_{\text{train}}/10}\sigma^2/U$ (see the definition of $\text{SNR}_{\text{train}}$ in Section II-C).

Based on (35), the constraint in (38) is replaced as $e_{\text{SE}}(T) \leq \text{BER}_{N_t,M}^{-1}(ber_{\text{target}})$. Then, (38) is changed as

$$\begin{aligned} \underset{\{U,N_t,M,p_s,N_r\}}{\text{maximize}} \quad & EE \\ \text{subject to} \quad & e_{\text{SE}}(T) \leqslant \text{BER}_{N_t,M}^{-1}(ber_{\text{target}}). \end{aligned} \quad (39)$$

Finally, (39) is solved as follows. First, $U \in [U_{\min}, U_{\max}]$, where $U_{\min}$ and $U_{\max}$ are the minimum and maximum numbers of UE. $U_{\max}$ is $\lfloor T_{\text{coh}}/N_t L - 1 \rfloor$ corresponding to the special case that almost the whole coherent time is occupied by CE (i.e., $N_t(U_{\max}+1)L \approx T_{\text{coh}}$). Based on the principles of SM [13], [14], N_t should be a power of two. Moreover, it cannot be too large due to the limited size of UE. Therefore, its possible values are set as $N_t \in \{1,2,4,8\}$. QPSK, 16-QAM, and 64-QAM are considered; thus, $M \in \{4,16,64\}$. Based on the definition of SNR_{sig} in Section II-B, the signal transmission power p_s is $10^{\text{SNR}_{\text{sig}}/10}\sigma^2/U$, where SNR_{sig} is increased from $\text{SNR}_{\min}$ dB to $\text{SNR}_{\max}$ dB with step size SNR_{step} dB. $N_r \in [N_r^{\min}, N_r^{\max}]$ where $N_r^{\min} = \alpha^{\min}U$ and $N_r^{\max} = \alpha^{\max}U$. $\alpha^{\min}$ and $\alpha^{\max}$ are the minimum and maximum values of the ratio $\alpha = N_r/U$. Note that $\alpha^{\max}$ is set to be large enough to ensure that the MSE constraint in (39) is satisfied at least when $N_r = N_r^{\max}$.

Second, for every realization of $\{U, N_t, M, p_s\}$, SE (33) is applied to determine the minimum N_r that guarantees the MSE constraint in (39) is satisfied. Then, EE is calculated based on (37). If EE is larger than $EE_{\max}$, we will update the current optimal parameter set $\mathbb{P}_{\text{opt}} = \{U, N_t, M, p_s, N_r\}$, and set $EE_{\max} = EE$. The algorithm details are listed in Tables I and II.

Finally, the caveat is that the algorithm in Table I is actually overoptimistic. Due to the inaccuracy of the Gaussian approximation on $\mathbf{w}_{\text{I}}$ in Section III, $e_{\text{GAMPD}}(G)$ could be slightly larger than $e_{\text{SE}}(T)$ (as explained in Section V-B). Therefore, the parameters' outputs in Table I may cause $e_{\text{GAMPD}}(G) > \text{BER}_{N_t,M}^{-1}(ber_{\text{target}})$ (i.e., $ber > ber_{\text{target}}$) although they ensure $e_{\text{SE}}(T) \leqslant \text{BER}_{N_t,M}^{-1}(ber_{\text{target}})$ is satisfied. Therefore, (39) only provides an approximate solution to (38). The data bits recovered by GAMPD could violate the BER constraint $ber \leqslant ber_{\text{target}}$. Fortunately, they do not deviate from the expected BER significantly. For example, in Section VII-D, the expected BER is chosen as 0.01, and most of the actual BERs are among 0.011–0.018. Therefore, Table I still provides an effective tool for designing energy-efficient massive SM-MIMO with predefined performance expectation.

TABLE I
EE OPTIMIZATION ALGORITHM

Initialization:
Initialize the system parameters including ber_{target}, σ^2, SNR_{train}, $p_{\text{circ}}^{\text{UE}}$, T_{coh}, and $p_{\text{circ}}^{\text{BS}}$. $EE_{\max}$ records current maximum value of EE and is initialized as 0. Set values for $U_{\min}$, $U_{\max}$, $SNR_{\min}$, SNR_{step}, $SNR_{\max}$, $\alpha^{\min}$ and $\alpha^{\max}$; calculate $N_r^{\min} = \alpha^{\min}U$ and $N_r^{\max} = \alpha^{\max}U$;
Iteration:
For every possible $\{U, N_t, M, SNR_{\text{sig}}\}$ where $U \in \{U_{\min}:1:U_{\max}\}$, $N_t \in \{1,2,4,8\}$, $M \in \{4,16,64\}$, $SNR_{\text{sig}} \in \{SNR_{\min}:SNR_{\text{step}}:SNR_{\max}\}$:
i) Calculate $p_t = 10^{SNR_{\text{train}}/10}\sigma^2/U$ and $p_s = 10^{SNR_{\text{sig}}/10}\sigma^2/U$;
ii) Exploit the bisection method $\text{BISEC}(U, N_t, M, p_s, p_t)$ in Table II to determine the minimum N_r;
iii) Calculate EE as (37). If $EE > EE_{\max}$, set $EE_{\max} = EE$ and update the optimal parameter set $\mathbb{P}_{\text{opt}}$ as $\{U, N_t, M, p_s, N_r\}$.
Termination:
Output the optimal system parameter set $\mathbb{P}_{\text{opt}}$.

TABLE II
BISECTION METHOD $\text{BISEC}(U, N_t, M, p_s, p_t)$ TO DETERMINE THE MINIMUM NUMBER OF BS ANTENNAS

Initialization:
Initialize the lower and upper bound for N_r as $N_r^{\text{low}} = N_r^{\min}$ and $N_r^{\text{up}} = N_r^{\max}$, and set N_r as $N_r = \lfloor (N_r^{\text{up}} + N_r^{\text{low}})/2 \rfloor$.
Iteration:
i) Calculate $e_{\text{SE}}(T)$ by repeating the SE iteration (33) until (34) is satisfied.
ii) If $e_{\text{SE}}(T) \leq \text{BER}_{N_t,M}^{-1}(ber_{\text{target}})$, set $N_r^{\text{up}} = N_r$; otherwise, set $N_r^{\text{low}} = N_r$.
iii) If $N_r^{\text{up}} - N_r^{\text{low}} > 1$, go to ***Iteration*** for the next iteration; otherwise, go to ***Termination*** for output.
Termination:
Output N_r as the minimum number of BS antennas.

VII. Simulation

In Part A, GAMPD is compared to MMSE, ML [13], [14], and AS [39]. In Part B, we research the impacts of the CE errors on GAMPD. In Part C, we test the BER prediction strategy in Section V-B. Finally, we show how to design energy-efficient massive SM-MIMO based on the algorithm in Table I.

For clarity, the common system parameters are summarized as follows. In one SC symbol, the data part contains $N = 64$ samples with time duration $T_s = 5 \times 10^{-8}$ s. The CP length is $L = 10$ and is equal to the number of channel taps. The circuit noise variance is $\sigma^2_{\text{circ}} = -80$ dBm, and the channel fading coefficient is $\rho = 2 \times 10^{-11}$. Therefore, $\sigma^2 = \sigma^2_{\text{circ}}/\rho$ is 0.5 W. tol in (28) and TOL in (34) are set as 10^{-3}. The uncoded BERs are obtained by averaging across all the U UE sharing the same parameter set $\{p_s, N_t, M\text{-QAM}\}$. Without special notice, $\text{SNR}_{\text{train}}$ is always fixed as 15 dB.

