PAR-Aware Large-Scale Multi-User MIMO-OFDM Downlink

Christoph Studer and Erik G. Larsson

Linköping University Post Print

N.B.: When citing this work, cite the original article.

PAR-Aware Large-Scale Multi-User MIMO-OFDM Downlink

Christoph Studer, *Member, IEEE*, and Erik G. Larsson, *Senior Member, IEEE*

Abstract—We investigate an orthogonal frequency-division multiplexing (OFDM)-based downlink transmission scheme for large-scale multi-user (MU) multiple-input multiple-output (MIMO) wireless systems. The use of OFDM causes a high peak-to-average (power) ratio (PAR), which necessitates expensive and power-inefficient radio-frequency (RF) components at the base station. In this paper, we present a novel downlink transmission scheme, which exploits the massive degrees-of-freedom available in large-scale MU-MIMO-OFDM systems to achieve low PAR. Specifically, we propose to jointly perform MU precoding, OFDM modulation, and PAR reduction by solving a convex optimization problem. We develop a corresponding fast iterative truncation algorithm (FITRA) and show numerical results to demonstrate tremendous PAR-reduction capabilities. The significantly reduced linearity requirements eventually enable the use of low-cost RF components for the large-scale MU-MIMO-OFDM downlink.

Index Terms—Multi-user wireless communication, multiple-input multiple-output (MIMO), orthogonal frequency-division multiplexing (OFDM), peak-to-average (power) ratio (PAR) reduction, precoding, convex optimization.

I. INTRODUCTION

LARGE-SCALE multiple-input multiple-output (MIMO) wireless communication is a promising means to meet the growing demands for higher throughput and improved quality-of-service of next-generation multi-user (MU) wireless communication systems [2]. The vision is that a large number of antennas at the base-station (BS) would serve a large number of users concurrently and in the same frequency band, but with the number of BS antennas being much larger than the number of users [3], say a hundred antennas serving ten users. Large-scale MIMO systems also have the potential to reduce the operational power consumption at the transmitter and enable the use of low-complexity schemes for suppressing MU interference (MUI). All these properties render large-scale MIMO a promising technology for next-generation wireless communication systems.

While the theoretical aspects of large-scale MU-MIMO systems have gained significant attention in the research community, e.g., [2]–[6], much less is known about practical transmission schemes. As pointed out in [7], practical implementations

of large-scale MIMO systems will require the use of low-cost and low-power radio-frequency (RF) components. To this end, reference [7] proposed a novel MU precoding scheme for frequency-flat channels, which relies on per-antenna constant-envelope (CE) transmission to enable efficient implementation using non-linear RF components. Moreover, the CE precoder of [7] forces the peak-to-average (power) ratio (PAR) to unity, which is not necessarily optimal as in practice there is always a trade-off between PAR, error-rate performance, and power-amplifier efficiency.

Practical wireless channels typically exhibit frequency-selective fading and a low-PAR precoding solution suitable for such channels would be desirable. Preferably, the solution should be such that the complexity required in each (mobile) terminal is small (due to stringent area and power constraints), whereas heavier processing could be afforded at the BS. Orthogonal frequency-division multiplexing (OFDM) [8] is an attractive and well-established way of dealing with frequency-selective channels. In addition to simplifying the equalization at the receiver, OFDM also facilitates per-tone power and bit allocation, scheduling in the frequency domain, and spectrum shaping. However, OFDM is known to suffer from a high PAR [9], which necessitates the use of linear RF components (e.g., power amplifiers) to avoid out-of-band radiation and signal distortions. Unfortunately, linear RF components are, in general, more costly and less power efficient than their non-linear counterparts, which would eventually result in exorbitant costs for large-scale BS implementations having hundreds of antennas. Therefore, it is of paramount importance to reduce the PAR of OFDM-based large-scale MU-MIMO systems to facilitate corresponding low-cost and low-power BS implementations.

To combat the challenging linearity requirements of OFDM, a plethora of PAR-reduction schemes have been proposed for point-to-point single-antenna and MIMO wireless systems, e.g., [10]–[16]. For MU-MIMO systems, however, a straightforward adaptation of these schemes is non-trivial, mainly because MU systems require the removal of MUI using a precoder [17]. PAR-reduction schemes suitable for the MU-MISO and MU-MIMO downlink were described in [18] and [19], respectively, and rely on Tomlinson-Harashima precoding. Both schemes, however, require specialized signal processing in the (mobile) terminals (e.g., modulo reduction), which prevents their use in conventional MIMO-OFDM systems, such as IEEE 802.11n [20].

A. Contributions

In this paper, we develop a novel downlink transmission scheme for large-scale MU-MIMO-OFDM wireless systems, which only affects the signal processing at the BS while leaving the processing required at each terminal untouched. The key idea of the proposed scheme is to exploit the excess of degrees-of-freedom (DoF) offered by equipping the BS with a large number of antennas and to *jointly* perform MU precoding, OFDM modulation, and PAR reduction, referred to as PMP in the remainder of the paper. Our contributions can be summarized as follows:

- We formulate PMP as a convex optimization problem, which jointly performs MU precoding, OFDM modulation, and PAR reduction at the BS.
- We develop and analyze a novel optimization algorithm, referred to as fast iterative truncation algorithm (FITRA), which is able to find the solution to PMP efficiently for the (typically large) dimensions arising in large-scale MU-MIMO-OFDM systems.
- We present numerical simulation results to demonstrate the capabilities of the proposed MU-MIMO-OFDM downlink transmission scheme. Specifically, we analyze the trade-offs between PAR, error-rate performance, and out-of-band radiation, and we present a comparison with conventional precoding schemes.

B. Notation

Lowercase boldface letters stand for column vectors and uppercase boldface letters designate matrices. For a matrix $\mathbf{A}$, we denote its transpose, conjugate transpose, and largest singular value by $\mathbf{A}^T$, $\mathbf{A}^H$, and $\sigma_{\max}(\mathbf{A})$, respectively; $\mathbf{A}^\dagger = \mathbf{A}^H \left(\mathbf{A}\mathbf{A}^H\right)^{-1}$ stands for the pseudo-inverse of $\mathbf{A}$ and the entry in the kth row and ℓth column is $[\mathbf{A}]_{k,\ell}$. The $M \times M$ identity matrix is denoted by $\mathbf{I}_M$, the $M \times N$ all-zeros matrix by $\mathbf{0}_{M\times N}$, and $\mathbf{F}_M$ refers to the $M \times M$ discrete Fourier transform (DFT) matrix. The kth entry of a vector $\mathbf{a}$ is designated by $[\mathbf{a}]_k$; the Euclidean (or ℓ_2) norm is denoted by $\|\mathbf{a}\|_2$, $\|\mathbf{a}\|_\infty = \max_k |[\mathbf{a}]_k|$ stands for the ℓ_∞-norm, and the $\ell_{\widetilde{\infty}}$-norm [21] is defined as $\|\mathbf{a}\|_{\widetilde{\infty}} = \max\{\|\Re\{\mathbf{a}\}\|_\infty, \|\Im\{\mathbf{a}\}\|_\infty\}$ with $\Re\{\mathbf{a}\}$ and $\Im\{\mathbf{a}\}$ representing the real and imaginary part of $\mathbf{a}$, respectively. Sets are designated by upper-case calligraphic letters; the cardinality and complement of the set $\mathcal{T}$ is $|\mathcal{T}|$ and $\mathcal{T}^c$, respectively. For $x \in \mathbb{R}$ we define $[x]^+ = \max\{x, 0\}$.

C. Outline of the Paper

The remainder of the paper is organized as follows. Section II introduces the system model and summarizes important PAR-reduction concepts. The proposed downlink transmission scheme is detailed in Section III and the fast iterative truncation algorithm (FITRA) is developed in Section IV. Simulation results are presented in Section V and we conclude in Section VI.

II. PRELIMINARIES

We start by introducing the system model that is considered in the remainder of the paper. We then provide a brief overview of (linear) MU precoding schemes and, finally, we summarize the fundamental PAR issues arising in OFDM-based communication systems.

A. System Model

We consider an OFDM-based MU-MIMO downlink scenario as depicted in Fig. 1. The BS is assumed to have a significantly larger number of transmit antennas N than the number $M \ll N$ of independent terminals (users); each terminal is equipped with a single antenna only. The signal vector $\mathbf{s}_w \in \mathcal{O}^M$ contains information for each of the M users, where $w = 1, \ldots, W$ indexes the OFDM tones, W corresponds to the total number of OFDM tones, $\mathcal{O}$ represents the set of scalar complex-valued constellations, and $[\mathbf{s}_w]_m \in \mathcal{O}$ corresponds to the symbol at tone w to be transmitted to user m.[1] We normalize the symbols to satisfy $\mathbb{E}\{|[\mathbf{s}_w]_m|^2\} = 1/M$. To shape the spectrum of the transmitted signals, OFDM systems typically specify certain unused tones (e.g., at both ends of the spectrum [8]). Hence, we set $\mathbf{s}_w = \mathbf{0}_{M\times 1}$ for $w \in \mathcal{T}^c$ where $\mathcal{T}$ designates the set of tones used for data transmission.

