Quantized Penalty Gradient Algorithm For Massive MIMO Systems with Low-Resolution ADCs

Qiqiang Chen, Zheng Wang, *Senior Member, IEEE*, Chenhao Qi, *Senior Member, IEEE*, Feng Shu, *Senior Member, IEEE*, and Yongming Huang, *Fellow, IEEE*

***Abstract*—In this paper, we propose a quantized penalty gradient (QPG) detection algorithm for massive multiple-input multiple-output (MIMO) systems with low-resolution analog-to-digital converters (ADCs). To tackle the challenges of maximum likelihood (ML) detection under discrete constraints, we reformulate the detection problem into an unconstrained optimization by introducing two customized penalty functions that promote alignment between the estimated signals and target constellation set. Based on this, the QPG algorithm is developed to efficiently solve the resulting problem, achieving competitive detection performance with only second-order computational complexity. We further provide a theoretical analysis establishing the Lipschitz continuity of the objective function, which guarantees the monotonic descent property of QPG and ensures its convergence. Moreover, we prove that QPG efficiently finds the local minima with an accessible linear convergence rate, thus leading to an explicit trade-off between detection performance and computational complexity. Finally, simulation results confirm the significant performance gains of QPG over the conventional quantized detectors across various channel conditions, while maintaining low computational complexity.**

***Index Terms*—Massive MIMO, quantized MIMO detection, low-resolution ADCs, penalty function.**

I. INTRODUCTION

MASSIVE multiple-input multiple-output (MIMO) systems have emerged as a cornerstone technology for beyond fifth-generation (B5G) wireless communications, offering significantly enhanced capacity, ultra-fast data rates, and superior energy efficiency [1]–[4]. Nevertheless, the substantial increase of antenna number at the base station (BS) introduces considerable challenges, particularly in terms of power consumption and hardware costs [5]–[7]. Specifically, as one of the most power consuming components, the power consumption of the analog-to-digital converter (ADC) units scales exponentially with the resolution and linearly with the sampling rate [8]. To mitigate these expenses, low-resolution ADCs, e.g., 1-3 bits, have emerged as an efficient solution [9]–[33]. However, despite their energy efficiency, low-resolution ADCs introduce severe nonlinear distortions due to the quantization, transforming the wireless channel into a non-linear and non-Gaussian system [9]. Thus, data detection in such systems becomes far more challenging than conventional full-resolution ADC architectures.

Linear detection algorithms for quantized massive MIMO systems have been extensively studied [10]–[14]. A straightforward approach is to ignore the quantization effects and directly apply conventional linear receivers designed for infinite-resolution systems [10], [11]. However, such quantization-unaware methods suffer from significant performance degradation when deployed in quantized systems. To mitigate this issue, a more refined strategy uses the Bussgang-based linear approximation, which accounts for quantization by modeling the distortion as additive white Gaussian noise (AWGN), uncorrelated with the quantizer's output [12]. This leads to improved performance and reduced error floors compared to conventional methods. Based on this model, the Bussgang-based minimum mean-squared error (BMMSE) [13] and successive interference cancellation (BSIC) algorithms [14] have been extensively studied. Nevertheless, the assumption of uncorrelated Gaussian quantization noise limits the accuracy of Bussgang model, especially at high signal-to-noise ratio (SNR) [6]. Consequently, Bussgang-based detectors exhibit high detection error floors under such conditions.

Instead of linearizing the input-output relation, state-of-the-art receivers adopt nonlinear optimizations of the likelihood function to improve the detection performance [15]–[27]. Specifically, the maximum likelihood (ML) detectors for quantized systems have been formulated using Gaussian cumulative distribution function (CDF) [15], [16]. Given the complexity of the original ML detection in massive systems, a near-ML (nML) algorithm has been proposed in [16] as an efficient alternative. Similarly, sphere decoding (SD) techniques were introduced in [17], [18], while [19] applies the weighted Hamming distance (WHD) for final decision. The detectors in [20]–[22] adopts a two-stage approach to improve performance, where the second stage refines estimates via an exhaustive neighborhood searching. However, the complexity of these methods grows exponentially with the number of transmit antennas, rendering them impractical for massive MIMO systems. To address hardware impairments, iterative detection methods have been developed, including Newton method (NM) [23], alternating direction method of multipliers (ADMM) [24] and forward-backward splitting (FBS) [25].