A. GAMPD Versus MMSE, ML, and AS

In the first experiment, GAMPD, stagewised MMSE, and the ML detector [13], [14] for SM-MIMO are compared with each other. Flat-fading channel is considered to ensure the computational feasibility of ML. Moreover, AS [39], which is another important strategy for utilizing the multiple UE antennas, is also added as a reference system. In AS, instead of activating the TA randomly as SM-MIMO, the BS determines and notifies the UE which TA they should activate under the criteria of maximizing the MU-MIMO channel capacity. Then every UE transmits QAM symbols from its assigned TA, and the BS recovers these QAM symbols by MMSE method. The simulation parameters are $N_t = 4$ and $U = 4$. N_r is varied as 18 and 64. For fairness, GAMPD, MMSE, ML, and AS are compared under the same system throughput (i.e., 16 bit/transmission). Therefore, we use QPSK in GAMPD, MMSE and ML, and 16-QAM in AS. AS requires higher order QAM because it loses the spatial constellation of SM for data transmission.

Fig. 4 indicates that GAMPD can approach to the performance of ML. For example, at BER 10^{-2}, the SNR loss is 0.8 dB when $N_r = 18$ and shrinks as 0.25 dB when N_r is increased as 64. GAMPD, MMSE, and ML perform similarly when $N_r = 64$ is much larger than $UN_t = 16$. Similar phenomenon has been observed in classical massive MIMO [1] where different detectors performed similarly if the "massive" condition (i.e., the number of BS antennas is much larger than the number of single-antenna users) is satisfied. This is an important motivation for adopting massive antennas at the BS. When $N_r = 18$, AS scheme can even outperform ML when SNR is lower than -1 dB. Unfortunately, this low-SNR region is not useful for information transmission because the BER is ultrahigh (> 0.1). Once SNR exceeds 0 dB, ML and GAMPD outperform AS, and their performance gains are gradually enlarged with the increase in SNR_{sig}. When $N_r = 64$, AS shows apparent SNR loss compared with GAMPD, ML, and MMSE. For example, at BER 10^{-2}, the SNR loss is about 3.4–4.4 dB. In addition, in AS, choosing and notifying TA indexes to the users call for extra communication cost and overhead. Therefore, AS is less attractive in massive SM-MIMO.

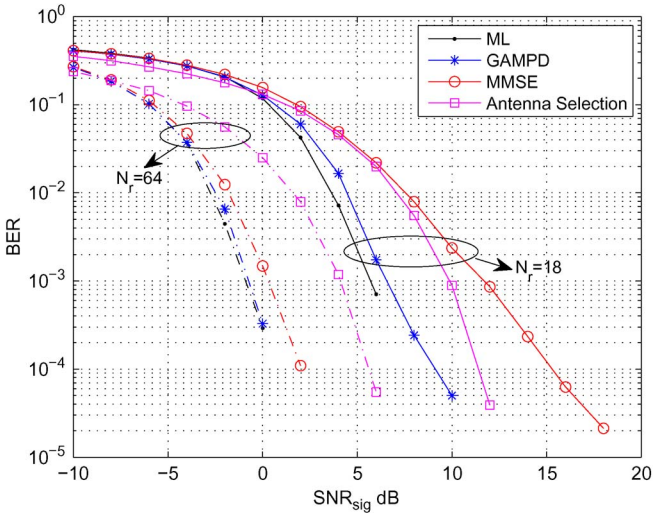


Fig. 4. BER performance of ML, GAMPD, MMSE, and antenna selection (AS). The first three detectors adopt QPSK, and AS uses 16-QAM. Consequently, they reach the same throughput, i.e., 16 bits per channel usage. A flat-fading channel is considered. $U = 4$, $N_t = 4$, and N_r is varied as 18 and 64.

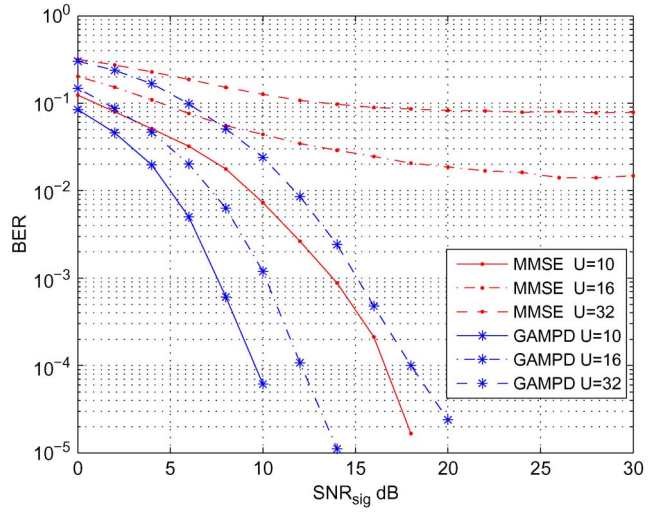


Fig. 5. BER performance of (solid blue curves) the GAMPD and (dashed red curves) the MMSE detector under frequency-selective channels. $N_r = 128$, $N_t = 8$, and 16-QAM. U has three possible values, i.e., 10, 16, and 32, which correspond to the three different cases $UN_t < N_r$, $UN_t = N_r$, and $UN_t > N_r$.

In the second experiment, we compare GAMPD with MMSE under the frequency-selective case (ML is not used due to its prohibitive complexity, and AS is omitted because it is undesirable, as indicated by the previous experiment). The simulation parameters are $N_r = 128$, $N_t = 8$, and 16-QAM. Three possible values of U are 10, 16, and 32, which correspond to the three cases that the total number of TAs (UN_t) is smaller than, equal to, and larger than N_r, respectively.

Fig. 5 shows that GAMPD always outperforms MMSE. With the increase in U, the BER performance of GAMPD degrades gracefully. In contrast, the performance of MMSE deteriorates quickly. The performance gain of GAMPD over MMSE is enlarged with the increase in U. For example, at BER 10^{-2}, the SNR gain is 4.3 dB when U is 10. When $U = 16$, MMSE shows "error floor" at BER $10^{-1.65}$ once $\text{SNR}_{\text{sig}} > 25$ dB.

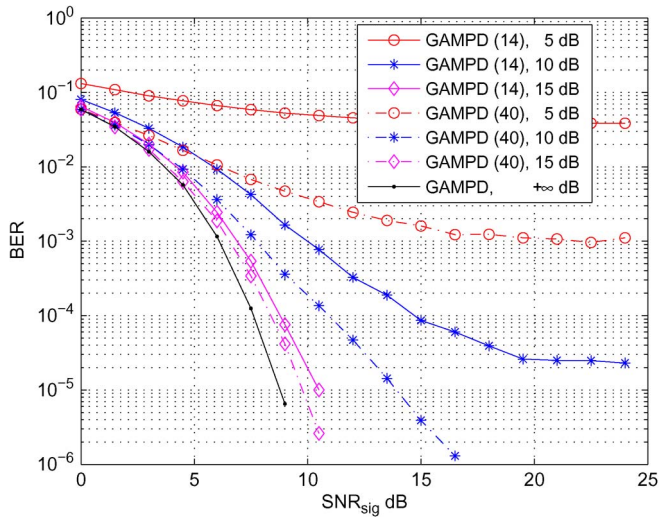


Fig. 6. BER performances of the two GAMPDs for the linear system (14) and (40) under different $\mathrm{SNR}_{\mathrm{train}}$'s (i.e., 5, 10, 15, and $+\infty$ dB). $N_r = 128$, $U = 10$, $N_t = 4$, and 16-QAM.