In order to remove MUI, the signal vectors $\mathbf{s}_w$, $\forall w$ are passed through a precoder, which generates W vectors $\mathbf{x}_w \in \mathbb{C}^N$ according to a given precoding scheme (see Section II-B). Since precoding causes the transmit power $P = \sum_{w=1}^W \|\mathbf{x}_w\|_2^2$ to depend on the signals $\mathbf{s}_w$, $\forall w$ and the channel state, we normalize the precoded vectors $\mathbf{x}_w$, $\forall w$ prior to transmission as

$$\hat{\mathbf{x}}_w = \mathbf{x}_w / \sqrt{\sum_{w=1}^W \|\mathbf{x}_w\|_2^2}, \ w = 1, \ldots, W, \qquad (1)$$

which ensures unit transmit power. We emphasize that this normalization is an essential step in practice (i.e., to meet regulatory power constraints). To simplify the presentation, however, the normalization is omitted in the description of the precoders to follow (but normalization employed in all simulation results shown in Section V). Hence, in what follows $\mathbf{x}_w$ and $\hat{\mathbf{x}}_w$ are treated interchangeably.

The (normalized) vectors $\mathbf{x}_w$, $\forall w$ are then re-ordered (from user orientation to transmit-antenna orientation) according to the following one-to-one mapping:

$$[\mathbf{x}_1 \cdots \mathbf{x}_W] = [\mathbf{a}_1 \cdots \mathbf{a}_N]^T. \qquad (2)$$

Here, the W-dimensional vector $\mathbf{a}_n$ corresponds to the (frequency-domain) signal to be transmitted from the nth antenna. The time-domain samples are obtained by applying the inverse DFT (IDFT) according to $\hat{\mathbf{a}}_n = \mathbf{F}_W^H \mathbf{a}_n$ followed by parallel-to-serial (P/S) conversion. Prior to modulation and transmission over the wireless channel, a cyclic prefix (CP) is added to the (time-domain) samples $\hat{\mathbf{a}}_n$, $\forall n$ to avoid ISI [8].

To simplify the exposition, we specify the input-output relation of the wireless channel in the frequency domain only. Concretely, we consider[2]

$$\mathbf{y}_w = \mathbf{H}_w \mathbf{x}_w + \mathbf{n}_w, \quad w = 1, \ldots, W, \qquad (3)$$

---

[1]For the sake of simplicity of exposition, we employ the same constellation for all users. An extension to the general case where different constellations are used by different users is straightforward.

[2]We assume perfect synchronization and a CP that is longer than the maximum excess delay of the frequency-selective channel.
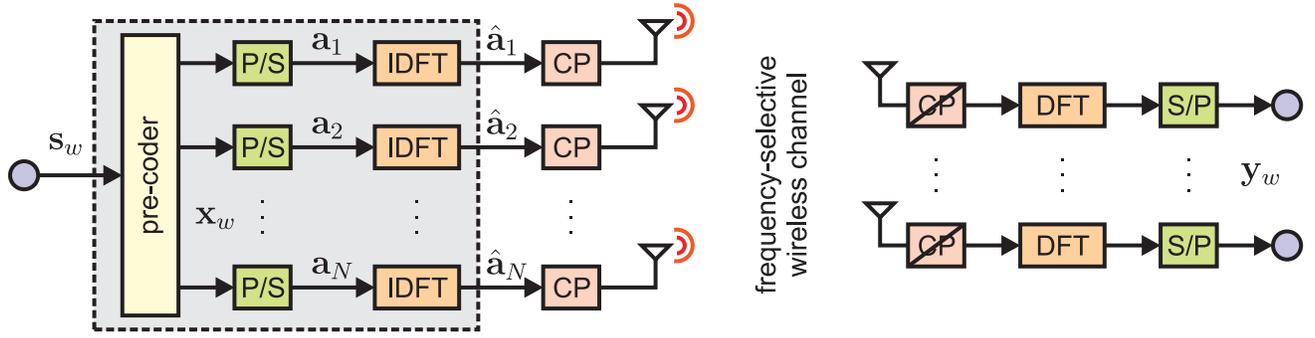
Fig. 1. Large-scale MU-MIMO-OFDM downlink (left: BS with N transmit antennas; right: M independent single-antenna terminals). The proposed downlink transmission scheme, referred to as PMP, combines MU precoding, OFDM modulation, and PAR reduction (highlighted by the dashed box in the BS).

where $\mathbf{y}_w$ denotes the wth receive vector, $\mathbf{H}_w \in \mathbb{C}^{M \times N}$ represents the MIMO channel matrix associated with the wth OFDM tone, and $\mathbf{n}_w$ is an M-vector of i.i.d. complex Gaussian noise with zero-mean and variance N_0 per entry. The average receive signal-to-noise-ratio (SNR) is defined by $\mathsf{SNR} = 1/N_0$. Finally, each of the M user terminals performs OFDM demodulation to obtain the received (frequency-domain) signals $[\mathbf{y}_w]_m$, $w = 1, \ldots, W$ (see Fig. 1).

B. MU Precoding Schemes

In order to avoid MUI, precoding must be employed at the BS. To this end, we assume the channel matrices $\mathbf{H}_w$, $\forall w$ to be known perfectly at the transmit-side.[3] Linear precoding now amounts to transmitting $\mathbf{x}_w = \mathbf{G}_w \mathbf{s}_w$, where $\mathbf{G}_w \in \mathbb{C}^{N \times M}$ is a suitable precoding matrix. One of the most prominent precoding schemes is least-squares (LS) precoding (or linear zero-forcing precoding), which corresponds to $\mathbf{G}_w = \mathbf{H}_w^\dagger$. Since $\mathbf{H}_w \mathbf{H}_w^\dagger = \mathbf{I}_M$, transmitting $\mathbf{x}_w = \mathbf{H}_w^\dagger \mathbf{s}_w$ perfectly removes all MUI, i.e., it transforms (3) into M independent single-stream systems $\mathbf{y}_w = \mathbf{s}_w + \mathbf{n}_w$. Note that LS precoding is equivalent to transmitting the solution $\dot{\mathbf{x}}_w$ to the following convex optimization problem:

$$\text{(LS)} \quad \underset{\tilde{\mathbf{x}}}{\text{minimize}} \ \|\tilde{\mathbf{x}}\|_2 \quad \text{subject to } \mathbf{s}_w = \mathbf{H}_w \tilde{\mathbf{x}}.$$

This formulation inspired us to state the MU-MIMO-OFDM downlink transmission scheme proposed in Section III as a convex optimization problem.

Several other linear precoding schemes have been proposed in the literature, such as matched-filter (MF) precoding, minimum-mean square-error (MMSE) precoding [17], or more sophisticated non-linear schemes, such as dirty-paper coding [22]. In the remainder of the paper, we will occasionally consider MF precoding, which corresponds to $\mathbf{G}_w = \mathbf{H}_w^H$. Since $\mathbf{H}_w \mathbf{H}_w^H$ is, in general, not a diagonal matrix, MF is normally unable to remove the MUI. Nevertheless, MF precoding was shown in [6] to be competitive for large-scale MIMO in some operating regimes and in [3] to perfectly remove MUI in the large-antenna limit, i.e., when $N \to \infty$.

C. Peak-to-Average Ratio (PAR)

The IDFT required at the transmitter causes the OFDM signals $\hat{\mathbf{a}}_n$, $\forall n$ to exhibit a large dynamic range [8]. Such signals are susceptible to non-linear distortions (e.g., saturation or clipping) typically induced by real-world RF components. To avoid unwanted out-of-band radiation and signal distortions altogether, linear RF components and PAR-reduction schemes are key to successfully deploy OFDM in practical systems.

1) PAR Definition: The dynamic range of the transmitted OFDM signals is typically characterized through the peak-to-average (power) ratio (PAR). Since many real-world RF-chain implementations process and modulate the real and imaginary part independently, we define the PAR at the nth transmit antenna as[4]

$$\mathsf{PAR}_n = \frac{2W \|\hat{\mathbf{a}}_n\|_{\widetilde{\infty}}^2}{\|\hat{\mathbf{a}}_n\|_2^2}. \tag{4}$$

As a consequence of standard vector-norm relations, (4) satisfies $1 \le \mathsf{PAR}_n \le 2W$. Here, the upper bound corresponds to the worst-case PAR and is achieved for signals having only a single (real or imaginary) non-zero entry. The lower bound corresponds to the best case and is realized by transmit vectors whose (real and imaginary) entries have constant modulus. To minimize distortion due to hardware non-linearities, the transmit signals should have a PAR that is close to one; this can either be achieved by CE transmission [7] or by using sophisticated PAR-reduction schemes.

