

RLS Precoding for Massive MIMO Systems with Nonlinear Front-End

Ali Bereyhi*, Saba Asaad*, Ralf R. Müller*, and Symeon Chatzinotas†

*Institute for Digital Communications, Friedrich-Alexander Universität Erlangen-Nürnberg,

†Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg,

{ali.bereyhi, saba.asaad, ralf.r.mueller}@fau.de, symeon.chatzinotas@uni.lu

Abstract—To keep massive MIMO systems cost-efficient, power amplifiers with rather small output dynamic ranges are employed. They may distort the transmit signal and degrade the performance. This paper proposes a *distortion aware* precoding scheme for realistic scenarios in which RF chains have nonlinear characteristics. The proposed scheme utilizes the method of regularized least-squares (RLS) to jointly compensate the channel impacts and the distortion imposed by the RF chains.

To construct the designed transmit waveform with low computational complexity, an iterative algorithm based on approximate message passing is developed. This algorithm is shown to track the achievable average signal distortion of the proposed scheme tightly, even for practical system dimensions. The results demonstrate considerable enhancement compared to the state of the art.

Index Terms—Precoding, nonlinear power amplifiers, approximate message passing, regularized least-squares, massive MIMO.

I. INTRODUCTION

Theoretical analyses depict that linear multiuser multiple-input multiple-output (MIMO) precoding techniques are efficient in the large-system limit [1]. This result, along with the low complexity of these schemes, has introduced linear precoding as the dominant approach for signal pre-processing in massive MIMO [2]. Investigations in this respect however rely on the characteristics of transceiver components which are often described by simplified models. An exemplary component, which is the focus of this paper, is the *power amplifier (PA)* used in the transmit radio frequency (RF) chains.

For sake of simplicity, PAs are often treated as linear components. The linear model, however, is not valid in general. In fact, such a characterization is a rather good *approximation* when the peak-to-average power ratio (PAPR) of the signal is less than the input *back-off* of the PA. Such a constraint can be easily violated in massive MIMO systems. In fact, the expense of a PA is proportional to its linearity characteristics. To keep massive MIMO settings cost-efficient, RF chains are implemented via PAs with low back-offs. This increases the nonlinear distortion on the RF transmit signal and degrades the performance.

The nonlinear distortion caused by PAs can be effectively resolved via signal pre-processing. In this respect, one can pre-distort the transmit waveform, either on the *sample* or *symbol*

basis, such that the receive signal will be of the desired shape. For given characteristics of PAs, this approach is analytically tractable with standard signal processing techniques, e.g. [3]. In multiuser MIMO settings, the transmit signal can be directly pre-distorted at the precoding stage. To this end, one can include the RF chains in the channel model, and design the precoder for the end-to-end channel. An example of this approach was studied in [4] where the authors developed a *symbol-level* precoding scheme [5] for MIMO settings with nonlinear PAs. Another example is [6] in which the end-to-end precoding was studied in sample domain considering waveform optimization and pulse shaping. Despite the promising performance, this approach often results in high computational complexity, due to the *nonlinearity* of the end-to-end channel.

Contributions

Regardless of the channel model, precoding is effectively addressed via the method of regularized least-squares (RLS). For the classic linear model, RLS leads to generalized least square error (GLSE) scheme introduced and analyzed in [7]–[9]. This paper extends this RLS-based methodology to MIMO settings with nonlinear RF front-ends. To this end, a multiuser MIMO setting is considered in which the transmit RF chains have a *generic* input-output characteristic. For this setting, end-to-end precoding is addressed by the RLS method considering a general set of constraints on the transmit signal.

The computational complexity of such an approach may be rather high for a large MIMO system, if it is implemented in a straightforward manner. To address this issue, we develop an algorithm based on approximate message passing (AMP) [10]. The complexity of this algorithm scales linearly with the number of transmit antennas, and it tightly tracks the large-system performance of the proposed scheme for practical dimensions.