This work was supported in part by National Science and Technology Major Projects of China under Grant 2025ZD1301800, in part by National Natural Science Foundation of China under Grant 62371124, 61720106003, 62225107, U22A2002, in part by the Hainan Province Science and Technology Special Fund under Grant ZDYF2024GXJS292, and in part by the SEU Innovation Capability Enhancement Plan for Doctoral Students under Grant CXJH_SEU 26102. *(Corresponding author: Zheng Wang.)*

Q. Chen, Z. Wang, C. Qi and Y. Huang are with School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: q.chen@seu.edu.cn; wznuaa@gmail.com; qch@seu.edu.cn; huangym@seu.edu.cn).

F. Shu is with the School of Information and Communication Engineering, Hainan University, Haikou 570228, China (email: shufeng0101@163.com).

Other nonlinear detection methods based on probabilistic inference, such as generalized approximate message passing (GAMP) and Bayesian inference, have also been explored in [26], [27].

Recently, gradient-based methods have emerged as effective low-complexity solutions for quantized massive MIMO detection [16], [28]–[31]. However, directly processing the Gaussian CDF in the likelihood function suffers from numerical instability, especially at high SNRs [28]. To this end, the Gaussian CDF is replaced with a sigmoid function in [29], which relaxes the discrete constellation constraint into an unconstrained optimization problem. Based on this approximation, the gradient descent (GD) detection method in [29] achieves low complexity. Subsequently, the detection performance is further improved in [30] by incorporating a neighborhood search after GD iterations. Moreover, the approximate likelihood model is generalized and its properties and convergence behavior are analyzed in [31]. Despite these advances, the performance of the GD-based methods still remains limited due to their disregard for the discrete constellation constraint.

To bridge the performance gap with ML detection while maintaining low complexity, we propose a quantized penalty gradient (QPG) algorithm for massive MIMO systems equipped with low-resolution ADCs. The main contributions of this work are summarized as follows:

- Firstly, we introduce two customized penalty terms into the ML function and reformulate the discrete constraints into an unconstrained optimization problem. These penalty functions are carefully crafted to align the estimates and the discrete constellation set, thus bringing extra nonlinear gains into the detection performance.
- Secondly, with respect to the established objective function, we develop the QPG algorithm as a low-complexity but high-accuracy detector for quantized massive MIMO systems. It achieves competitive detection performance with only second-order computational complexity. Furthermore, we establish the monotonic descent and a linear convergence rate of the proposed QPG algorithm, which precisely quantifies the trade-off between detection performance and computational complexity.
- Finally, we perform detailed simulations over various channel conditions, comparing the proposed QPG method with existing quantized MIMO detectors in the literature. In addition, a comprehensive complexity analysis is presented to demonstrate the low computational cost of QPG.

The remainder of this paper is organized as follows: Section II provides a brief overview of ML detection problem in uplink quantized massive MIMO systems. In Section III, the proposed QPG algorithm is introduced, and its computational complexity is analyzed. Section IV establishes the Lipschitz continuity of the ML function. Subsequently, the monotonic decent property and linear convergence rate of the QPG algorithm are derived to guarantee its performance gains. Section V presents the simulation results for the proposed QPG detection in uplink quantized massive MIMO systems. Finally, Section VI concludes the paper.

Notation: Throughout this paper, matrices are denoted by uppercase boldface letters while column vectors are represented by lowercase boldface letters. The transpose and inverse of a matrix $\mathbf{B}$ are denoted by $\mathbf{B}^T$ and $\mathbf{B}^{-1}$, respectively. The (i, j)-th entry of $\mathbf{B}$ is represented as $b_{i,j}$ and $\mathbb{E}\{\cdot\}$ refers to the expectation. The Hadamard product, expressed as $\odot$, performs element-wise multiplication between two vectors. The operator $\operatorname{diag}\{\mathbf{B}\}$ extracts the diagonal entries of the square matrix $\mathbf{B}$, while $\operatorname{sign}(\cdot)$ denotes the element-wise sign function. The norm $\|\cdot\|$ denotes the ℓ_2-norm for both vectors and matrices. Superscripts indicating iterations are written in parentheses. Finally, $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote the real and the imaginary components, respectively.