2) PAR-Reduction Schemes for OFDM: Prominent PAR-reduction schemes for single-antenna communication systems are selected mapping (SM) [10], partial transmit sequences [11], active constellation extension (ACE) [12], and tone reservation (TR) [13], [15]. PAR-reduction schemes for point-to-point MIMO systems mostly rely on SM or ACE and have been described in, e.g., [14], [16]. For the MU-MIMO downlink, a method relying on Tomlinson-Harashima precoding and lattice reduction has been introduced recently in [19]; this method, however, requires dedicated signal-processing

---

[3]In large-scale MU-MIMO systems, channel-state information at the transmitter would probably be acquired through pilot-based training in the uplink and by exploiting reciprocity of the wireless channel [2], [3].

[4]Note that alternative PAR definitions exist in the literature, e.g., using the ℓ_∞-norm in the nominator instead of the $\ell_{\widetilde{\infty}}$-norm (and W instead of $2W$). Nevertheless, the relation $\frac{1}{2}\|\hat{\mathbf{a}}_n\|_\infty^2 \le \|\hat{\mathbf{a}}_n\|_{\widetilde{\infty}}^2 \le \|\hat{\mathbf{a}}_n\|_\infty^2$ shown in [21, Eq. 12] ensures that reducing the PAR as defined in (4) also reduces an ℓ_∞-norm-based PAR definition (and vice versa). Moreover, the theory and algorithms presented in this paper can, for example, be formulated to directly reduce an ℓ_∞-norm-based PAR.

algorithms at both ends of the wireless link (e.g., modulo reduction in the receiver). In contrast, the transmission scheme developed next aims at reducing the PAR by *only* exploiting the excess of transmit antennas available at the BS. This approach has the key advantage of being *transparent* to the receivers, i.e., it does not require any special signal-processing algorithms in the (mobile) terminals. Hence, the proposed precoding scheme can be deployed in existing MIMO-OFDM systems for which channel-state information is available at the transmitter, such as IEEE 802.11n [20].

III. DOWNLINK TRANSMISSION SCHEME

The main idea of the downlink transmission scheme developed next is to *jointly* perform MU precoding, OFDM modulation, and PAR reduction, by exploiting the DoF available in large-scale MU-MIMO systems. To convey the basic idea and to characterize its fundamental properties, we start by considering a simplified MIMO system. We then present the MU-MIMO-OFDM downlink transmission scheme in full detail and conclude by discussing possible extensions.

A. Basic Idea and Fundamental Properties

To convey the main idea of the proposed precoding method, let us consider an OFDM-free (narrow-band, flat-channel) MU-MIMO system with the real-valued input-output relation $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$ and an $M \times N$ channel matrix satisfying $M < N$. To eliminate MUI, the transmit-vector $\mathbf{x}$ must satisfy the precoding constraint $\mathbf{s} = \mathbf{H}\mathbf{x}$, which ensures that $\mathbf{y} = \mathbf{s} + \mathbf{n}$ when transmitting the vector $\mathbf{x}$. Since $M < N$, the equation $\mathbf{s} = \mathbf{H}\mathbf{x}$ is *underdetermined*; this implies that there are, in general, infinitely many solutions $\mathbf{x}$ satisfying the precoding constraint. Our hope is now to find a suitable vector $\dot{\mathbf{x}}$ having a small dynamic range (or low PAR).

A straightforward approach that reduces the dynamic range is to transmit the solution $\dot{\mathbf{x}}$ of the following optimization problem:

$$\text{(P-DYN)} \begin{cases} \underset{\alpha, \beta, \tilde{\mathbf{x}}}{\text{minimize}} & \alpha - \beta \\ \text{subject to} & \mathbf{s} = \mathbf{H}\tilde{\mathbf{x}}, \ \beta \le |[\tilde{\mathbf{x}}]_i| \le \alpha, \forall i. \end{cases}$$

Unfortunately, the second constraint $\beta \le |[\tilde{\mathbf{x}}]_i| \le \alpha, \forall i$ causes this problem to be non-convex and hence, finding the solution of (P-DYN) with efficient algorithms seems to be difficult.

1) Convex Relaxation: To arrive at an optimization problem that reduces the dynamic range and can be solved efficiently, we relax (P-DYN). Specifically, $\beta \le |[\tilde{\mathbf{x}}]_i| \le \alpha$ is replaced by $|[\tilde{\mathbf{x}}]_i| \le \alpha$, which leads to the following *convex* optimization problem:

$$\text{(P-INF)} \quad \underset{\tilde{\mathbf{x}}}{\text{minimize}} \ \|\tilde{\mathbf{x}}\|_\infty \quad \text{subject to } \mathbf{s} = \mathbf{H}\tilde{\mathbf{x}}.$$

Intuitively, as (P-INF) minimizes the magnitude of the largest entry of $\tilde{\mathbf{x}}$, we can expect that its solution $\dot{\mathbf{x}}$ exhibits low PAR. In fact, (P-INF) has potentially smaller PAR than a transmit vector resulting from LS precoding. To show this, we note that $\|\dot{\mathbf{x}}\|_\infty \le \|\mathbf{H}^\dagger \mathbf{s}\|_\infty$, where $\dot{\mathbf{x}}$ is the minimizer of (P-INF) and $\mathbf{H}^\dagger \mathbf{s}$ corresponds to the LS-precoded vector. Since $\mathbf{H}^\dagger \mathbf{s}$ is the ℓ_2-norm minimizer, we have $\|\mathbf{H}^\dagger \mathbf{s}\|_2 \le \|\dot{\mathbf{x}}\|_2$ and,

consequently, the PAR-levels of (P-INF) and of LS precoding satisfy

$$\mathsf{PAR}_{\text{P-INF}} = \frac{N\|\dot{\mathbf{x}}\|_\infty^2}{\|\dot{\mathbf{x}}\|_2^2} \le \frac{N\|\mathbf{H}^\dagger \mathbf{s}\|_\infty^2}{\|\mathbf{H}^\dagger \mathbf{s}\|_2^2} = \mathsf{PAR}_{\text{LS}},$$

which implies that the PAR associated with (P-INF) cannot be larger than that of LS precoding. We confirm this observation in Section V, where the proposed downlink transmission scheme is shown to achieve substantially lower PAR than for LS precoding.

2) Benefits of Large-Scale MIMO: To characterize the benefit of having a large number of transmit antennas at the BS on the PAR when using (P-INF), we first restate a key result from [23].

Proposition 1 ([23, Prop. 1]): Let $\mathbf{H}$ have full (column) rank and $1 \le M < N$. Generally, the solution $\dot{\mathbf{x}}$ to (P-INF) has $N - M + 1$ entries with magnitude equal to $\|\dot{\mathbf{x}}\|_\infty$. The $M - 1$ remaining entries might have smaller magnitude.

With this result, we are able to derive the following upper bound on the PAR on the solution $\dot{\mathbf{x}}$ to (P-INF):

$$\mathsf{PAR}_{\text{P-INF}} = \frac{N\|\dot{\mathbf{x}}\|_\infty^2}{\|\dot{\mathbf{x}}\|_2^2} \le \frac{N}{N - M + 1}. \tag{5}$$

Here, the following inequality is an immediate consequence of Proposition 1, i.e., we have

$$\begin{aligned} \|\dot{\mathbf{x}}\|_2^2 &= \sum_{\mathcal{X}} \|\dot{\mathbf{x}}\|_\infty^2 + \sum_{\mathcal{X}^c} |[\dot{\mathbf{x}}]_i|^2 \\ &\ge \sum_{\mathcal{X}} \|\dot{\mathbf{x}}\|_\infty^2 = (N - M + 1)\|\dot{\mathbf{x}}\|_\infty^2, \end{aligned}$$

where $\mathcal{X}$ is the set of indices associated with the $N - M + 1$ entries of $\dot{\mathbf{x}}$ for which $|[\dot{\mathbf{x}}]_i| = \|\dot{\mathbf{x}}\|_\infty$. It is now key to realize that for a constant number of users M and in the large-antenna limit $N \to \infty$, the bound (5) implies that $\mathsf{PAR}_{\text{P-INF}} \to 1$. Hence, for systems having a significantly larger number of transmit antennas than users—as is the case for typical large-scale MU-MIMO systems [2], [3], [5], [6]—a precoder that implements (P-INF) is able to achieve a PAR that is arbitrarily close to unity. This means that in the large-antenna limit of $N \to \infty$, (P-INF) yields constant-envelope signals, while being able to perfectly eliminate the MUI.

B. Joint Precoding, Modulation, and PAR Reduction (PMP)

The application of (P-INF) to each time-domain sample *after* OFDM modulation would reduce the PAR but, unfortunately, would *no longer* allow the equalization of ISI using conventional OFDM demodulation. In fact, such a straightforward PAR-reduction approach would necessitate the deployment of sophisticated equalization schemes in each terminal. To enable the use of conventional OFDM demodulation in the receiver, we next formulate the convex optimization problem, which jointly performs MU precoding, OFDM modulation, and PAR reduction.