Notations

Throughout the paper, scalars, vectors and matrices are represented with non-bold, bold lower case and bold upper case letters, respectively. \mathbf{I}_K is a $K \times K$ identity matrix, and \mathbf{H}^\top is the transpose of \mathbf{H} . The real axis and the complex plane are denoted by \mathbb{R} and \mathbb{C} , respectively. For $s \in \mathbb{C}$, $\Re\{s\}$, $\Im\{s\}$ and $\mathbf{s} := [\Re\{s\}, \Im\{s\}]^\top$ are the real part, imaginary part and augmented vector, respectively. $\mathbb{E}\{\cdot\}$ is the statistical expectation. For simplicity, $\{1, \dots, N\}$ is abbreviated by $[N]$. For any

This work has been accepted for presentation in the 20th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) 2019 in Cannes, France. The link to the final version in the Proceedings of SPAWC will be available later.

differentiable function $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \dots, f_n(\mathbf{x})]^\top$, the gradient is defined as $\nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x}) := [\nabla_{\mathbf{x}} f_1(\mathbf{x}), \dots, \nabla_{\mathbf{x}} f_n(\mathbf{x})]^\top$.

II. PROBLEM FORMULATION

We consider downlink transmission in a Gaussian broadcast MIMO channel with a single base station (BS) and K single-antenna users. The BS is equipped with an antenna array of size M , $L \leq M$ nonlinear RF chains, and a switching network which connects each subset of L antennas to the RF chains.

In the n -th transmission interval, the BS intends to transmit data symbols $s_1[n], \dots, s_k[n]$ to the user terminals (UTs). To this end, it constructs the transmit signal $\mathbf{x}[n] \in \mathbb{X}^M$ with L non-zero entries via a precoding scheme. Here, $\mathbb{X} \subseteq \mathbb{C}$ is the *precoding support* and contains all possible points which can be selected as constellation points by the transceiver. For instance, in the case of per-antenna constant envelope precoding, the precoding support is $\mathbb{X} = \{x \in \mathbb{C} : |x|^2 = P\}$.

The m -th entry of the transmit signal represents the symbol which is intended to be sent over antenna element m . Hence, the indices of non-zero entries in $\mathbf{x}[n]$ correspond to those antennas which are set active in transmission time interval n . Let $\mathbb{L}[n] \subseteq [M]$ denote the index set of non-zero entries in the transmit signal $\mathbf{x}[n]$. The switching network connects the RF chains to the antennas indexed by $\mathbb{L}[n]$. The precoded signal is then transmitted via the RF chains.

The system operates in the time division duplexing (TDD) mode which means that the uplink and downlink channels are reciprocal. It is assumed that the channel state information (CSI) is estimated at both the transmit and receive sides at the beginning of each coherence interval within an estimation loop whose duration is much shorter than the coherence interval. Hence, the BS knows the CSI prior to transmission.

A. Nonlinear RF Chains

The main component of an RF chain is the PA which has nonlinear input-output characteristics, in general. Several examples of nonlinear PA models can be followed in the literature; e.g. [11], [12]. For sake of generality, we consider a generic input-output characteristic for the RF chains: Let x be the symbol which is fed to an RF chain. The output of the RF chain is given by $w = f_{\text{RF}}(x)$ with $f_{\text{RF}}(\cdot) : \mathbb{X} \mapsto \mathbb{W}$ for some $\mathbb{W} \subseteq \mathbb{C}$. \mathbb{W} describes the set of all possible constellation points after being distorted by the RF chain. We refer to $f_{\text{RF}}(\cdot)$ as the *RF conversion function*. This function is considered to be of a general form describing various nonlinear PA models, e.g. the well-known amplitude-to-amplitude and amplitude-to-phase distortion model. Noting that the output of an RF chain, which is not fed by any signal, is zero, we have $f_{\text{RF}}(0) = 0$.

Considering the RF conversion model, the signal entry that is observed on an active antenna reads $w_m = f_{\text{RF}}(x_m)$, where $m \in \mathbb{L}[n]$. For passive antennas, we further have

$$w_m = 0 = f_{\text{RF}}(0) = f_{\text{RF}}(x_m), \quad (1)$$

where $m \in [M] \setminus \mathbb{L}[n]$. As a result, we can compactly represent the signal on the transmit antennas as $\mathbf{w}[n] = f_{\text{RF}}(\mathbf{x}[n])$. To

distinguish between $\mathbf{x}[n]$ and $\mathbf{w}[n]$, we refer to $\mathbf{w}[n]$ as the *RF transmit signal* in the transmission interval n .

Remark 1: The RF conversion function is often derived via interpolating methods. Hence, actual outputs slightly deviate from $f_{\text{RF}}(x_m[n])$. As a result, one can write

$$|w_m[n] - f_{\text{RF}}(x_m[n])|^2 \leq \epsilon$$

where $\epsilon \downarrow 0$ in the ideal case.