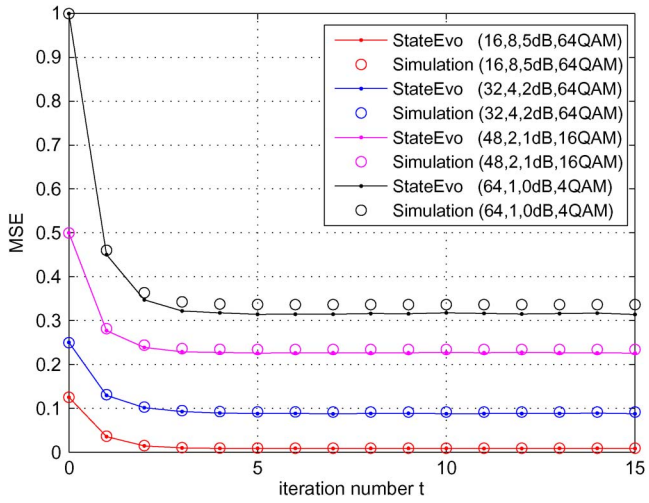


Fig. 7. Tracking the MSE performance of GAMPD. Bracketed numbers indicate the values of $(U, N_t, \mathrm{SNR}_{\mathrm{sig}}\,\mathrm{dB}, M\text{-QAM})$. N_r is 128. Solid curves show $\{e_{\mathrm{SE}}(t)\}_t$, and hollow markers show $\{e_{\mathrm{GAMPD}}(t)\}_t$.

This phenomenon is caused by the CE errors. When $U = 32$ $(N_r < UN_t)$, MMSE is trapped at the "error floor" $10^{-1.1}$ once $\mathrm{SNR}_{\mathrm{sig}}$ exceeds 20 dB. This should be attributed to both the CE errors and the fact that (14) becomes to be underdetermined. In contrast, when U is 16 or 32, GAMPD still works well. Its BERs can approach to 0 with the increase in $\mathrm{SNR}_{\mathrm{sig}}$. Compared with MMSE, it is more robust to CE errors and shows more superior detection performance. This is because GAMPD has exploited the sparsity and prior pdf of the transmitted signal $\mathbf{x}$ (see Section V for more details).

B. Impacts of the CE Errors

In Section III, the interference $\mathbf{w}_I = -\sqrt{p_s}\mathbf{E}\mathbf{x}$ caused by CE errors is assumed to be Gaussian and independent to $\mathbf{x}$. This assumption is relatively accurate when the CE errors are small (i.e., $\mathrm{SNR}_{\mathrm{train}}$ is high). To validate this conclusion, we also consider a virtual system

$$\mathbf{y} = \sqrt{p_s}\hat{\mathbf{H}}\mathbf{x} + \tilde{\mathbf{n}}. \tag{40}$$

Different from $\mathbf{n}$ in (14), $\tilde{\mathbf{n}} \sim \mathcal{CN}(\mathbf{0}_{N_rN}, \tilde{\sigma}^2\mathbf{I}_{N_rN})$ in (40) is true Gaussian noise. GAMPD is applied to (40) to recover $\mathbf{x}$. Naturally, it will outperform the GAMPD for (14) because $\mathbf{n}$ in (14) is actually correlated with $\sqrt{p_s}\hat{\mathbf{H}}\mathbf{x}$.

The simulation parameters are $N_r = 128$, $U = 10$, $N_t = 4$, and 16-QAM. Fig. 6 shows that, in general, GAMPD degrades with the decrease in $\mathrm{SNR}_{\mathrm{train}}$. At $\mathrm{SNR}_{\mathrm{train}} = 15$ dB, GAMPDs in (14) and (40) show minor performance degradation compared with the lower bound provided by GAMPD at $\mathrm{SNR}_{\mathrm{train}} = +\infty$ dB (i.e., no CE errors). In contrast, at 10 or 5 dB, the performance losses caused by CE errors are apparent. At the same time, at $\mathrm{SNR}_{\mathrm{train}} = 15$ dB, the GAMPD in (14) shows only minor performance degradation (0.3–0.7 dB) in comparison with its counterpart in (40). However, when $\mathrm{SNR}_{\mathrm{train}}$ is decreased to be 10 or 5 dB, it is apparently inferior to the latter. Therefore, the Gaussian approximation of $\mathbf{w}_I$ is effective when the SNR at the stage of CE $(\mathrm{SNR}_{\mathrm{train}})$ is moderately large. When $\mathrm{SNR}_{\mathrm{train}}$ is low, although GAMPD still works steadily, simply treating $\mathbf{w}_{\mathrm{I}}$ as Gaussian noise is not very powerful. In the future, more advanced techniques (e.g., the matrix-uncertain AMP [49] or the bilinear GAMP [50]) could be considered.

C. Predicting the BER Performance of GAMPD

In Section V-B, during the prediction of the uncoded BER performance of GAMPD, three important conclusions have been exploited. They are that SE approximately tracks the MSE performance of GAMPD, SE can predict when GAMPD is terminated, and there exists monotone relationship between the MSEs and BERs of GAMPD. Here, they are validated one by one as follows.

In the first experiment, SE is executed to track the MSE performance of GAMPD. We fixed $N_r = 128$. In Fig. 7, the bracketed numbers represent $(U, N_t, \mathrm{SNR}_{\mathrm{sig}}\,\mathrm{dB}, \text{M-QAM})$, the solid curves show $\{e_{\mathrm{SE}}(t)\}_t$ calculated by SE iteration (33), and the hollow circles show $\{e_{\mathrm{GAMPD}}(t)\}_t$ obtained by averaging the experimental results from repeating GAMPD 100 times randomly (data bits and channel are regenerated). These two types of curves coincide with each other approximately. Fig. 7 clearly shows that $\{e^t_{\mathrm{GAMPD}} \leqslant e^t_{\mathrm{SE}}\}_t$ and $\{e^t_{\mathrm{GAMPD}} \approx e^t_{\mathrm{SE}}\}_t$. The tracking errors come from the inaccuracy of the "Gaussian approximation" exerted on $\mathbf{w}_{\mathrm{I}}$ in Section III. More details can be found in Section V-B.

In the second experiment, we verify that G and e^G_{GAMPD} can be approximated as T and e^T_{SE}, respectively. The simulation parameters are set as $U = 24$, $N_t = 4$, and 16-QAM. N_r is computed as αU. For every possible $(\alpha, \mathrm{SNR}_{\mathrm{sig}})$, GAMPD is repeated 100 times randomly. Every time, G and e^G_{GAMPD} are recorded. Their averaged values are shown in Fig. 8(a) and (c), respectively. T and e^T_{SE} are obtained by repeating the SE iteration (33) until the termination condition (34) is satisfied. They are shown in Fig. 8(b) and (d). By comparing Fig. 8(a) with Fig. 8(b) [and Fig. 8(c) with Fig. 8(d)], one can easily verify the conclusion at the beginning of this paragraph.

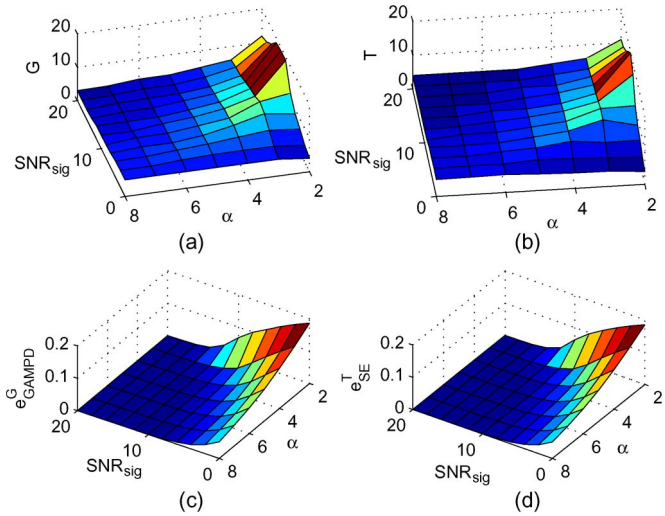


Fig. 8. Under different $\{\text{SNR}, \alpha\}s$, (a) and (c) show the averaged values of G and e^G_{GAMPD} from the simulation; (b) and (d) show T and e^T_{SE} obtained by repeating the SE iteration (33) until (34) is satisfied.