We start by specifying the necessary constraints. In order to remove MUI, the following *precoding constraints* must hold:

$$\mathbf{s}_w = \mathbf{H}_w \mathbf{x}_w, \ w \in \mathcal{T}. \tag{6}$$

To ensure certain desirable spectral properties of the transmitted OFDM signals, the inactive OFDM tones (indexed by $\mathcal{T}^c$) must satisfy the following *shaping constraints*:

$$\mathbf{0}_{N \times 1} = \mathbf{x}_w, \ w \in \mathcal{T}^c. \tag{7}$$

PAR reduction is achieved similarly to (P-INF), with the main difference that we want to minimize the $\ell_{\widetilde{\infty}}$-norm of the *time-domain* samples $\hat{\mathbf{a}}_n$, $\forall n$. In order to simplify notation, we define the (linear) mapping between the time-domain samples $\hat{\mathbf{a}}_n$, $\forall n$, and the wth (frequency-domain) transmit vector $\mathbf{x}_w$ as $\mathbf{x}_w = f_w(\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_N)$, where the linear function $f_w(\cdot)$ applies the DFT according to $\mathbf{a}_n = \mathbf{F}_W \hat{\mathbf{a}}_n$, $\forall n$ and performs the re-ordering defined in (2).

With (6) and (7), we are able to formulate the downlink transmission scheme as a convex optimization problem:

$$(\text{PMP}) \quad \begin{cases} \underset{\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_N}{\text{minimize}} & \max\{\|\tilde{\mathbf{a}}_1\|_{\widetilde{\infty}}, \dots, \|\tilde{\mathbf{a}}_N\|_{\widetilde{\infty}}\} \\ \text{subject to} & \mathbf{s}_w = \mathbf{H}_w f_w(\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_N), \ w \in \mathcal{T} \\ & \mathbf{0}_{N \times 1} = f_w(\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_N), \ w \in \mathcal{T}^c. \end{cases}$$

The vectors $\hat{\mathbf{a}}_n$, $\forall n$ which minimize (PMP) correspond to the time-domain OFDM samples to be transmitted from each antenna. Following the reasoning of Section III-A, we expect these vectors to have low PAR (see Section V for corresponding simulation results). In what follows, PMP refers to the general method of jointly performing precoding, modulation, and PAR reduction, whereas (PMP) refers to the actual optimization problem stated above.

C. Relaxation of (PMP)

The high dimensionality of (PMP) for large-scale MIMO systems necessitates corresponding efficient optimization algorithms. To this end, we relax the constraints of (PMP) to arrive at an optimization problem that can be solved efficiently using the algorithm developed in Section IV.

To simplify the notation, we aggregate all time-domain vectors in $\bar{\mathbf{a}} = [\, \hat{\mathbf{a}}_1^T \ \cdots \ \hat{\mathbf{a}}_N^T \,]^T$ and rewrite the constraints of (PMP) as a *single* linear system of equations. Specifically, both constraints in (PMP) can be rewritten as $\bar{\mathbf{b}} = \overline{\mathbf{C}}\bar{\mathbf{a}}$, where the vector $\bar{\mathbf{b}}$ is a concatenation of $\mathbf{s}_w$, $w \in \mathcal{T}$ and $|\mathcal{T}^c|$ all-zeros vectors of dimension N; the matrix $\overline{\mathbf{C}}$ implements the right-hand-side of the constraints (6) and (7), i.e., also includes the inverse Fourier transforms.[5] We can now re-state (PMP) in more compact form as

$$(\text{PMP}) \quad \underset{\bar{\mathbf{a}}}{\text{minimize}} \ \|\bar{\mathbf{a}}\|_{\widetilde{\infty}} \quad \text{subject to} \ \bar{\mathbf{b}} = \overline{\mathbf{C}}\bar{\mathbf{a}}.$$

In practice, it is desirable to relax the constraint $\bar{\mathbf{b}} = \overline{\mathbf{C}}\bar{\mathbf{a}}$. Firstly, from an implementation point-of-view, relaxing the constraints in (PMP) enables us to develop an efficient algorithm (see Section IV). Secondly, in the medium-to-low SNR regime, the effect of thermal noise at the receiver is comparable to that of MUI and out-of-band interference. Hence, relaxing the equation $\bar{\mathbf{b}} = \overline{\mathbf{C}}\bar{\mathbf{a}}$ to $\|\bar{\mathbf{b}} - \overline{\mathbf{C}}\bar{\mathbf{a}}\|_2 \leq \eta$ does not significantly degrade the performance for small values of η. To

---

[5]For the sake of simplicity of exposition, the actual structural details of the matrix $\overline{\mathbf{C}}$ are omitted.

develop an efficient algorithm for the large dimensions faced in large-scale MU-MIMO-OFDM systems (see Section IV), we state a relaxed version of (PMP) in Lagrangian form as

$$(\text{PMP-L}) \quad \underset{\bar{\mathbf{a}}}{\text{minimize}} \ \lambda\|\bar{\mathbf{a}}\|_{\widetilde{\infty}} + \|\bar{\mathbf{b}} - \overline{\mathbf{C}}\bar{\mathbf{a}}\|_2^2,$$

where $\lambda > 0$ is a regularization parameter. Note that (PMP-L) is an $\ell_{\widetilde{\infty}}$-norm regularized LS problem and λ allows one to trade fidelity to the constraints with the amount of PAR reduction (similarly to the parameter η); the associated trade-offs are investigated in Section V-D. Note that the algorithm developed in Section IV operates on real-valued variables. To this end, (PMP) and (PMP-L) must be transformed into *equivalent* real-valued problems. This transformation, however, is straightforward and we omit the details due to space limitations.

D. Extensions of PMP

The basic ideas behind PMP can be extended to several other scenarios. Corresponding examples are outlined in the next paragraphs.

1) Emulating Other Linear Precoders: By replacing the precoding constraints in (6) by

$$\mathbf{H}_w \mathbf{P}_w \mathbf{s}_w = \mathbf{H}_w \mathbf{x}_w, \ w \in \mathcal{T}, \tag{8}$$

where $\mathbf{P}_w$ is an $N \times M$ precoding matrix of choice, one can generalize PMP to a variety of linear precoders. We emphasize that this generalization allows one to trade MUI removal with noise enhancement and could be used to take into account imperfect channel-state information at the transmitter, e.g., by using a minimum mean-square error precoder (see, e.g., [17]).

2) Peak-Power Constrained Optimization: Instead of normalizing the power of the transmitted vectors as in (1), one may want to impose a predefined upper bound $P_{\max}$ on the transmit power already in the optimization problem. To this end, an additional constraint of the form $\|\bar{\mathbf{a}}\|_2^2 \leq P_{\max}$ could be added to (PMP), which ensures that—if a feasible solution exists—the transmit power does not exceed $P_{\max}$. This constraint maintains the convexity of (PMP) but requires the development of a novel algorithm, as the algorithm proposed in Section IV is unable to consider such peak-power constraints.

3) Combining PMP with Tone-Reservation (TR): In [15], the authors proposed to combine Kashin representations [23], [24] with TR to reduce the PAR in OFDM-based communication systems. The underlying idea is to obtain a time-domain signal that exhibits low PAR by exploiting the DoF offered by TR. We emphasize that PMP can easily be combined with TR, by removing certain precoding constraints (6). Specifically, only a subset $\mathcal{T}_d \subset \mathcal{T}$ is used for data transmission; the remaining tones $\mathcal{T}_d^c$ are reserved for PAR reduction. This approach offers additional DoF and is, therefore, expected to further improve the PAR-reduction capabilities of PMP.

4) Application to Point-to-Point MIMO Systems: The proposed transmission scheme can be used for point-to-point MIMO systems for which channel-state information is available at the transmitter, e.g., IEEE 802.11n [20]. In such systems, MUI does not need to be removed as the MIMO detector is able to separate the transmitted data streams;

hence, there is potentially more flexibility in the choice of the precoding matrices $\mathbf{P}_w$, $\forall w$, as opposed to in a MU-MIMO scenario, which requires the removal of MUI.

5) Application to Single-Carrier Systems: The idea of PMP, i.e., to simultaneously perform precoding, modulation, and PAR reduction, can also be adapted for single-carrier large-scale MIMO systems exhibiting ISI. To this end, one might want to replace the constraints in (P-INF) by[6]

$$
\begin{bmatrix}
\hat{\mathbf{s}}_1 \\
\hat{\mathbf{s}}_2 \\
\vdots \\
\hat{\mathbf{s}}_D \\
\hat{\mathbf{s}}_{D+1} \\
\vdots \\
\hat{\mathbf{s}}_Q
\end{bmatrix}
=
\begin{bmatrix}
\widehat{\mathbf{H}}_1 & \mathbf{0}_{M\times N} & \cdots & \mathbf{0}_{M\times N} \\
\widehat{\mathbf{H}}_2 & \widehat{\mathbf{H}}_1 & \cdots & \mathbf{0}_{M\times N} \\
\vdots & \vdots & \ddots & \vdots \\
\widehat{\mathbf{H}}_D & \widehat{\mathbf{H}}_{D-1} & \cdots & \widehat{\mathbf{H}}_1 \\
\mathbf{0}_{M\times N} & \widehat{\mathbf{H}}_D & \cdots & \widehat{\mathbf{H}}_2 \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{0}_{M\times N} & \mathbf{0}_{M\times N} & \cdots & \widehat{\mathbf{H}}_D
\end{bmatrix}
\begin{bmatrix}
\hat{\mathbf{x}}_1 \\
\hat{\mathbf{x}}_2 \\
\vdots \\
\hat{\mathbf{x}}_D \\
\hat{\mathbf{x}}_{D+1} \\
\vdots \\
\hat{\mathbf{x}}_Q
\end{bmatrix}
$$

and minimize the $\ell_{\widetilde{\infty}}$-norm of the vector $\overline{\mathbf{x}} = [\,\hat{\mathbf{x}}_1^T \,\cdots\, \hat{\mathbf{x}}_Q^T\,]^T$, which contains the PAR-reduced time-domain samples to be transmitted. The channel matrices $\widehat{\mathbf{H}}_t$ are associated to the delay (or tap) $t = 1,\ldots,D$, the information symbols are denoted by $\hat{\mathbf{s}}_q$, $q = 1,\ldots,Q$, and $Q \geq D$ refers to the number of transmitted information symbols per block. Alternatively to PMP, the CE precoding scheme developed in [7] can also be used with the constraints given above. A detailed investigation of both transmission schemes is, however, left for future work.