B. Channel Model

The RF transmit signal is sent to the UTs over a Gaussian broadcast MIMO channel which experiences quasi-static fading. The receive signal at user k for n -th interval is hence

$$y_k[n] = \sqrt{\beta_k} \mathbf{g}_k^\top \mathbf{w}[n] + z_k[n] \quad (2)$$

for $k \in [K]$. Here, $\mathbf{g}_k \in \mathbb{C}^M$ contains the fading coefficients of the uplink channel between user k and the BS. Moreover, β_k describes the path-loss and shadowing in the channel which is the same for all antenna elements at the BS. The random variable $z_k[n]$ denotes additive white Gaussian noise (AWGN) and is assumed to be zero-mean with variance σ^2 .

Considering the channel model, the vector of receive signals at the UTs in transmission interval n , i.e. $\mathbf{y}[n] = [y_1[n], \dots, y_K[n]]^\top$, is compactly represented as

$$\mathbf{y}[n] = \mathbf{H}^\top \mathbf{w}[n] + \mathbf{z}[n] \quad (3)$$

where $\mathbf{z}[n] = [z_1[n], \dots, z_K[n]]^\top$ is the noise vector and

$$\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_K] \quad (4)$$

with $\mathbf{h}_k = \sqrt{\beta_k} \mathbf{g}_k$ represents the uplink channel vector.

III. RLS-BASED PRECODING SCHEME

The ultimate aim of precoding is to pre-process signals, such that data can be recovered from the receive signal at UTs with minimal post-processing. This means that the vector of noise-free receive signals, i.e. $\mathbf{H}^\top \mathbf{w}[n]$, is desired to be close to the data vector, i.e. $\mathbf{s}[n] = [s_1[n], \dots, s_K[n]]^\top$. To formulate this interpretation of precoding, consider the following definition:

Definition 1 (RSS): Let \mathbf{s} be a data vector and \mathbf{v} represent its corresponding RF transmit signal. For a given scaling factor ρ , the residual sum of squares (RSS) at the UTs is defined as

$$\text{RSS}(\mathbf{v}|\rho, \mathbf{s}, \mathbf{H}) = \|\mathbf{H}^\top \mathbf{v} - \sqrt{\rho} \mathbf{s}\|^2. \quad (5)$$

The RSS determines the squared of the Euclidean distance between the noise-free receive signals and data symbols. Using this definition, the precoding problem is interpreted as

$$\mathbf{x}[n] = \underset{\mathbf{u} \in \mathbb{X}^M}{\text{argmin}} \text{RSS}(f_{\text{RF}}(\mathbf{u})|\rho, \mathbf{s}[n], \mathbf{H}) \text{ s.t. } \mathcal{C}(\mathbf{u}) \quad (6)$$

for some ρ . In (6), $\mathcal{C}(\mathbf{u})$ denotes the signal constraints required to be satisfied by the transmit signal. For example, when the number of the RF chains L is less than M , $\mathcal{C}(\mathbf{u})$ includes the sparsity constraint $\|\mathbf{u}\|_0 \leq L$. The formulation in (6) describes the RLS method which we discuss in the following sections. For simplicity, we drop the time index n in the sequel.

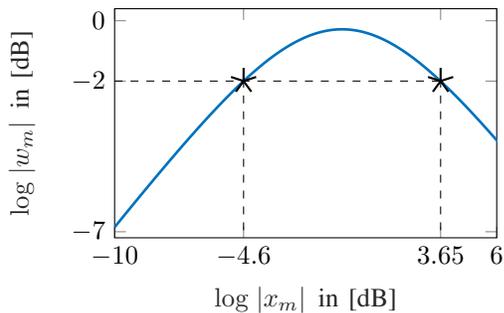


Fig. 1: An RF conversion function given by Saleh’s model with $\alpha = 2.092$ and $\beta = 1.247$ [11].