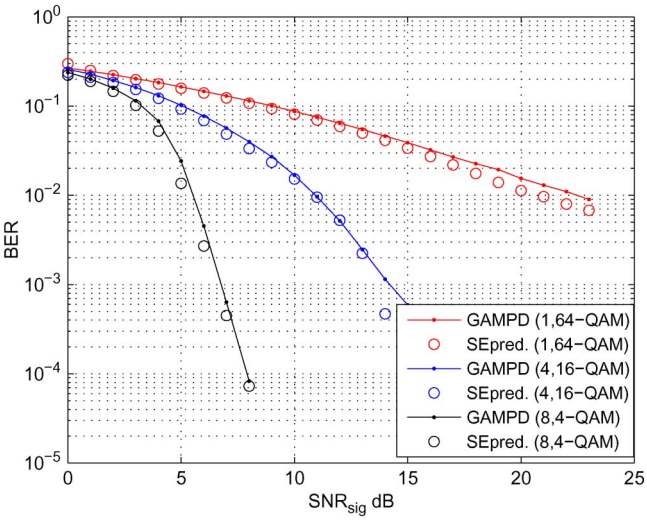


Fig. 10. BER predication. (Solid curves) BERs from simulation. (Hollow circles) BERs predicted by SE as in Section V-B. $N_r = 128$, and $U = 32$. The curves with the same color share the same $(N_t, M\text{-QAM})$.

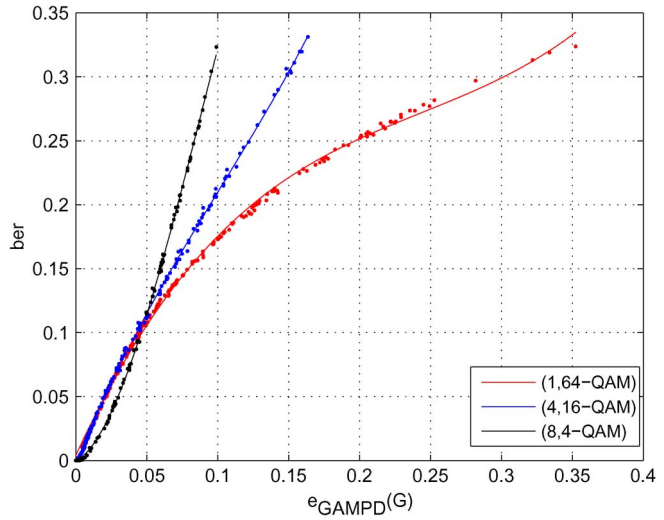


Fig. 9. Monotone relationship between ber and $e_{\text{GAMPD}}(G)$. The discrete points show $\{ber, e_{\text{GAMPD}}(G)\}$'s from random simulations where U, N_r, and SNR_{sig} are uniformly sampled from the ranges [12, 20], [64, 128], and [0, 15] dB, respectively. The solid lines show the fitted curves based on (41). $\{N_t, M\text{-QAM}\}$ are given in the legend.

The third experiment confirms the existence of the monotone relationship [i.e., (32)] between ber and $e_{\text{GAMPD}}(G)$. For clarity, Fig. 9 only shows the simulation results under four possible values of N_t, M (i.e., the bracketed numbers in the legends). GAMPD is randomly repeated 200 times. Every time, U, N_r, and SNR_{sig} are uniformly sampled from the ranges [12, 20], [64, 128], and [0 dB,15 dB], respectively. The values of $\{ber, e_{\text{GAMPD}}(G)$ are shown as the discrete points in Fig. 9. Then, they are exploited to approximate (32) as a third-order polynomial

$$ber = \max\Big\{a_0 + a_1 e_{\text{GAMPD}}(G) + a_2 e_{\text{GAMPD}}(G)^2 + a_3 e_{\text{GAMPD}}(G)^3, 0\Big\} \quad (41)$$

whose coefficients $\{a_k\}_{k=0}^3$ are obtained by the LS method. Equation (41) is shown as the solid curves in Fig. 9. Fig. 9 reveals that the discrete points concentrate around the solid curves closely. This indicates that the monotone relationship (32) does exist and can be well approximated by (41).

Finally, in the fourth experiment, the uncoded BERs of GAMPD are forecasted as in Section V-B. We fixed $N_r = 128$ and $U = 32$. Similar to Fig. 9, only four values of (N_t, M) are considered. In Fig. 10, the solid curves show the BER from simulations, and the hollow circles show the predicted BER by SE. They coincide with each other well, particularly when the BER is not very small (larger than 10^{-3}). Small BERs are failed to be predicted, but they are usually not interesting. For example, in the next section, we are interested in the BER 10^{-2}. However, the caution here is that the predicted BERs are generally lower than the actual BERs. In other words, the BER prediction algorithm is overoptimistic. This is caused by inaccuracy of the Gaussian approximation of $\mathbf{w}_{\text{I}}$. See Section V-B for more explanations.

D. Designing Energy-Efficient Massive SM-MIMO

Here, we show how to design energy-efficient massive SM-MIMO based on the algorithm in Table I in Section VI. Specially, we focus on how the channel coherence time (T_{coh}) and the BS circuit power $(p^{\text{BS}}_{\text{circ}})$ influence the choice of $\mathbb{P}_{\text{opt}} = \{U, N_t, M, p_s, N_r\}$. T_{coh} is focused because its value controls the tradeoff between the time length for CE $(N_t(U+1)L)$ and that for data transmission $(T_{\text{coh}} - N_t(U+1)L)$. $p^{\text{BS}}_{\text{circ}}$ is concerned because high BS circuit power consumption [40]–[42], [45] could be the bottleneck of the large-scale antenna systems.

The parameters in Table I are initialized as: $ber_{\text{target}} = 10^{-2}$, $U_{\min} = 30$, $U_{\max} = 50$, $\text{SNR}_{\min} = 0\,\text{dB}$, $\text{SNR}_{\text{step}} = 0.1\,\text{dB}$, $\text{SNR}_{\max} = 20\,\text{dB}$, $\alpha^{\min} = 1$, and $\alpha^{\max} = 40$. T_{coh} is varied from 35 to 600 SC symbols. $p^{\text{BS}}_{\text{circ}}$ is increased from 0.1 to 4 W. $p^{\text{UE}}_{\text{circ}}$ is fixed as 0.15 W.

Fig. 11(a)–(e) shows the optimized values of N_r, N_t, U, SNR_{sig}, and $EE_{\max}$ under different pairs of $(T_{\text{coh}}, p^{\text{BS}}_{\text{circ}})$'s.