IV. FAST ITERATIVE TRUNCATION ALGORITHM

A common approach to solve optimization problems of the form (PMP) and (PMP-L) is to use interior-point methods [25]. Such methods, however, often result in prohibitively high computational complexity for the problem sizes faced in large-scale MIMO systems. Hence, to enable practical implementation, more efficient algorithms are of paramount importance. While a large number of computationally efficient algorithms for the ℓ_1-norm regularized LS problem have been developed in the compressive-sensing and sparse-signal recovery literature, e.g., [26], efficient solvers for the ℓ_∞-norm regularized LS problem (PMP-L), however, seem to be missing.

A. Summary of ISTA/FISTA

In this section, we summarize the framework developed in [27] for ℓ_1-norm-based LS, which builds the basis of the algorithm derived in Section IV-B for solving (PMP-L).

1) ISTA: The goal of the iterative soft-thresholding algorithm (ISTA) developed in [27] is to compute the solution $\hat{\mathbf{x}}$ to real-valued convex optimization problems of the form

$$(\text{P}) \quad \min_{\mathbf{x}} F(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}),$$

where $g(\mathbf{x})$ is a real-valued continuous convex function that is possibly non-smooth and $h(\mathbf{x})$ is a smooth convex function, which is continuously differentiable with the Lipschitz constant L. The resulting algorithms are initialized by an arbitrary

[6]Note that the exact structure of the Toeplitz matrix depends on the pre- and post-ambles of the used block-transmission scheme.

vector $\mathbf{x}_0$. The main ingredient of ISTA is the proximal map defined as [27]

$$\mathbf{p}_L(\mathbf{y}) = \arg\min_{\mathbf{x}}\left\{ g(\mathbf{x}) + \frac{L}{2}\left\| \mathbf{x} - \left(\mathbf{y} - \frac{1}{L}\nabla h(\mathbf{y}) \right) \right\|_2^2 \right\}, \quad (9)$$

which constitutes the main iteration step defined as:

$$\mathbf{x}_k = \mathbf{p}_L(\mathbf{x}_{k-1}), \quad k = 1,\ldots,K.$$

Here, K denotes the maximum number of iterations. We emphasize that (9) has a simple closed-form solution for ℓ_1-norm regularized LS, leading to a low-complexity first-order algorithm, i.e., an algorithm requiring matrix-vector multiplications and simple shrinkage operations only. This property renders ISTA an attractive solution for PMP, as the involved matrices $\overline{\mathbf{C}}$ and its adjoint $\overline{\mathbf{C}}^H$ exhibit a structure that enables fast matrix-vector multiplication (see Section III-C).

2) Fast Version of ISTA: As detailed in [27], ISTA exhibits sub-linear convergence, i.e., $F(\mathbf{x}_k) - F(\mathbf{x}^*) \simeq O(1/k)$, where $\mathbf{x}^*$ designates the optimal solution to (P). In order to improve the convergence rate, a fast version of ISTA, referred to as FISTA, was developed in [27]. The main idea of FISTA is to evaluate the proximal map (9) with a (linear) combination of the previous two points $(\mathbf{x}_{k-1}, \mathbf{x}_{k-2})$ instead of $\mathbf{x}_{k-1}$ only (see [27] for the details), which improves the convergence rate to $F(\mathbf{x}_k) - F(\mathbf{x}^*) \simeq O(1/k^2)$ and builds the foundation of the algorithm for solving (PMP-L) described next.

B. Fast Iterative Truncation Algorithm (FITRA)

To simplify the derivation of the first-order algorithm for solving (PMP-L), we describe the algorithm for solving the Lagrangian variant of (P-INF) defined as follows:

$$(\text{P-INF-L}) \quad \min_{\tilde{\mathbf{x}}} \lambda\|\tilde{\mathbf{x}}\|_\infty + \|\mathbf{s} - \mathbf{H}\tilde{\mathbf{x}}\|_2^2.$$

First, we must compute the (smallest) Lipschitz constant L for the function $h(\mathbf{x}) = \|\mathbf{s} - \mathbf{H}\mathbf{x}\|_2^2$ and then, evaluate the proximal map (9) for the functions $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_\infty$ and $h(\mathbf{x})$.

1) FITRA: The (smallest) Lipschitz constant of the gradient $\nabla h(\mathbf{x})$ corresponds to $L = 2\sigma_{\max}^2(\mathbf{H})$, which can, for example, be calculated efficiently using the power method [28]. To compute the proximal map (9) for (P-INF-L), we define the auxiliary vector

$$\mathbf{w} = \mathbf{y} - \frac{1}{L}\nabla h(\mathbf{y}) = \mathbf{y} - \frac{2}{L}\mathbf{H}^T(\mathbf{H}\mathbf{y} - \mathbf{x})$$

which enables us to re-write the proximal map in more compact form as

$$\mathbf{p}_L(\mathbf{y}) = \arg\min_{\tilde{\mathbf{x}}}\left\{ \lambda\|\tilde{\mathbf{x}}\|_\infty + \frac{L}{2}\|\tilde{\mathbf{x}} - \mathbf{w}\|_2^2 \right\}. \quad (10)$$

Unfortunately, (10) does—in contrast to ℓ_1-norm regularized LS—not have a simple closed-form solution for (P-INF-L). Nevertheless, standard algebraic manipulations enable us to evaluate the proximal map efficiently using the following two-step approach: First, we compute

$$\alpha = \arg\min_{\tilde{\alpha}}\left\{ \lambda\tilde{\alpha} + \frac{L}{2}\sum_{i=1}^{N}\left([|[\mathbf{w}]_i| - \tilde{\alpha}]^+ \right)^2 \right\}, \quad (11)$$

Algorithm 1 Fast Iterative Truncation Algorithm (FITRA)

1: **initialize** $\mathbf{x}_0 \leftarrow \mathbf{0}_{N \times 1}$, $\mathbf{y}_1 \leftarrow \mathbf{x}_0$, $t_1 \leftarrow 1$, $L \leftarrow 2\sigma_{\max}^2(\mathbf{H})$
2: **for** $k = 1, \ldots, K$ **do**
3: $\quad \mathbf{w} \leftarrow \mathbf{y}_k - \frac{2}{L}\mathbf{H}^T(\mathbf{H}\mathbf{y}_k - \mathbf{s})$
4: $\quad \alpha \leftarrow \arg\min\limits_{\tilde{\alpha}} \left\{ \lambda\tilde{\alpha} + \frac{L}{2}\sum_{i=1}^{N}\left([|[\mathbf{w}]_i| - \tilde{\alpha}]^{+}\right)^2 \right\}$
5: $\quad \mathbf{x}_k \leftarrow \mathrm{trunc}_\alpha(\mathbf{w})$
6: $\quad t_{k+1} \leftarrow \frac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right)$
7: $\quad \mathbf{y}_{k+1} \leftarrow \mathbf{x}_k + \frac{t_k - 1}{t_{k+1}}(\mathbf{x}_k - \mathbf{x}_{k-1})$
8: **end for**
9: **return** $\mathbf{x}_K$

for which general-purpose scalar optimization algorithms, such as the bisection method [29], can be used. Then, we apply element-wise truncation (clipping) of $\mathbf{w}$ to the interval $[-\alpha, \alpha]$ according to $\mathbf{p}_L(\mathbf{x}) = \mathrm{trunc}_\alpha(\mathbf{w})$. The truncation operator applied to the scalar $x \in \mathbb{R}$ is defined as

$$\mathrm{trunc}_\alpha(x) = \min\{\max\{x, -\alpha\}, +\alpha\}.$$

The resulting first-order algorithm, including the methods proposed in [27] to improve the convergence rate (compared to ISTA), is detailed in Algorithm 1 and referred to as the fast iterative truncation algorithm (FITRA).

2) Convergence Rate: The following proposition is an immediate consequence of the convergence results for ISTA/FISTA in [27, Thm. 4.4] and characterizes the convergence rate of FITRA analytically.