A. GLSE Precoding at the RF Stage

The RF transmit signal \mathbf{w} is directly found by solving

$$\mathbf{w} = \underset{\mathbf{v} \in \mathbb{W}^M}{\operatorname{argmin}} \operatorname{RSS}(\mathbf{v} | \rho, \mathbf{s}, \mathbf{H}) \text{ s.t. } \tilde{\mathcal{C}}(\mathbf{v}) \quad (7)$$

where $\tilde{\mathcal{C}}(\mathbf{v})$ contains the signal constraints in $\mathcal{C}(\mathbf{u})$ projected on \mathbb{W}^M with respect to the RF conversion function. Following the RLS method, this problem is equivalently solved by

$$\mathbf{w} = \underset{\mathbf{v} \in \mathbb{W}^M}{\operatorname{argmin}} \|\mathbf{H}^T \mathbf{v} - \sqrt{\rho} \mathbf{s}\|^2 + c(\mathbf{v}) \quad (8)$$

for some penalty $c(\cdot)$. The precoder in (8) recovers the GLSE precoding scheme at the RF stage [7]–[9]. The key differences here are: 1) The penalty $c(\mathbf{v})$ is chosen with respect to the RF conversion function $f_{\text{RF}}(\cdot)$. 2) The entries of the RF transmit signal should be projected back on the precoding support \mathbb{X} .

As the RF conversion function is known to the BS, the first task is tractable. However, the backward projection of the RF transmit entries to the precoding support cannot be uniquely done, as the $f_{\text{RF}}(\cdot)$ is not necessarily bijective.

B. Projecting RF Signals on the Precoding Support

Consider the precoded signal at the RF stage, i.e. \mathbf{w} . To project back the RF transmit entries on the precoding support, we need to solve $w_m = f_{\text{RF}}(x_m)$ for $m \in [M]$. Considering the typical characteristics of PAs, it is clear that there are multiple solutions for x_m . We hence need to select a desired solution among the available ones with respect to a metric. To clarify this latter statement, let us consider a simple RF conversion function. In Fig. 1, the output amplitude, i.e. $|w_m|$ of a sample PA is sketched against the amplitude of input symbol x_m using Saleh’s model in [11] with $\alpha = 2.0922$ and $\beta = 1.2466$. As the figure depicts, $\log|w_m| = -2$ dB is achieved at the output of the PA, when the input amplitude is set either $\log|x_m| = -4.6$ dB or $\log|x_m| = 3.65$ dB. This behavior comes from the significant nonlinearity of the PA in the saturation region. As we wish to restrict the power consumed in the system, we set the input symbol to the one whose amplitude is smaller, i.e., we set $\log|x_m| = -4.6$ dB.

For a generic $f_{\text{RF}}(\cdot)$, the backward projection of RF transmit entries on the precoding support can be formulated via the RLS method. Considering the pre-amplifier power constraint

as the selection measure, the approach for choosing the input symbol in the given example is formulated as

$$x_m = \underset{u \in \mathbb{X}}{\operatorname{argmin}} |u|^2 \text{ s.t. } w_m = f_{\text{RF}}(u). \quad (9)$$

Following Remark 1, one can reformulate (9) as

$$x_m = \underset{u \in \mathbb{X}}{\operatorname{argmin}} |u|^2 \text{ s.t. } |w_m - f_{\text{RF}}(u)|^2 \leq \epsilon. \quad (10)$$

The approach in (10) is equivalently given in form of a scalar RLS problem. The regularization term in this problem is quadratic, as the metric for choosing the input symbol is its power. Nevertheless, in a generic case, the metric can be of a different form taking into account also the phase of the input symbol.

Following the above discussion, the projection approach in (10) is represented in a general form as

$$x_m = \underset{u \in \mathbb{X}}{\operatorname{argmin}} |w_m - f_{\text{RF}}(u)|^2 + \theta r(u) \quad (11)$$

where $r(\cdot)$ is a regularization term describing the metric. For instance, in the case that we choose the transmit symbol based on its power, we have $r(u) = |u|^2$. θ is a tunable factor which controls the consumed power.

Remark 2: Note that the signal power before and after amplification defines two different parameters which are implicitly related. The power of transmit entries, i.e. signal before amplification, specifies the power consumed in the system while the power at the output of transmit RF chains mainly quantifies the multiuser interference at each UT. The RLS approach in (11), along with the GLSE scheme in (8), guarantees that both of these parameters are minimized given the set of available constraints on the transmit signal. In this respect, θ can be seen as a Lagrange multiplier which tunes the trade-off between the consumed power and the RSS.

IV. ALGORITHMIC IMPLEMENTATION VIA AMP

The computational complexity of the proposed scheme is dominated by the precoding in the RF stage. In fact, the projection on the precoding support deals with M parallel copies of a scalar optimization problem which is addressed tractably. The calculation of the RF transmit signal, however, requires to solve an optimization problem of size M within each transmission interval¹. In massive MIMO settings with large antenna arrays, such a task is not computationally tractable in practice. To address this issue, we propose an iterative algorithm based on AMP whose complexity scales linearly with M .