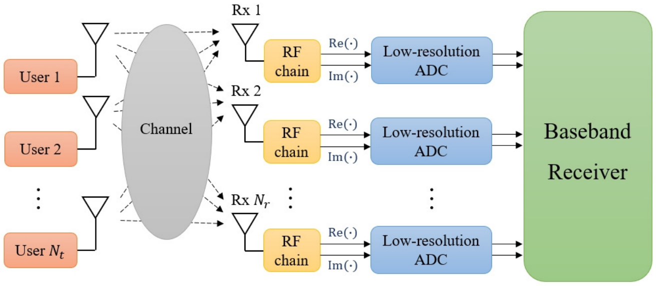


Fig. 1. Illustration of the quantized massive MIMO system with N_t single-antenna users and a BS equipped with N_r antennas, N_r radio-frequency (RF) chains, and $2N_r$ low-resolution ADCs.

II. Preliminary

This section begins with an introduction to the uplink massive MIMO system model with low-resolution ADCs. Subsequently, a detailed explanation of the quantized signal detection framework is provided.

A. System Model with Low-Resolution ADCs

We consider an uplink massive MIMO system, as shown in Fig. 1, where N_t single antenna users communicate with a BS equipped with N_r receive antennas $(N_r \geq N_t)$. Let $\bar{\mathbf{x}} \in \mathcal{O}^{N_t}$ denote the transmitted vector from the discrete complex M-quadrature amplitude modulation (QAM) constellation set $\mathcal{O}^{N_t}$. $\bar{\mathbf{H}} \in \mathbb{C}^{N_r \times N_t}$ represents the channel matrix, which is assumed to be flat fading. Let $\bar{\mathbf{r}} \in \mathbb{C}^{N_r}$ be the unquantized received signal at BS, which is given by

$$\bar{\mathbf{r}} = \bar{\mathbf{H}}\bar{\mathbf{x}} + \bar{\mathbf{n}}, \tag{1}$$

where $\bar{\mathbf{n}} \in \mathbb{C}^{N_r}$ denotes the addictive white Gaussian noise (AWGN) with zero mean and covariance matrix $\sigma_{\bar{n}}^2 \mathbf{I}_{N_r}$. Each received analog signal is then quantized by a pair of low-resolution ADCs, yielding the quantized received signal as

$$\bar{\mathbf{y}} = \mathcal{Q}_b(\bar{\mathbf{r}}) = \mathcal{Q}_b(\Re\{\bar{\mathbf{r}}\}) + j\mathcal{Q}_b(\Im\{\bar{\mathbf{r}}\}), \tag{2}$$

where $\mathcal{Q}_b(\cdot)$ denotes the b-bit quantization function.

To facilitate algorithm design using real-valued inputs, we transform the complex-valued model in (1) into an equivalent real-valued system of dimensions $2N_r \times 2N_t$ as follows

$$\begin{aligned} \mathbf{H} &= \begin{bmatrix} \Re\{\bar{\mathbf{H}}\} & -\Im\{\bar{\mathbf{H}}\} \\ \Im\{\bar{\mathbf{H}}\} & \Re\{\bar{\mathbf{H}}\} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} \Re\{\bar{\mathbf{x}}\} \\ \Im\{\bar{\mathbf{x}}\} \end{bmatrix}, \\ \mathbf{r} &= \begin{bmatrix} \Re\{\bar{\mathbf{r}}\} \\ \Im\{\bar{\mathbf{r}}\} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \Re\{\bar{\mathbf{y}}\} \\ \Im\{\bar{\mathbf{y}}\} \end{bmatrix}, \quad \mathbf{n} = \begin{bmatrix} \Re\{\bar{\mathbf{n}}\} \\ \Im\{\bar{\mathbf{n}}\} \end{bmatrix}. \end{aligned} \tag{3}$$

For simplicity of notation, we define $N = 2N_r$ and $K = 2N_t$ from this point onward. In this way, complex constellation $\mathcal{O}^{N_t}$ is mapped to a real-valued $\sqrt{M}$-amplitude-shift keying (ASK) constellation set $\mathcal{X}^K$, where $\mathcal{X} = \{\pm 1, \pm 3, \ldots, \pm(\sqrt{M}-1)\}$. Accordingly, $\mathbf{x} \in \mathcal{X}^K$, $\mathbf{r}, \mathbf{n} \in \mathbb{R}^N$, and $\mathbf{H} \in \mathbb{R}^{N \times K}$, while the noise vector satisfies $\mathbf{n} \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I}_N)$ with variance $\sigma_n^2 = \sigma_{\bar{n}}^2/2$. Thus, the received signal at the BS can be expressed as