Proposition 2: The convergence rate of FITRA (as detailed in Algorithm 1) satisfies

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{(k+1)^2},$$

where $\mathbf{x}^*$ denotes the solution to (P-INF-L), $\mathbf{x}_k$ is the FITRA estimate at iteration k, $\mathbf{x}_0$ the initial value at iteration $k = 0$, and $F(\mathbf{x}) = \lambda\|\mathbf{x}\|_\infty + \|\mathbf{s} - \mathbf{H}\mathbf{x}\|_2^2$.

We emphasize that continuation strategies, e.g., [30], potentially reduce the computational complexity of FITRA; the investigation of such methods is left for future work.

C. Related Work

An algorithm to compute an approximation to (P-INF) relying an iterative truncation procedure similar to FITRA was proposed in [24]. The main differences between these algorithms are as follows: The algorithm in [24] requires the matrix $\mathbf{H}$ to be a tight frame and relies on a constant (and pre-defined) truncation parameter, which depends on $\mathbf{H}$ and cannot be computed efficiently. In the present application, however, the matrix $\mathbf{H}$ is, in general, not a tight frame and depends on the channel realization; this requires to chose the truncation parameter in [24] heuristically and hence, convergence of this method is no longer guaranteed. FITRA, in contrast, does not require the matrix $\mathbf{H}$ to be a tight frame, avoids manual tuning of the truncation parameter, and is guaranteed to converge to the solution of (P-INF-L).

V. SIMULATION RESULTS

In this section, we demonstrate the efficacy of the proposed joint precoding, modulation, and PAR reduction approach, and provide a comparison to conventional MU precoding schemes.

A. Simulation Parameters

Unless explicitly stated otherwise, all simulation results are for a MU-MIMO-OFDM system having $N = 100$ antennas at the BS and serving $M = 10$ single-antenna terminals. We employ OFDM with $W = 128$ tones and use a spectral map $\mathcal{T}$ as specified in the 40 MHz-mode of IEEE 802.11n [20].[7] We consider coded transmission, i.e., for each user, we independently encode 216 information bits using a convolutional code (rate-$1/2$, generator polynomials [133_o 171_o], and constraint length 7), apply random interleaving (across OFDM tones), and map the coded bits to a 16-QAM constellation (using Gray labeling).

To implement (PMP-L), we use FITRA as detailed in Algorithm 1 with a maximum number of $K = 2000$ iterations and a regularization parameter of $\lambda = 0.25$. In addition to LS and MF precoding, we also consider the performance of a baseline precoding and PAR-reduction method. To this end, we employ LS precoding followed by truncation (clipping) of the entries of the time-domain samples $\hat{\mathbf{a}}_n$, $\forall n$. We use a clipping strategy where one can specify a target PAR, which is then used to compute a clipping level for which the PAR in (4) of the resulting time-domain samples is no more than the chosen target PAR.

The precoded and normalized vectors are then transmitted over a frequency-selective channel modeled as a tap-delay line with $T = 4$ taps. The time-domain channel matrices $\widehat{\mathbf{H}}_t$, $t = 1, \ldots, T$, that constitute the impulse response of the channel, have i.i.d. circularly symmetric Gaussian distributed entries with zero mean and unit variance. To detect the transmitted information bits, each user m performs soft-output demodulation of the received symbols $[\mathbf{y}_w]_m$, $w = 1, \ldots, W$ and applies a soft-input Viterbi decoder.

B. Performance Measures

To compare the PAR characteristics of different precoding schemes, we use the complementary cumulative distribution function (CCDF) defined as

$$\mathrm{CCDF}(\mathsf{PAR}) = \mathbb{P}\{\mathsf{PAR}_n > \mathsf{PAR}\}.$$

We furthermore define the "PAR performance" as the maximum PAR level PAR^* that is met for 99% of all transmitted OFDM symbols, i.e., given by $\mathrm{CCDF}(\mathsf{PAR}^*) = 1\%$. The error-rate performance is measured by the average (across users) symbol-error rate (SER); a symbol is said to be in error if at least one of the information bits per received OFDM symbol is decoded in error. The "SNR operating point" corresponds to the minimum SNR required to achieve 1% SER. In order to characterize the amount of signal power that is

---

[7] We solely consider $|\mathcal{T}| = 108$ data-carrying tones; the tones reserved for pilot symbols in IEEE 802.11n [20] are ignored in all simulations.
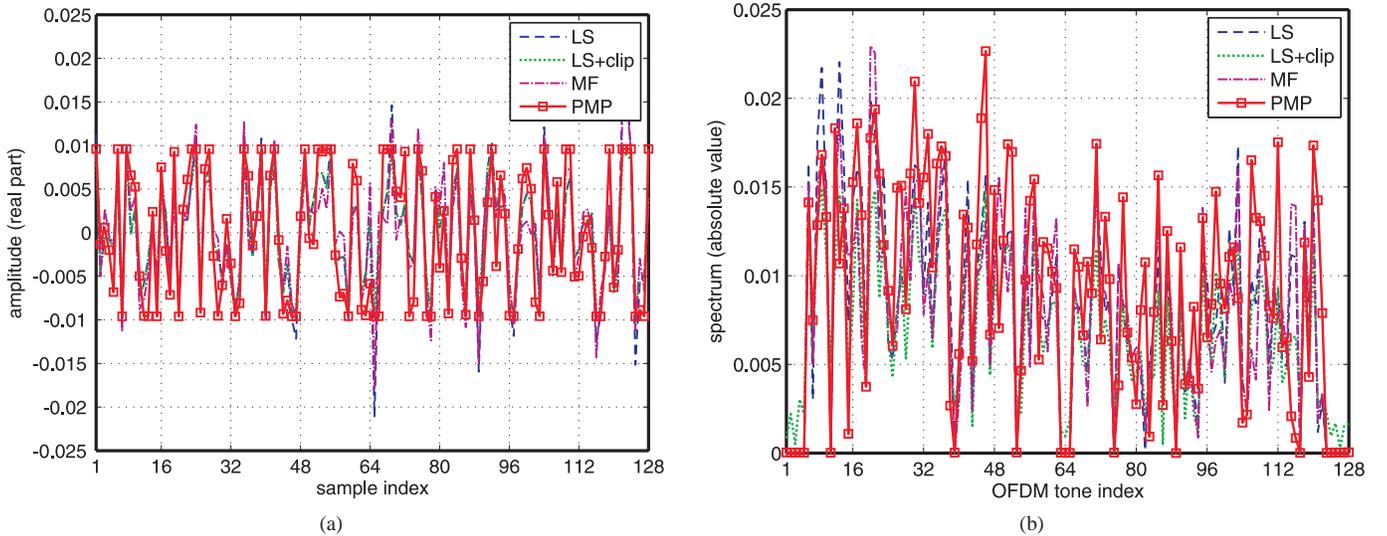
(a)

(b)

Fig. 2. Time/frequency representation for different precoding schemes. The target PAR for LS+clip is $4\,\text{dB}$ and $\lambda = 0.25$ for PMP relying on FITRA. (a) Time-domain signals (PAR: LS $= 10.4\,\text{dB}$, LS+clip $= 4.0\,\text{dB}$, MF $= 10.1\,\text{dB}$, and PMP $= 1.9\,\text{dB}$). Note that PMP generates a time-domain signal of substantially smaller PAR than LS and MF. (b) Frequency-domain signals (OBR: LS $= -\infty\,\text{dB}$, LS+clip $= -11.9\,\text{dB}$, MF $= -\infty\,\text{dB}$, and PMP $= -52.9\,\text{dB}$). Note that LS, MF, and PMP preserve the spectral properties. LS+clip suffers from substantial OBR (visible at both ends of the spectrum).
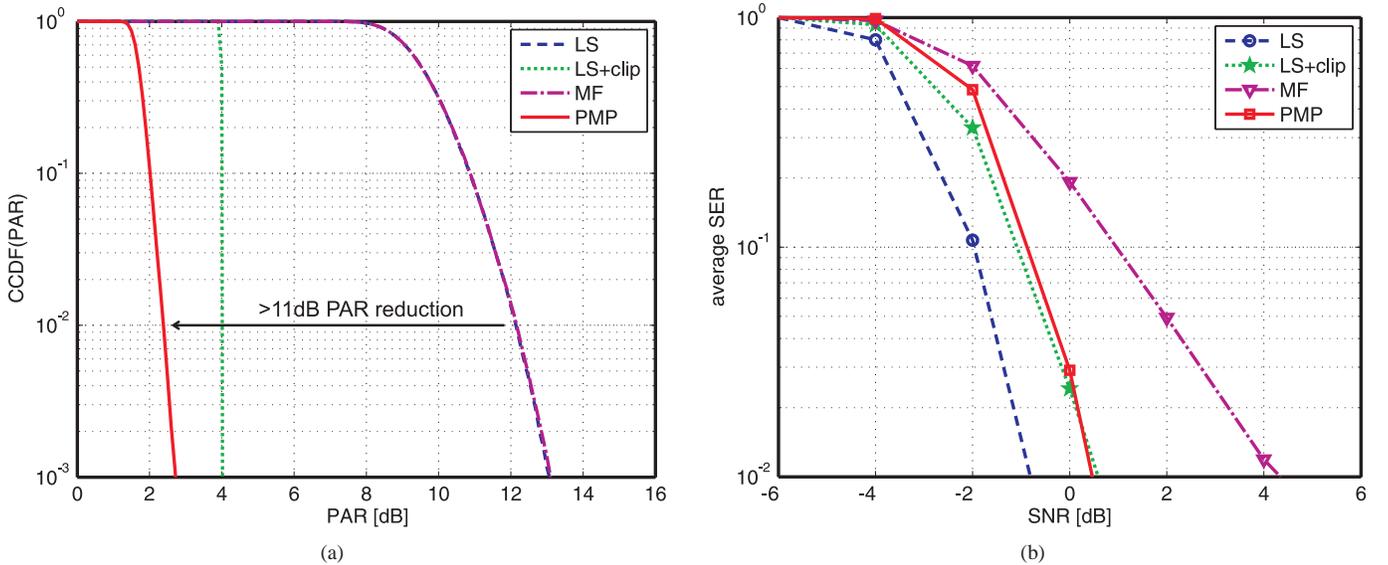