A. RF Stage Precoding via AMP

The GLSE precoding in (8) is mathematically equivalent to a *max-sum* problem in the Bayesian framework. This equivalence was studied in [13] where an iterative algorithm based on generalized AMP (GAMP) [10], [14] was proposed. This algorithm is adapted to the problem in (8) in Algorithm 1. In this algorithm, \mathbf{w} is constructed after T iterations. Here,

¹In practice, the update rate can be reduced to once per coherence time interval by block-wise precoding; see [7].

Algorithm 1 AMP-based Precoding in the RF Stage

Initiate For $k \in [K]$, let $\mathbf{y}_k(0) = \mathbf{0}$. For $m \in [M]$, set

$$\mathbf{w}_m(1) = \underset{\mathbf{v} \in \mathbb{W}}{\operatorname{argmin}} c(\mathbf{v}) \quad (12a)$$

$$\mathbf{R}_m^w(1) = [\nabla^2 c(\mathbf{w}_m(1))]^{-1} \quad (12b)$$

while $1 \leq t < T$

for $k \in [K]$ and $m \in [M]$ update

$$\mathbf{R}_k^v(t) = \sum_{m=1}^M \mathbf{Q}_{mk} \mathbf{R}_m^w(t) \mathbf{Q}_{mk}^\top \quad (13a)$$

$$\mathbf{v}_k(t) = \sum_{m=1}^M \mathbf{Q}_{mk} \mathbf{w}_m(t) - \mathbf{R}_k^v(t) \mathbf{y}_k(t-1) \quad (13b)$$

$$\mathbf{y}_k(t) = \mathbf{g}_{\text{out}}(\mathbf{v}_k(t), \mathbf{s}_k, \mathbf{R}_k^v(t)) \quad (13c)$$

$$\mathbf{R}_k^y(t) = -\nabla_{\mathbf{v}} \mathbf{g}_{\text{out}}(\mathbf{v}_k(t), \mathbf{s}_k, \mathbf{R}_k^v(t)) \quad (13d)$$

$$\mathbf{R}_m^u(t) = \left[\sum_{k=1}^K \mathbf{Q}_{mk}^\top \mathbf{R}_k^y(t) \mathbf{Q}_{mk} \right]^{-1} \quad (13e)$$

$$\mathbf{u}_m(t) = \mathbf{w}_m(t) + \mathbf{R}_m^u(t) \left[\sum_{k=1}^K \mathbf{Q}_{mk}^\top \mathbf{y}_k(t) \right] \quad (13f)$$

$$\mathbf{w}_m(t+1) = \mathbf{g}_{\text{in}}(\mathbf{u}_m(t), \mathbf{R}_m^u(t)) \quad (13g)$$

$$\mathbf{R}_m^w(t+1) = [\nabla_{\mathbf{u}} \mathbf{g}_{\text{in}}(\mathbf{u}_m(t), \mathbf{R}_m^u(t))] \mathbf{R}_m^u(t) \quad (13h)$$

end for

end while

Output: $\mathbf{w}_m(T)$ for $m \in [M]$.

- $\mathbf{w}_m(t)$ and \mathbf{s}_k are the augmented forms of the m -th RF symbol in iteration t , i.e. $w_m(t)$, and s_k , respectively.
- $\mathbf{R}_m^w(t)$, $\mathbf{R}_k^v(t)$, $\mathbf{R}_k^y(t)$ and $\mathbf{R}_m^u(t)$ are two-dimensional real square matrices, and \mathbf{Q}_{mk} is defined as

$$\mathbf{Q}_{mk} := \begin{bmatrix} \Re\{h_{mk}\} & -\Im\{h_{mk}\} \\ \Im\{h_{mk}\} & \Re\{h_{mk}\} \end{bmatrix} \quad (14)$$

where h_{mk} is the entry (m, k) of \mathbf{H} .

- $\mathbf{g}_{\text{out}}(\cdot)$ is the output thresholder given by

$$\mathbf{g}_{\text{out}}(\mathbf{v}, \mathbf{s}, \mathbf{R}) := \nabla_{\mathbf{v}} \min_{\mathbf{z} \in \mathbb{C}} G_{\text{out}}(\mathbf{z}, \mathbf{v}, \mathbf{s}, \mathbf{R}) \quad (15)$$

where $G_{\text{out}}(\cdot)$ is

$$G_{\text{out}}(\mathbf{z}, \mathbf{v}, \mathbf{s}, \mathbf{R}) = q(\mathbf{z} - \mathbf{v}, \mathbf{R}) + \|\mathbf{z} - \sqrt{\rho} \mathbf{s}\|^2 \quad (16)$$

with $q(\cdot)$ being $q(\mathbf{x}, \mathbf{R}) = (\mathbf{x}^\top \mathbf{R}^{-1} \mathbf{x})/2$.