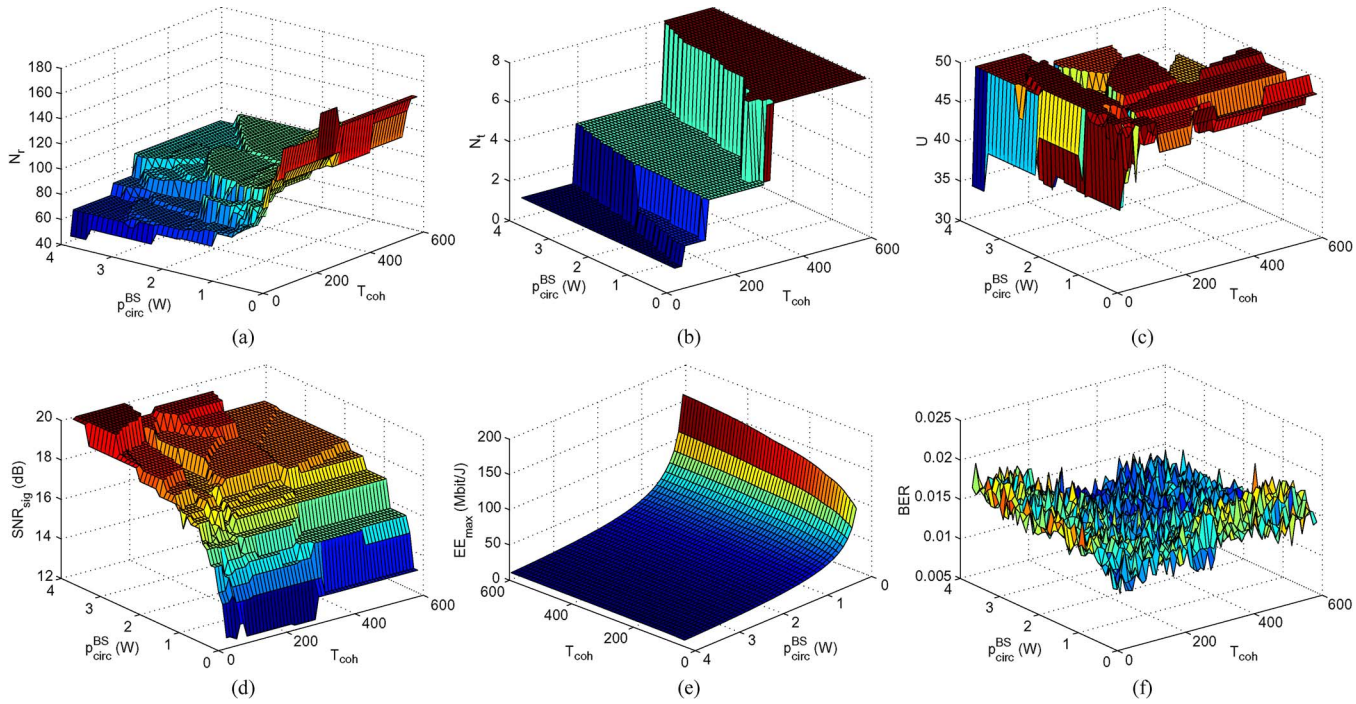


Fig. 11. (a)–(d) Optimal parameters N_r, N_t, U, and $\mathrm{SNR}_{\mathrm{sig}}$ under different pairs of $(T_{\mathrm{coh}}, p_{\mathrm{circ}}^{\mathrm{BS}})$. (e) Maximum EE $(EE_{\max})$ in Table I. (f) Simulated BERs in SM-MIMO with the optimized parameters (targeted BER is 0.01). The values of T_{coh} represent the numbers of SC symbols.

Note that the signal transmission power p_s can be calculated as $10^{\mathrm{SNR}_{\mathrm{sig}}/10}/U$. Numerical results indicate that 16-QAM is always adopted. Therefore, $M \equiv 16$ and is not shown in Fig. 11. Section VI has pointed out that, by adopting the system parameters' outputs in Table I, GAMPD can approximately satisfy the BER constraint $\mathrm{BER} \leq 10^{-2}$. To verify this conclusion, GAMPD is repeated 100 times randomly under different $(T_{\mathrm{coh}}, p_{\mathrm{circ}}^{\mathrm{BS}})$'s. The averaged BER performance is shown in Fig. 11(f).

Fig. 11(a)–(d) provide some important insights on system parameter design. First, under fixed N_t, with the decrease in the circuit power $p_{\mathrm{BS}}^{\mathrm{circ}}$, the number of the BS antennas will increase, and the transmission power at the UE side keeps decreasing. For example, when T_{coh} is equal to 275 SC symbols and $p_{\mathrm{BS}}^{\mathrm{circ}}$ is decreased from 4 to 0.1 W, N_t is fixed as 4 [see Fig. 11(b)], Fig. 11(a) shows that N_r gradually increases from 90 to 163, and Fig. 11(d) indicates that $\mathrm{SNR}_{\mathrm{sig}}$ is decreased from 19.4 to 12.1 dB (i.e., Up_s, the total transmission power of the UE, is decreased from 43.5 to 8.1 W). This observation is in accordance to a famous conclusion that the transmission power can be significantly decreased after adopting massive antennas at the BS [1], [2], [7]. It also indicates that the "massive" gain is more apparent when the circuit power consumption at the BS is low.

Second, with the increase in the channel coherence time and decrease in the BS circuit power, more TAs (i.e., higher order spatial constellation) should be adopted at the UE. This is due to the following two reasons. First, more TAs (i.e., higher N_t) indicate not only higher data transmission rate ($\log_2(N_t M)$ where $M = 16$) but also longer channel training time whose length is $N_t(U+1)L$. In (39), the rate gains from increasing N_t overweigh the negative effects from consuming more coherence time for CE only when T_{coh} is long enough. Second, with the decrease in $p_{\mathrm{BS}}^{\mathrm{circ}}$, Fig. 11(a) indicates that more BS antennas would be adopted; consequently, higher N_t could be supported. For example, Fig. 11(b) shows that, when $p_{\mathrm{BS}}^{\mathrm{circ}}$ is 3.5 (1) W, N_r is increased from 53 (74) to 99 (112). N_t is jumped from 1 to 4 when T_{coh} exceeds 185 (135) and further reaches 8 when T_{coh} is larger than 425 (345).

Third, the more TAs the UE have, the more BS antennas are required. In other words, compared with massive MIMO [1], [2], massive SM-MIMO calls for more BS antennas, which are used to remove the additional spatial ambiguity (from the random activation of the TAs). For example, based on Fig. 11(a) and (b), when N_t is 4 and 8, the averaged number of the BS antennas is 93 and 106, respectively. In contrast, when N_t is 1 (i.e., the case of massive MIMO [1], [2]), this number is only 68.

Fourth, to reach high EE, massive users are preferred to be supported simultaneously. Fig. 11(c) shows that, although the possible values of U come from [30, 50], the actual U exceeds 40 with probability 0.92, and its average value is 46. Similar phenomenon has been observed in massive MIMO [42], where the authors pointed out that supporting huge number of users is an EE optimal strategy.

After adopting our optimized parameters from Table I, Fig. 11(e) shows that massive SM-MIMO becomes more and more energy efficient with the increase in the coherence time and the decrease in the BS circuit power. For example, at $(T_{\mathrm{coh}}, p_{\mathrm{circ}}^{\mathrm{BS}}) = (35, 4\,\mathrm{W})$, the EE is 9 Mb/J. In contrast, at $(T_{\mathrm{coh}}, p_{\mathrm{circ}}^{\mathrm{BS}}) = (600, 0.1\,\mathrm{W})$, the EE is increased as 162 Mb/J. Fig. 11(f) shows that, under different $(T_{\mathrm{coh}}, p_{\mathrm{circ}}^{\mathrm{BS}})$'s, the simulated BERs concentrate in [0.011, 0.018] and are a little larger than the expected 0.01. This is not surprising because the

BER prediction method in Section V-B is overoptimistic due to the "Gaussian approximation" in Section III. Please refer to Section VI for more details. Fortunately, the actual BERs do not deviate from the targeted BER 0.01 very much. This deviation is still acceptable. For example, for every 1000 bits, the desired number of bit-flipping errors is 10 but could be 18 when actual BER is 0.018. Moreover. by slightly increasing $\mathrm{SNR}_{\mathrm{sig}}$, 0.01 could be reached. At the same time, SM-MIMO could be still energy efficient. Therefore, the optimized system parameters still provide important guidelines for designing energy-efficient massive SM-MIMO.