(a)

(b)

Fig. 3. PAR and SER performance for various precoding schemes. The target PAR for LS+clip is $4\,\text{dB}$ and $\lambda = 0.25$ for PMP relying on FITRA. (a) PAR performance (the curves of LS and MF overlap). Note that PMP effectively reduces the PAR compared to LS and MF precoding. (b) Symbol error-rate (SER) performance. Note that the signal normalization causes $1\,\text{dB}$ SNR-performance loss for PMP compared to LS precoding. The loss of MF is caused by residual MUI; the loss of LS+clip is caused by normalization and residual MUI.

transmitted outside the active tones $\mathcal{T}$, we define the out-of-band (power) ratio (OBR) as follows:

$$\mathsf{OBR} = \frac{|\mathcal{T}| \sum_{w \in \mathcal{T}^c} \|\mathbf{x}_w\|_2^2}{|\mathcal{T}^c| \sum_{w \in \mathcal{T}} \|\mathbf{x}_w\|_2^2}.$$

Note that for LS and MF precoding, we have $\mathsf{OBR} = 0$, as they operate independently on each of the W tones; for PMP or LS followed by clipping, we have $\mathsf{OBR} > 0$ in general.

C. Summary of PMP Properties

Figures 2 and 3 summarize the key characteristics of PMP and compare its PAR-reduction capabilities and error-rate performance to those of LS and MF precoding, as well as to

LS precoding followed by clipping (denoted by "LS+clip" in the following). Fig. 2(a) shows the real part of a time-domain signal $\hat{\mathbf{a}}_1$ for all precoding schemes (the imaginary part behaves similarly). Clearly, PMP results in time-domain signals having a significantly smaller PAR than that of LS and MF; for LS+clip the target PAR corresponds to $4\,\text{dB}$. The frequency-domain results shown in Fig. 2(b) confirm that LS, MF, and PMP maintain the spectral constraints. For LS+clip, however, the OBR is $-11.9\,\text{dB}$, which is a result of ignoring the spectral constraints (see the non-zero OFDM tones at both ends of the spectrum in Fig. 2(b)). Fig. 3(a) shows the PAR-performance characteristics for all considered precoding schemes. One can immediately see that PMP reduces the
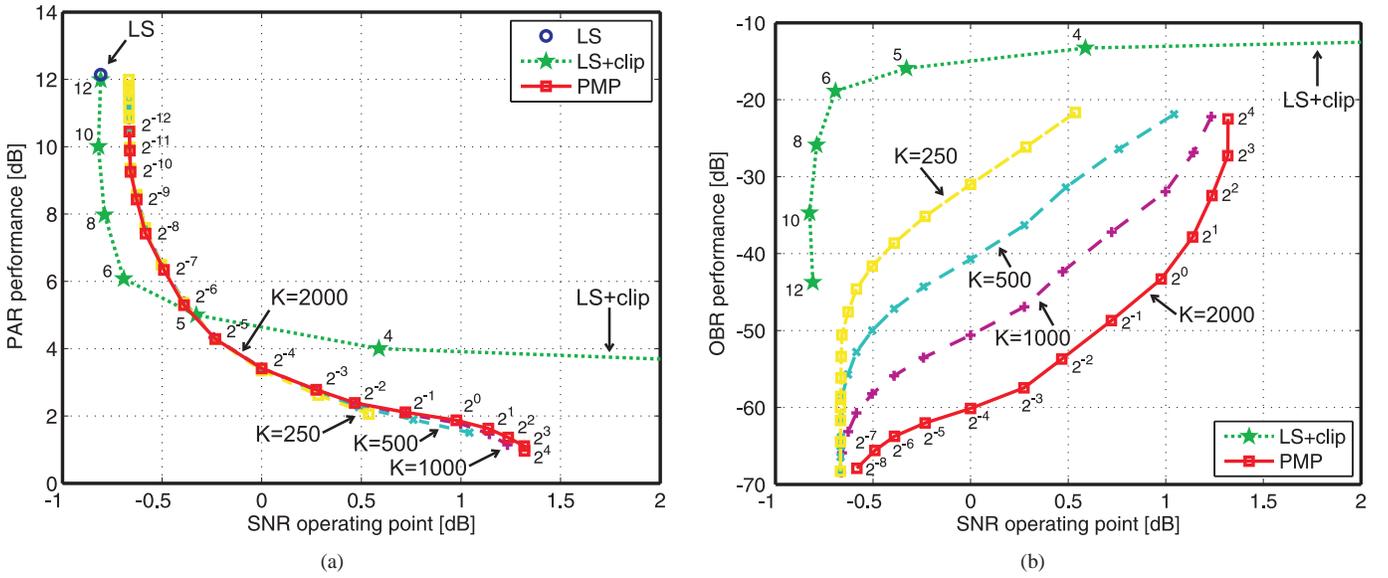
Fig. 4. SNR, PAR, and OBR performance trade-offs of PMP. The numbers next to the trade-off curve for FITRA correspond to the regularization parameter λ used in (PMP-L). The LS+clip curves are parametrized by the target PAR in dB. (a) PAR/SNR trade-off (parts of the FITRA curves overlap). (b) OBR/SNR trade-off (all curves labeled with K correspond to FITRA).

PAR by more than $11\,\mathrm{dB}$ compared to LS and MF precoding (at $\mathsf{CCDF(PAR)} = 1\%$); as expected, LS+clip achieves $4\,\mathrm{dB}$ PAR deterministically. In order to maintain a constant transmit power, the signals resulting from PMP require a stronger normalization (roughly $1\,\mathrm{dB}$) than the signals from LS precoding; this behavior causes the SNR-performance loss compared to LS (see Fig. 3(b)). The performance loss of MF and LS+clip is mainly caused by residual MUI.

D. SNR, PAR, and OBR Trade-Offs

As observed in Fig. 3, PMP is able to significantly reduce the PAR but results in an SNR-performance loss compared to LS precoding. Hence, there exists a trade-off between PAR and SER, which can be controlled by the regularization parameter λ of (PMP-L). Fig. 4(a) characterizes this trade-off for $\lambda = 2^v$ with $v \in \{-12, \ldots, 4\}$. In addition to the performance of LS and MF precoding, we show the behavior of LS+clip for various target-PAR values.

Fig. 4(a) shows that PMP is able to cover a large trade-off region that can be tuned by the regularization parameter λ of (PMP-L). In particular, for a given number of FITRA iterations $K = 2000$, decreasing λ approaches the performance of LS precoding—increasing λ reduces the PAR but results in a graceful degradation of the SNR operating point.[8] Hence, (PMP-L) allows one to adjust the PAR to the linearity properties of the RF components, while keeping the resulting SNR-performance loss at a minimum. As shown in Fig. 4(a), LS+clip achieves a similar trade-off characteristic as PMP; for less aggressive values of the target PAR, LS+clip even seems to outperform PMP.

It is important to realize that even if LS+clip outperforms PMP in terms of the PAR/SNR trade-off in the high-PAR regime, LS+clip results in substantial out-of-band interference; this important drawback is a result of ignoring the shaping constraints (7). In particular, we can observe from Fig. 4(b) that reducing the PAR for LS+clip quickly results in significant OBR, which renders this scheme useless in practice. By way of contrast, the OBR of PMP is significantly lower and degrades gracefully when lowering the PAR. Furthermore, we see that reducing the maximum number of FITRA iterations K increases the OBR. Hence, the regularization parameter λ together with the maximum number of FITRA iterations K determine the PAR, OBR, and SNR performance of PMP. We finally note that for $K = 2000$ the computational complexity of FITRA is one-to-two orders of magnitude larger than that of LS precoding. The underlying reason is the fact that LS precoding solves N *independent* problems, whereas PMP requires the solution to a *joint* optimization problem among all N transmit antennas.