- $\mathbf{g}_{\text{in}}(\cdot)$ is the input thresholding function and reads

$$\mathbf{g}_{\text{in}}(\mathbf{u}, \mathbf{R}) := \underset{\mathbf{w} \in \mathbb{W}}{\operatorname{argmin}} G_{\text{in}}(\mathbf{w}, \mathbf{u}, \mathbf{R}) \quad (17)$$

where $G_{\text{in}}(\cdot)$ is given by

$$G_{\text{in}}(\mathbf{w}, \mathbf{u}, \mathbf{R}) = q(\mathbf{u} - \mathbf{w}, \mathbf{R}) + c(\mathbf{w}). \quad (18)$$

B. Projection on the Precoding Support

To project the precoded RF signal \mathbf{w} back on the precoding support, we follow the approach proposed in Section III-B. To this end, transmit entry x_m for $m \in [M]$ is calculated via (11) for some θ and regularization term $r(\cdot)$. To tune θ , we note

- 1) As $\theta \downarrow 0$, (11) determines the minimizer of $r(\cdot)$ over the set of points u at which $w_m = f_{\text{RF}}(u)$. This tuning is suitable for RF conversion functions which accurately model the input-output characteristic of the RF chains, i.e. when $\epsilon \downarrow 0$ in Remark 1.
- 2) For $\theta \uparrow \infty$, (11) finds the minimizer of $r(\cdot)$ over \mathbb{C} . Such a setting corresponds to RF conversion models with high error, i.e. when ϵ is significantly large.

As a result, θ is tuned such that it monotonically increases with ϵ , where ϵ is error of the analytic model given by $f_{\text{RF}}(\cdot)$.

V. NUMERICAL INVESTIGATIONS

In this section, we investigate the performance of the proposed algorithm through numerical simulations. For this aim, we first specify the configuration which is being simulated.

A. System Configuration

We consider the case in which independent and identically distributed (i.i.d.) zero-mean and unit-variance Gaussian data symbols s_1, \dots, s_K are to be transmitted over the downlink channel. It is assumed that the channel experiences i.i.d. Rayleigh fading. This means that the entries of \mathbf{H} are i.i.d. Gaussian with zero-mean and variance $1/M$. The characteristics of each system component are illustrated in the sequel.

1) *RF chains*: The conversion function of an RF chain is assumed to be fully described via its PA, and the input-output characteristics of the PA is represented by the amplitude-to-amplitude and amplitude-to-phase model which reads

$$f_{\text{RF}}(x) = f_A(|x|) \exp\{j f_\Phi(|x|)\} \frac{x}{|x|}. \quad (19)$$

In this model, $f_A(\cdot)$ is the amplitude-to-amplitude conversion function which specifies the amplitude of the RF signal at the output of the PA. $f_\Phi(\cdot)$ further determines the nonlinear phase shift at the output which is a function of the input amplitude.

Well-known analytic formulations for $f_A(\cdot)$ and $f_\Phi(\cdot)$ are given by Saleh's model [11], [12] in which

$$f_A(\omega) = \frac{\alpha_A \omega}{1 + \beta_A \omega^2}, \quad f_\Phi(\omega) = \frac{\alpha_\Phi \omega^2}{1 + \beta_\Phi \omega^2}. \quad (20)$$

Here, (α_A, β_A) and $(\alpha_\Phi, \beta_\Phi)$ are non-negative scalars which are determined for a specific PA numerically via the method of least-squares. The model is assumed to have average error ϵ . This means that for the true output symbol w , we have

$$\mathbb{E}_x \{|w - f_{\text{RF}}(x)|^2\} \leq \epsilon \quad (21)$$

where x is the true input symbol, and $\mathbb{E}_x\{\cdot\}$ averages over all possible realizations of x .

2) *Precoder*: From the example in Fig. 1, we know that the PA's output is saturated at some level. This means that the RF transmit entries always satisfy $|w_m|^2 \leq P_{\text{out}}$ for some power P_{out} which is specified for each PA. As a result, we consider the RF constellation set as $\mathbb{W} = \{w \in \mathbb{C} : |w| \leq \sqrt{P_{\text{out}}}\}$.