In summary, long-enough coherence time and low-enough BS circuit power consumption are the two most important prerequisites to the success of massive SM-MIMO. Table I provides an effective tool for designing energy-efficient SM-MIMO systems.

VIII. Conclusion

In this paper, we have researched the multiuser detection (MUD) and system parameter design in a single carrier (SC) massive SM-MIMO over frequency-selective channels. For MUD, we proposed a low-complexity generalized approximate message passing detector (GAMPD), which exploits both the prior probability distribution and sparsity of the transmitted signal and contains only parallelized matrix–vector multiplications as its most complex operations. By the tool of state evolution, we predict the uncoded BER performance of GAMPD. To design energy-efficient massive SM-MIMO, we constructed a practical algorithm to optimize the key system parameters under the constraint that some predefined BER could be reached approximately. Finally, numerical results validated the superiority of GAMPD and the effectiveness of the parameter optimization algorithm and indicated that long channel coherence time and low BS circuit power were two key prerequisites for the success of massive SM-MIMO.

Appendix A
Implementation of the Matrix Inversion in (16)

First, we introduce how to calculate $\left(\mathbf{H}^H\mathbf{H}+\lambda\mathbf{I}_{UN_tN}\right)^{-1}$, where λ is an arbitrary positive number, and $\mathbf{H}=\mathrm{TOEP}(\{\mathbf{h}^u_{n_rn_t}\}^{u\in[U]}_{n_r\in[N_r],\ n_t\in[N_t]})$ is defined as (A.1), shown at the bottom of the page.

In (A.1), $\mathbf{H}^u_{n_rn_t}=\mathrm{toep}(\mathbf{h}^u_{n_rn_t},N)$ is a $N\times N$ Toeplitz matrix toep$(\cdot)$ is defined by (6). $\mathbf{H}^u_{n_rn_t}$ can be decomposed as $\mathbf{H}^u_{n_rn_t}=\mathbf{F}_N^{\dagger}\mathbf{D}^u_{n_rn_t}\mathbf{F}_N$, where $\mathbf{F}_N$ is the N dimensional Fourier matrix containing $1/\sqrt{N}e^{-i(2\pi/N)mn}$ as its (m,n)th element, and $\mathbf{D}^u_{n_rn_t}=\mathrm{diag}(\mathbf{d}^u_{n_rn_t})$ is a diagonal matrix. $\mathbf{d}^u_{n_rn_t}$ is the FFT result of the zero-padded sequence $[(\mathbf{h}^u_{n_rn_t})^T\mathbf{0}^T_{N-L}]^T$.

Consequently, we have

$$\mathbf{H}=\left(\mathbf{I}_{N_r}\otimes\mathbf{F}_N^{\dagger}\right)\mathbf{D}\left(\mathbf{I}_{UN_t}\otimes\mathbf{F}_N\right) \tag{A.2}$$

where $\mathbf{D}$ is constructed by replacing $\mathbf{H}^u_{n_rn_t}$ in $\mathbf{H}$ as $\mathbf{D}^u_{n_rn_t}$. Through row and column permutations, $\mathbf{D}$ can be transformed as $\mathbf{D}=\mathbf{P}_L\mathbf{G}\mathbf{P}_R$. $\mathbf{P}_L$ and $\mathbf{P}_R$ are the left and right permutation matrices, respectively (their expressions are omitted here due to space constraints). $\mathbf{G}$ is a block diagonal matrix $\mathrm{diag}(\{\mathbf{G}_n\}_{n=0}^{N-1})$ where $\mathbf{G}_n\in\mathbb{C}^{N_r\times UN_t}$ contains $\mathbf{d}^u_{n_rn_t}(n)$ as its (n_r,uN_t+n_t)th element.

Consequently, we have

$$\mathbf{H}=\left(\mathbf{I}_{N_r}\otimes\mathbf{F}_N^{\dagger}\right)\mathbf{P}_L\mathbf{G}\mathbf{P}_R\left(\mathbf{I}_{UN_t}\otimes\mathbf{F}_N\right) \tag{A.3}$$

$$\mathbf{H}^H\mathbf{H}+\lambda\mathbf{I}_{UN_tN}=\left(\mathbf{I}_{UN_t}\otimes\mathbf{F}_N^{\dagger}\right)\times\mathbf{P}_R^T\left(\mathbf{G}^{\dagger}\mathbf{G}+\lambda\mathbf{I}_{UN_tN}\right)\mathbf{P}_R\left(\mathbf{I}_{UN_t}\otimes\mathbf{F}_N\right). \tag{A.4}$$

In the derivation of (A.3) and (A.4), we have exploited $\mathbf{P}_L^T\mathbf{P}_L=\mathbf{I}_{N_rN}$, $(\mathbf{F}_N\otimes\mathbf{I}_{N_r})(\mathbf{F}_N^{\dagger}\otimes\mathbf{I}_{Nr})=\mathbf{I}_{N_rN}$, $(\mathbf{I}_{UN_t}\otimes\mathbf{F}_N^{\dagger})\mathbf{P}_R^T\mathbf{P}_R(\mathbf{I}_{UN_t}\otimes\mathbf{F}_N)=\mathbf{I}_{UN_tN}$, $(\mathbf{A}\otimes\mathbf{B})^{\dagger}=\mathbf{A}^{\dagger}\otimes\mathbf{B}^{\dagger}$, and $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D})=\mathbf{AC}\otimes\mathbf{BD}$ [46].

Then, the inversion of (A.4) is

$$\left(\mathbf{H}^H\mathbf{H}+\lambda\mathbf{I}_{UN_tN}\right)^{-1}=\left(\mathbf{I}_{UN_t}\otimes\mathbf{F}_N^{\dagger}\right)\times\mathbf{P}_R^T\left(\mathbf{G}^{\dagger}\mathbf{G}+\lambda\mathbf{I}_{UN_tN}\right)^{-1}\mathbf{P}_R\left(\mathbf{I}_{UN_t}\otimes\mathbf{F}_N\right). \tag{A.5}$$

$\mathbf{G}$ is a block diagonal matrix; thus, $\left(\mathbf{G}^{\dagger}\mathbf{G}+\lambda\mathbf{I}_{UN_tN}\right)^{-1}$ can be decomposed as N small-scaled matrix inversions $\{(\mathbf{G}_n^{\dagger}\mathbf{G}_{\mathbf{n}}+\lambda\mathbf{I}_{UN_t})^{-1}\}_{n=0}^{N-1}$.

Finally, similar to (A.5), the matrix inversion $(\hat{\mathbf{H}}^H\hat{\mathbf{H}}+\sigma^2(1/p_t+1/p_s)\mathbf{I}_{UN_tN})^{-1}$ in (16) can be accelerated after replacing $\mathbf{H}$ as $\hat{\mathbf{H}}$, and λ as $\sigma^2(1/p_t+1/p_s)$.

$$\mathbf{H}=\mathrm{TOEP}\left(\left\{\mathbf{h}^u_{n_rn_t}\right\}^{u\in[U]}_{n_r\in[N_r],n_t\in[N_t]}\right)=\begin{bmatrix}\mathbf{H}^0_{00} & \cdots & \mathbf{H}^0_{0(N_t-1)} & \mathbf{H}^1_{00} & \cdots & \mathbf{H}^1_{0(N_t-1)} & \cdots & \mathbf{H}^{U-1}_{00} & \cdots & \mathbf{H}^{U-1}_{0(N_t-1)}\\ \mathbf{H}^0_{10} & \cdots & \mathbf{H}^0_{1(N_t-1)} & \mathbf{H}^1_{10} & \cdots & \mathbf{H}^1_{1(N_t-1)} & \cdots & \mathbf{H}^{U-1}_{10} & \cdots & \mathbf{H}^{U-1}_{1(N_t-1)}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \mathbf{H}^0_{(N_r-1)0} & \cdots & \mathbf{H}^0_{(N_r-1)(N_t-1)} & \mathbf{H}^1_{(N_r-1)0} & \cdots & \mathbf{H}^1_{(N_r-1)(N_t-1)} & \cdots & \mathbf{H}^{U-1}_{(N_r-1)0} & \cdots & \mathbf{H}^{U-1}_{(N_r-1)(N_t-1)}\end{bmatrix} \tag{A.1}$$