E. Impact of Antenna Configuration and Channel Taps

We finally investigate the impact of the antenna configuration to the PAR performance of PMP and LS precoding. To illustrate the impact of the channel model, we also vary the number of non-zero channel taps $T \in \{2, 4, 8\}$. Fig. 5 shows that increasing the number of transmit antennas yields improved PAR performance for PMP; this behavior was predicted analytically in (5) for the (narrow-band) system considered in Section III-A2. Increasing the number of channel taps T also has a beneficial impact on the PAR if using PMP. An intuitive explanation for this behavior is that having a large number of taps increases the number of DoF, which can then be exploited by PMP to reduce the PAR. For LS precoding, however, the resulting PAR is virtually independent of the number of channel taps.[9] In summary, PMP is suitable for MU-MIMO systems

---

[8]For $\lambda > 0$, a small SNR gap remains; for $\lambda = 0$, however, (PMP-L) corresponds to LS precoding and the gap vanishes.

[9]MF and LS+clip exhibit the same behavior; the corresponding curves are omitted in Fig. 5.
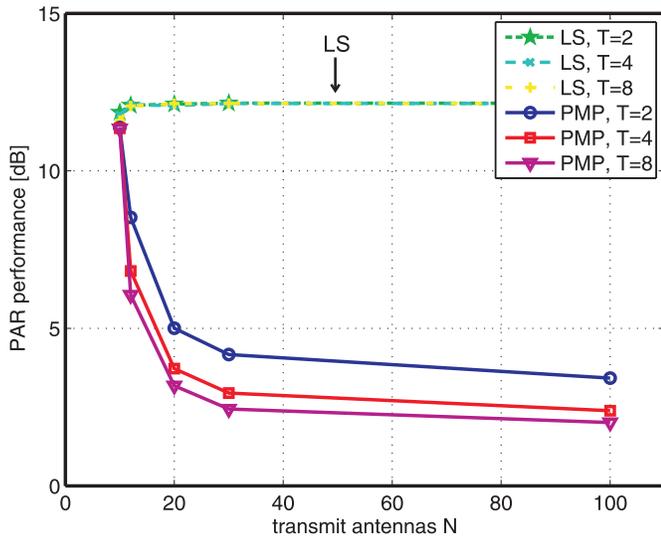
Fig. 5. PAR performance of PMP and LS precoding depending on the number of transmit antennas N and the number of non-zero channel taps T; the number of users $M = 10$ is held constant and $\lambda = 0.25$ for PMP relying on FITRA. (The curves for LS precoding overlap.)

offering a large number of DoF, but also enables substantial PAR reduction for small-scale MIMO systems and channels offering only a small amount of frequency-diversity.

VI. CONCLUSIONS AND OUTLOOK

The proposed joint precoding, modulation, and PAR reduction framework, referred to as PMP, facilitates an explicit trade-off between PAR, SNR performance, and out-of-band interference for the large-scale MU-MIMO-OFDM downlink. As for the constant-envelope precoder in [7], the fundamental motivation of PMP is the large number of DoF offered by systems where the number of BS antennas is much larger than the number of terminals (users). Essentially, the downlink channel matrix has a high-dimensional null-space, which enables us to design transmit signals with "hardware-friendly" properties, such as low PAR. In particular, PMP yields per-antenna constant-envelope OFDM signals in the large-antenna limit, i.e., for $N \to \infty$. PMP is formulated as a convex optimization problem for which a novel efficient numerical technique, called the fast iterative truncation algorithm (FITRA), was devised.

Numerical experiments showed that PMP is able to reduce the PAR by more than 11 dB compared to conventional precoding methods, without creating significant out-of-band interference; this substantially alleviates the linearity requirements of the radio-frequency (RF) components. Furthermore, PMP only affects the signal processing at the BS and can therefore be deployed in existing MIMO-OFDM wireless communication systems, such as IEEE 802.11n [20].

In addition to the extensions outlined in Section III-D, there are many possibilities for future work. Analytical PAR-performance guarantees of PMP are missing; the development of such results is a challenging open research topic. Moreover, a detailed analysis of the impact of imperfect channel state information on the performance of PMP is left for future work. Finally, reducing the computational complexity of FITRA, e.g., using continuation [30], is part of ongoing work.

REFERENCES

[1] C. Studer and E. G. Larsson, "PAR-aware multi-user precoder for the large-scale MIMO-OFDM downlink," in *Proc. of the 9th International Symposium on Wireless Communication Systems (ISWCS)*, Paris, France, August 2012.

[2] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Scaling up MIMO: opportunities and challenges with very large arrays," *arXiv:1201.3210v1*, Jan. 2012.

[3] T. L. Marzetta, "Non-cooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Comm.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[4] ——, "How much training is required for multi-user MIMO?" in *Proc. 40th Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, USA, Oct. 2006, pp. 359–363.

[5] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO: How many antennas do we need?" in *Proc. IEEE 49th Ann. Allerton Conf. on Comm. Control, and Computing (Allerton)*, Monticello, IL, USA, Sept. 2011, pp. 545–550.

[6] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *arXiv:1112.3810v1*, Dec. 2011.

[7] S. K. Mohammed and E. G. Larsson, "Per-antenna constant envelope precoding for large multi-user MIMO systems," *arXiv:1111.3752v1*, Jan. 2012.

[8] R. van Nee and R. Prasad, *OFDM for wireless multimedia communications*. Artech House Publ., 2000.

[9] S. H. Han and J. H. Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," *IEEE Wireless Comm.*, vol. 12, no. 2, pp. 1536–1284, Apr. 2005.

[10] R. W. Bäuml, R. F. H. Fischer, and J. B. Huber, "Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping," *IEE Elec. Letters*, vol. 32, no. 22, pp. 2056–2057, Oct. 1996.

[11] S. H. Müller and J. B. Huber, "OFDM with reduced peak-to-average power ratio by optimum combination of partial transmit sequences," *IEE Elec. Letters*, vol. 33, no. 5, pp. 368–369, Feb. 1997.

[12] B. S. Krongold and D. L. Jones, "PAR reduction in OFDM via active constellation extension," in *IEEE Int. Conf. on Acoustics, Speech, and Sig. Proc. (ICASSP)*, vol. 4, Hong Kong, China, Apr. 2003, pp. 525–528.

[13] ——, "An active-set approach for OFDM PAR reduction via tone reservation," *IEEE Trans. Sig. Proc.*, vol. 52, no. 2, pp. 495–509, Feb. 2004.

[14] R. F. H. Fischer and M. Hoch, "Directed selected mapping for peak-to-average power ratio reduction in MIMO OFDM," *IEE Elec. Letters*, vol. 42, no. 2, pp. 1289–1290, Oct. 2006.

[15] J. Illic and T. Strohmer, "PAPR reduction in OFDM using Kashin's representation," in *Proc. IEEE 10th Workshop on Sig. Proc. Advances in Wireless Comm. (SPAWC)*, Perugia, Italy, June 2009, pp. 444–448.

[16] T. Tsiligkaridis and D. L. Jones, "PAPR reduction performance by active constellation extension for diversity MIMO-OFDM systems," *J. Electrical and Computer Eng.*, no. 930368, 2010.

[17] R. F. H. Fischer, *Precoding and Signal Shaping for Digital Transmission*. Wiley, New York, 2002.

[18] S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A low-complexity precoder for large multiuser MISO systems," in *IEEE Vehicular Tech. Conf (VTC)*, vol. Spring, Marina Bay, Singapore, May 2008, pp. 797–801.

[19] C. Siegl and R. F. H. Fischer, "Selected basis for PAR reduction in multi-user downlink scenarios using lattice-reduction-aided precoding," *EURASIP J. on Advanced Sig. Proc.*, vol. 17, pp. 1–11, July 2011.

[20] *IEEE Draft Standard; Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications; Amendment 4: Enhancements for Higher Throughput*, P802.11n/D3.0, Sep. 2007.

[21] D. Seethaler and H. Bölcskei, "Performance and complexity analysis of infinity-norm sphere-decoding," *IEEE Trans. Inf. Th.*, vol. 56, no. 3, pp. 1085–1105, Mar. 2010.

[22] U. Erez and S. ten Brink, "A close-to-capacity dirty paper coding scheme," *IEEE Trans. Inf. Th.*, vol. 51, no. 10, pp. 3417–3432, Oct. 2005.

[23] J.-J. Fuchs, "Spread representations," in *Proc. 45th Asilomar Conf. on Signals, Systems, and Comput.*, Pacific Grove, CA, USA, 2011.

[24] Y. Lyubarskii and R. Vershynin, "Uncertainty principles and vector quantization," *IEEE Trans. Inf. Th.*, vol. 56, no. 7, pp. 3491–3501, Jul. 2010.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

[26] E. J. Candès and M. B. Wakin, "An introdutction to compressive sampling," *IEEE Sig. Proc. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[27] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[28] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins Univ. Press, 1996.

[29] G. E. Forsythe, M. A. Malcolm, and C. B. Moler, *Computer Methods for Mathematical Computations*. Prentice-Hall, 1977.

[30] T. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for ℓ_1-regularized minimization with applications to compressed sensing," Dept. Computat. Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR07-07, 2007.