The precoding support \mathbb{X} is further set to $\mathbb{X} = \mathbb{C}$. Following the discussions in Section IV-B, we need to set θ in (11) monotonically increasing in ϵ for backward projection of the RF transmit signal on \mathbb{X} . We consider the simple choice of $\theta = \epsilon$ and set $r(u) = |u|^2$.

For the sake of simplicity, we assume full transmit complexity, i.e. $L = M$, with limited average transmit power; hence, $c(\mathbf{w}) = \lambda \|\mathbf{w}\|^2$ for some λ . Nevertheless, scenarios with partially active arrays, i.e. $L \leq M$, are straightforwardly addressed by modifying $c(\cdot)$; see discussions in [7].

B. Performance Metrics

Following the discussions in Section III, we know that the RSS defined in Definition 1 quantifies the performance of the precoder, effectively. We hence define the performance metric with respect to the RSS. To this end, let $\mathbf{w} \in \mathbb{W}^M$ be the signal precoded directly at the RF stage via (8). The transmit signal $\mathbf{x} \in \mathbb{C}^M$ is calculated entry-wise from \mathbf{w} using (11). The true RF signal is then given by $\tilde{\mathbf{w}} = f_{\text{RF}}(\mathbf{x})$ which is in general different² from \mathbf{w} . In this case, the average RSS predicted by the RF-stage GLSE precoder is

$$D(\rho) = \frac{1}{K} \|\mathbf{H}^T \mathbf{w} - \sqrt{\rho} \mathbf{s}\|^2$$

for the given scaling factor ρ . However, the average RSS which is achieved in practice is

$$\tilde{D}(\rho) = \frac{1}{K} \|\mathbf{H}^T \tilde{\mathbf{w}} - \sqrt{\rho} \mathbf{s}\|^2.$$

$D(\rho)$ and $\tilde{D}(\rho)$ in general address the average distortion imposed by the multiuser interference at each UT. For effective design of the precoder and small ϵ , we have $\tilde{D}(\rho) \approx D(\rho)$.

C. Numerical Results

Fig. 2 plots $D(\rho)$ and $\tilde{D}(\rho)$ against the number of transmit antennas per user, i.e. $\xi = M/K$, for $\rho = 1$. The transmit array size is set to $M = 64$, and the parameters of the PA read

$$(\alpha_A, \beta_A) = (2.159, 1.152), \quad (\alpha_\Phi, \beta_\Phi) = (4.003, 9.104)$$

with $\epsilon = 0.05$. Considering the dynamic range of the PA, the peak output power on the RF stage is set to $P_{\text{out}} = 1$.

To compare the results with the benchmark, λ is tuned, such that the PAPR of the RF transmit signal is $\log \text{PAPR} = 5$ dB. The results are given for Algorithm 1, as well as directly solving (8) via CVX [15], [16].

It is observed that $D(\rho)$ closely matches $\tilde{D}(\rho)$ which indicates the efficiency of the backward projection. The figure further depicts the accuracy of Algorithm 1, as its results are tightly consistent with the direct simulations. For sake of comparison, we sketch two other plots: The first plot shows the

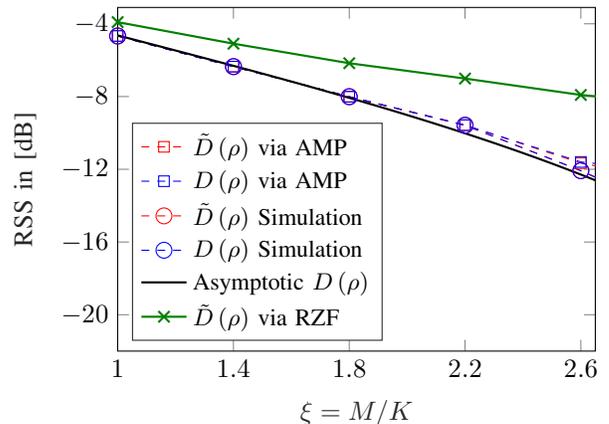


Fig. 2: Average RSS vs. per-user number of antennas for $\rho = 1$. The PAPR of the RF signal is set to $\log \text{PAPR} = 5$ dB.

asymptotic value of $D(\rho)$ derived in [7]. This plot is closely tracked by the finite-dimension simulations. The second plot shows $\tilde{D}(\rho)$ achieved via regularized zero forcing (RZF) precoding [17]. The RZF precoder in this case is tuned, such that the output PAPR remains $\log \text{PAPR} = 5$ dB. As the plot demonstrates, RZF precoding exhibits degraded performance. This is a result of unwanted distortion imposed by nonlinear characteristics of the RF chains. Due to the page limit, further numerical investigations are skipped and will be given in the extended version of the manuscript.