APPENDIX B
ACCELERATING THE MATRIX–VECTOR MULTIPLICATIONS IN (22) AND (25) OF GENERALIZED APPROXIMATE MESSAGE PASSING DETECTOR

In (22) and (25) of GAMPD, there exist two matrix–vector multiplications with the forms $\mathbf{Hx}$ and $\mathbf{H}^{\dagger}\mathbf{y}$ ($\mathbf{x} \in \mathbb{C}^{UN_tN}$ and $\mathbf{y} \in \mathbb{C}^{N_rN}$ are two arbitrary complex vectors). They can be accelerated by exploiting the Toeplitz-like structure of $\mathbf{H}$ as follows.

First, we focus on $\mathbf{Hx}$. Based on (A.3), we have

$$\mathbf{Hx} = \overbrace{\left(\mathbf{I}_{N_r} \otimes \mathbf{F}_N^{\dagger}\right) \underbrace{\mathbf{P}_L \mathbf{G} \mathbf{P}_R \underbrace{\left(\mathbf{I}_{UN_t} \otimes \mathbf{F}_N\right)\mathbf{x}}_{\text{step1}}}_{\text{step2}}}^{\text{step3}}.$$

step1 and *step3* can be executed by FFT and inverse FFT, respectively. In *step2*, the vector from *step1* is first permutated by $\mathbf{P}_R$ and then multiplied with the block diagonal matrix $\mathbf{G}$. Finally, the multiplication result is permutated by $\mathbf{P}_L$.

Similar to $\mathbf{Hx}$, $\mathbf{H}^{\dagger}\mathbf{y}$ can be also accelerated.

APPENDIX C
IRRELEVANCE OF T TO THE PROBLEM DIMENSIONS

T is the number of iterations needed by SE (33) to reach the termination condition (34). It will be proved to be irrelevant to the problem dimensions (the values of N_r and U).

First, we prove that, when $\mathrm{SNR}_{\mathrm{train}}$ is relatively large, there exists $\sigma_{SE}^2 \approx (10^{-\mathrm{SNR}_{\mathrm{train}}/10} + 10^{-\mathrm{SNR}_{\mathrm{sig}}/10})/\alpha$ where $\alpha = N_r/U$. Based on the definitions of $\mathrm{SNR}_{\mathrm{sig}}$ and $\mathrm{SNR}_{\mathrm{train}}$ in Section II-B and C, we have $p_s = 10^{\mathrm{SNR}_{\mathrm{sig}}/10}\sigma^2/U$ and $p_t = 10^{\mathrm{SNR}_{\mathrm{train}}/10}\sigma^2/U$. By substituting these two equations into the expression of σ_{SE}^2 in (33), we have

$$\begin{aligned}\sigma_{SE}^2 &= \frac{U}{N_r}\left(10^{-\frac{\mathrm{SNR}_{\mathrm{train}}}{10}} + 10^{-\frac{\mathrm{SNR}_{\mathrm{sig}}}{10}}\right)\frac{1}{\left(1+10^{-\frac{\mathrm{SNR}_{\mathrm{train}}}{10}}\right)} \\ &\approx \frac{1}{\alpha}\left(10^{-\frac{\mathrm{SNR}_{\mathrm{train}}}{10}} + 10^{-\frac{\mathrm{SNR}_{\mathrm{sig}}}{10}}\right) \qquad \text{(C.1)}\end{aligned}$$

where $\alpha = N_r/U$, and the approximation comes from the facts that $\mathrm{SNR}_{\mathrm{train}}$ is usually large (see Section III) and $10^{-\mathrm{SNR}_{\mathrm{train}}/10}$ is much smaller than 1 (e.g., $10^{-\mathrm{SNR}_{\mathrm{train}}/10} \approx 0.03$ under $\mathrm{SNR}_{\mathrm{train}} = 15$ dB).

Second, by substituting (C.1) into (33), (33) is approximated as

$$e_{\mathrm{SE}}(t+1) \approx \mathrm{E}\left(\mathrm{VAR}\left(x|r, \frac{N_t}{\alpha}e_{\mathrm{SE}}(t) + \frac{1}{\alpha}\left(10^{-\frac{\mathrm{SNR}_{\mathrm{train}}}{10}} + 10^{-\frac{\mathrm{SNR}_{\mathrm{sig}}}{10}}\right)\right)\right).$$

Because N_t, M-QAM, $\mathrm{SNR}_{\mathrm{train}}$, and $\mathrm{SNR}_{\mathrm{sig}}$ have been fixed, $\{e_{\mathrm{SE}}(t)\}_t$ are determined by α exclusively and so is T. In other words, under fixed α, T is irrelevant to the values of both U and N_r (i.e., the problem dimensions).

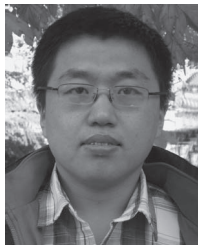
ACKNOWLEDGMENT

The authors thank the reviewers for their insightful comments that helped us improve the quality of this paper. S. Wang would like to thank Dr. Z. Gao for his warmhearted help with English writing.

REFERENCES

[1] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
[2] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
[3] F. B. Vodafone, R. W. Heath, Jr., A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.
[4] C. X. Wang *et al.*, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122–130, Feb. 2014.
[5] M. D. Renzo, H. Haas, A. Ghrayeb, S. Sugiura, and L. Hanzo, "Generalized MIMO: Challenges, opportunities, and implementation," *Proc. IEEE*, vol. 102, no.1, pp. 56–103, Jan. 2014.
[6] D. Tse and P. Viswanath. *Fundamentals of Wireless Communications*, Cambridge, U.K.: Cambridge Univ. Press, 2005.
[7] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.
[8] C. Shepard *et al.*, "Argos: Practical many-antenna base stations," in *ACM Int. Conf. Mobile Computing and Networking (MobiCom)*, Istanbul, Turkey, Aug. 2012.
[9] T. L. Marzetta, "Noncooperative cellular wireless with unlimited umbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
[10] A. Mohammadi and F. M. Ghannouchi, "Single RF front-end MIMO transceivers," *IEEE Commun. Mag.*, vol. 49, no. 12, pp. 104–109, Dec. 2011.
[11] A. Kalis, A. G. Kanatas, and C. B. Papadias, "A novel approach to MIMO transmission using a single RF front end," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 972–980, Aug. 2008.
[12] P. Raviteja, T. L. Narasimhan, and A. Chockalingam, "Large-scale multiuser SM-MIMO versus massive MIMO," in *Proc. ITA*, Feb. 2014, pp. 1–9.
[13] R. Y. Mesleh, H. Haas, S. Sinanovic, C. W. Ahn, and S. Yun, "Spatial modulation," *IEEE Trans. Veh. Technol.*, vol. 57, no. 4, pp. 2228–2241, Jul. 2008.
[14] M. Di Renzo, H. Haas, and P. M. Grant, "Spatial modulation for multiple-antenna wireless systems: A survey," *IEEE Commun. Mag.*, vol. 49, no. 12, pp. 182–191, Dec. 2011.
[15] S. Payami and F. Tufvesson, "Channel measurements and analysis for very large array systems at 2.6 GHz," in *Proc. 6th EUCAP*, Mar. 2012, pp. 433–437.
[16] S. Payami and F. Tufvesson, "Delay spread properties in a measured massive MIMO system at 2.6 GHz." in *Proc. IEEE 24th In. Symp. Pers., Indoor, Mobile Radio Commun.*, Sep. 2013, pp. 53–57.
[17] P. Som and A. Chockalingam, "Spatial modulation and space shift keying in single carrier communication," in *Proc. IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Sep. 2012, pp. 1991–1996.
[18] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," 2010, arXiv: 1010.5141[cs.IT].
[19] A. Younis, S. Sinanovic, M. D. Renzo, R. Mesleh, and H. Haas, "Generalized sphere decoding for spatial modulation," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2805–2815, Jul. 2013.
[20] C.-M. Yu, *et al.*, "Compressed sensing detector design for space shift keying in MIMO systems," *IEEE Commun. Lett.*, vol. 16, no. 10, pp. 1556–1559, Oct. 2012.
[21] 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.

[22] U. S. Kamilov, V. K. Goyal, and S. Rangan, "Message-passing dequantization with application to compressed sensing," *IEEE Trans. Signal Process.*, vol. 60, pp. 6270–6281, Dec. 2012.
[23] C. Studer and E. G. Larsson, "PAR-Aware large-scale multiuser MIMO-OFDM downlink," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 303–313, Feb. 2013.
[24] S. C. Wang, Y. Z. Li, and J. Wang, "Convex optimization based downlink precoding for large-scale MIMO," in *Proc. IEEE WCNC*, Apr. 2014, pp. 230–235.
[25] F. Fernandes, A. Ashikhmin, and T. L. Marzetta, "Inter-cell interference in noncooperative TDD large scale antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 192–201, Feb. 2013.
[26] S. Sugiura, and L. Hanzo, "Effects of channel estimation on spatial modulation," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 805–808, Dec. 2012.
[27] X. Wu, M. Di Renzo, and H. Haas, "Channel estimation for spatial modulation," in *Proc. IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun.*, 2013, pp. 306–310.
[28] S. Sesla, I. Toufik, and M. Baker, *LTE—The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. Hoboken, NJ, USA: Wiley, 2011, pp. 330–332.
[29] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[30] A. V. Oppenheim, A. S. Willsky, and S. Hamid, *Signals and Systems*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1996.
[31] I. Barhumi, G. Leus, and M. Moonen, "Optimal training design for MIMO OFDM systems in mobile wireless channels," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp. 1615–1624, Jun. 2003.
[32] H. Minn and N. Al-Dhahir, "Optimal training signals for MIMO OFDM channel estimation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 5, pp. 1158–1168, May 2006.
[33] S. C. Wang, Y. Z. Li, and J. Wang, "Low-complexity multiuser detection for uplink large-scale MIMO," in *Proc. IEEE WCNC*, Apr. 2014, pp. 236–241.
[34] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," *J. Phys. A, Math. Gen.*, vol. 36, no. 43, Oct. 2003, 11111.
[35] D. Guo and C.-C. Wang, "Random sparse linear systems observed via arbitrary channels: A decoupling principle," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2007, pp. 946–950.
[36] F. Krzakala, M. Mezard, F. Sausset, Y. Sun, and L. Zdeborova, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," *J. Statist. Mech., Theory Experiment*, vol. 2012, Aug. 2012, P08009.
[37] S. C. Wang, Y. Z. Li, and J. Wang, "Sparse Bayesian learning for compressed sensing under measurement matrix uncertainty," in *Proc. WCSP*, Oct. 2013, pp. 1–5.
[38] Z. Yang, L. Xie, and C. Zhang, " Bayesian compressed sensing with new sparsity-inducing prior," in *Proc. Inf. Theory,* 2012, pp. 1–11. [Online]. Available: http://arxiv.org/pdf/1208.6464
[39] Y. Zhang *et al.*, "Joint antenna and user selection algorithm for uplink of multiuser MIMO systems using sequential Monte Carlo optimization," in *Proc. IEEE/SP 14th Workshop SSP*, Aug. 2007, pp. 493–496.
[40] D. Ha, K. Lee, and J. Kang, "Energy efficiency analysis with circuit power consumption in massive MIMO systems," in *Proc. IEEE 24th Int. Symp. PIMRC*, 2013, pp. 938–942.
[41] S. C. Wang, Y. Z. Li, and J. Wang, "Multiuser detection for uplink large-scale MIMO under one-bit quantization," in *Proc. IEEE ICC*, Jun. 2014, pp. 4460–4465.
[42] E. Bjornson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" in *Proc. Inf. Theory*, Mar. 2014, pp. 1–16. [Online]. Available: http://arxiv.org/abs/1403.6150
[43] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," *IEEE J. Sel. Areas Commun.*, pp. 1089–1098, vol. 22, no. 6, Aug. 2004.
[44] J. Jeganathan, A. Ghrayeb, L. Szczecinski, A. Ceron, "Space shift keying modulation for MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3692–3703, Jul. 2009.
[45] S. C. Wang, Y. Z. Li, and J. Wang, "Convex optimization based multiuser detection for uplink large-scale MIMO under low-resolution quantization," in *IEEE ICC*, Jun. 2014, pp. 4789–4794.
[46] K. B. Petersen and M. S. Pedersen, The Matrix Cookbook. Feb. 2006. [Online] Available: http://www.mit.edu/~wingated/stuff_i_use/matrix_cookbook.pdf
[47] N. Serafimovski1, S. Sinanovic, M. Di Renzo, and H. Haas, "Multiple access spatial modulation," *EURASIP J. Wireless Commun. Netw.*, vol. 2012, no. 299, pp. 1–20, Sep. 2012.
[48] S. Wu, *et al.*, "Low-complexity iterative detection for large-scale multiuser MIMO-OFDM systems using approximate message passing," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 902–915, Oct. 2014.
[49] J. T. Parker, V. Cevher, and P. Schniter, "Compressive sensing under matrix uncertainties: an approximate message passing approach," in *Proc. 45th Asilomar Conf. Signals Syst. Comput.*, Pacific Grove, CA, USA, 2011, pp. 804–808.
[50] J. 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.



Shengchu Wang received the B.S. degree in electronic engineering from Beijing Institute of Technology, Beijing, China, in 2008, and the M.S. degree in electronic engineering from Tsinghua University, Beijing, in 2011, where he is currently working toward the Ph.D. degree

His research studies focus on designing baseband algorithms for massive multiple-input–multiple-output based on compressed sensing and convex optimization.

Yunzhou Li (M'14) received the Ph.D. degree from Tsinghua University, Beijing, China, in 2004.

He is currently an Associate Professor with Tsinghua University. He is mainly focused on signal processing technologies in wireless and mobile communications, including spatial-time signal processing, channel estimation, multiuser detection, and synchronization algorithms for code-division–multiple-access/orthogonal frequency-division multiplexing systems. He is also interested in the analysis, optimization design, and enhancement of cellular systems and wireless local area networks.

Ming Zhao (M'14) received the B.S. and Ph.D. degrees in electronics engineering from Tsinghua University, Beijing, China, in 1993 and 1998, respectively.

He is currently a Professor with the Research Institute of Information Technology, Tsinghua University. His current interests include wireless mobile communication, distributed wireless communication systems, and the theory and technology of broadband wireless communication.

Jing Wang (M'14) received the B.S. and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1983 and 1986, respectively.

Since 1986, he has been with the faculty of Tsinghua University, where he currently is a Professor with the School of Information Science and Technology. He serves as the Vice Director of Tsinghua National Laboratory for Information Science and Technology. He has published more than 150 conference and journal papers. His research interests are in the area of wireless communications, including transmission and networking technologies of Fourth-Generation (4G)/Beyond 4G.