VI. CONCLUSIONS

The proposed precoding scheme for massive MIMO settings with nonlinear front-ends utilizes the RLS method to jointly invert the channel and compensate distortions caused by nonlinear RF chains. An AMP-based algorithm implements the proposed scheme with low complexity. Numerical investigations show performance enhancement compared to the classic precoding techniques. Although the main focus of this paper was on the PA, the results are straightforwardly extended to other non-ideal transceiver components.

REFERENCES

- [1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. on Wireless Com.*, vol. 9, no. 11, pp. 3590–3600, 2010.
- [2] J. Hoydis, S. Ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. on Sel. Areas in Com.*, vol. 31, no. 2, pp. 160–171, 2013.
- [3] R. Piazza, M. B. Shankar, and B. Ottersten, "Data predistortion for multicarrier satellite channels based on direct learning," *IEEE Trans. on Sig. Proc.*, vol. 62, no. 22, pp. 5868–5880, 2014.
- [4] D. Spano, M. Alodeh, S. Chatzinotas, and B. Ottersten, "Symbol-level precoding for the non-linear multiuser MISO downlink channel," *IEEE Trans. on Sig. Proc.*, 2017.
- [5] M. Alodeh, D. Spano, A. Kalantari, C. G. Tsinos, D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Symbol-level and multicast precoding for multiuser multiantenna downlink: A state-of-the-art, classification, and challenges," *IEEE Com. Sur. & Tut.*, vol. 20, pp. 1733–1757, 2018.
- [6] D. Spano, M. Alodeh, S. Chatzinotas, and B. Ottersten, "Faster-than-nyquist signaling through spatio-temporal symbol-level precoding for the multiuser MISO downlink channel," *IEEE Trans. on Wireless Com.*, vol. 17, no. 9, pp. 5915–5928, 2018.

²Note that (11) solves exactly $\mathbf{w} = f_{\text{RF}}(\mathbf{x})$ only when $\theta = 0$.

- [7] A. Bereyhi, M. A. Sedaghat, R. R. Müller, and G. Fischer, "GLSE precoders for massive MIMO systems: Analysis and applications," *arXiv preprint arXiv:1808.01880*, 2018.
- [8] A. Bereyhi, M. A. Sedaghat, S. Asaad, and R. Mueller, "Nonlinear precoders for massive MIMO systems with general constraints," in *21 Int. ITG Wor. on Smart Antennas (WSA)*, 2017.
- [9] A. Bereyhi, M. A. Sedaghat, and R. R. Müller, "Asymptotics of nonlinear LSE precoders with applications to transmit antenna selection," in *IEEE Int. Symp. on Inf. Th. (ISIT)*, 2017.
- [10] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *IEEE Int. Symp. on Inf. Th. (ISIT)*, 2018.
- [11] A. A. Saleh, "Frequency-independent and frequency-dependent nonlinear models of TWT amplifiers," *IEEE Trans. on Com.*, vol. 29, no. 11, pp. 1715–1720, 1981.
- [12] M. O'Droma, S. Meza, and Y. Lei, "New modified saleh models for memoryless nonlinear power amplifier behavioural modelling," *IEEE Com. Letters*, vol. 13, no. 6, 2009.
- [13] A. Bereyhi, M. A. Sedaghat, and R. R. Müller, "Precoding via approximate message passing with instantaneous signal constraints," in *Int. Zurich Seminar on Inf. and Com. (IZS)*, 2018.
- [14] S. Rangan, P. Schniter, E. Riegler, A. K. Fletcher, and V. Cevher, "Fixed points of generalized approximate message passing with arbitrary matrices," *IEEE Trans. on Inf. Th.*, vol. 62, no. 12, pp. 7464–7474, 2016.
- [15] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," <http://cvxr.com/cvx>, Mar. 2014.
- [16] —, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lec. Notes in Control and Inf. Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.
- [17] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication- Part I: channel inversion and regularization," *IEEE Trans. on Com.*, vol. 53, no. 1, pp. 195–202, 2005.