$$\mathbf{r} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{4}$$

$$\mathbf{y} = \mathcal{Q}_b(\mathbf{r}). \tag{5}$$

Specifically, r_i is mapped to a quantized label $y_i = a_l$ if $r_i \in (c_{l-1}, c_l]$, where $\mathcal{C} = \{c_0, \ldots, c_{2^b}\}$ defines the quantization thresholds and $\mathcal{A} = \{a_1, a_2, \ldots, a_{2^b}\}$ denotes the quantization alphabet. These values are determined based on a non-uniform quantization scheme optimized for a standard Gaussian input [34]. To account for varying signal energy, both thresholds and quantization levels are scaled by the standard deviation of the received signal, given by

$$\sigma_r = \sqrt{\frac{P_t + \sigma_{\bar{n}}^2}{2}}, \tag{6}$$

where $P_t = K \cdot \mathbb{E}\{|x_i|^2\}$ denotes the transmitter power [32]. Throughout this work, the detection problem is formulated with respect to a fixed channel realization $\mathbf{H}$ and the corresponding quantized observation $\mathbf{y}$ obtained from a particular transmission instance. The objective of the detection algorithm is to recover the transmitted $\sqrt{M}$-ASK symbol vector $\mathbf{x}$ for this specific realization.

B. Signal Detection-Maximum Likelihood Framework

The maximum likelihood (ML) detection problem for low-resolution massive MIMO systems is formulated in [15] as

$$\hat{\mathbf{x}}_{\text{ML}} = \arg\min_{\mathbf{x} \in \mathcal{X}^K} \sum_{i=1}^{N} -\log\left[\Phi\left(a_i^{\text{up}}\right) - \Phi\left(a_i^{\text{low}}\right)\right], \tag{7}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard normal distribution $\mathcal{N}(0,1)$, and $\rho = 1/\sigma_n^2$ is the inverse of noise variance. The vector $\mathbf{h}_i^T$ refers to the i-th row of channel matrix $\mathbf{H}$. $a_i^{\text{up}} = \sqrt{2\rho}\left(q_i^{\text{up}} - \mathbf{h}_i^T\mathbf{x}\right)$ and $a_i^{\text{low}} = \sqrt{2\rho}\left(q_i^{\text{low}} - \mathbf{h}_i^T\mathbf{x}\right)$, with q_i^{up} and q_i^{low} denoting the upper and lower quantization thresholds to y_i, respectively. In the special case of 1-bit quantization, the general ML problem in (7) simplifies to

$$\hat{\mathbf{x}}_{\text{ML,1-bit}} = \arg\min_{\mathbf{x} \in \mathcal{X}^K} \sum_{i=1}^{N} -\log\Phi\left(\sqrt{2\rho}\, v_i \mathbf{h}_i^T\mathbf{x}\right), \tag{8}$$

where $v_i = \text{sign}(y_i)$ captures the polarity of the quantized output, i.e., $v_i = +1$ if $y_i \geq 0$ and $v_i = -1$ otherwise [16].

To ease the evaluation of Gaussian CDF $\Phi(\nu)$, we approximate it using Sigmoid function as

$$\Phi(\nu) \approx \sigma(c\nu) = \frac{1}{1 + e^{-c\nu}}, \tag{9}$$

where $c = 1.702$ is a constant, and the approximation error is uniformly bounded by $|\Phi(\nu) - \sigma(c\nu)| \leq 0.0095$. This leads to the following SNR-explicit signal detection problem [29]

$$\min_{\mathbf{x} \in \mathcal{X}^K} h(\mathbf{x}) = \begin{cases} \sum_{i=1}^{N} -\log\sigma\left(c\sqrt{2\rho}\, v_i \mathbf{h}_i^T\mathbf{x}\right), & \text{if } b = 1, \\ \sum_{i=1}^{N} -\log\left[\sigma\left(c a_i^{\text{up}}\right) - \sigma\left(c a_i^{\text{low}}\right)\right], & \text{if } b \geq 2. \end{cases} \tag{10}$$

Although $h(\mathbf{x})$ closely tracks the original ML problem, its gradient scales with ρ, complicating step-size selection and increasing computational cost. To simplify optimization, [31] removes the positive factor $c\sqrt{2\rho}$ and adopts the following SNR-normalized surrogate

$$\min_{\mathbf{x} \in \mathcal{X}^K} f(\mathbf{x}) = \begin{cases} \sum_{i=1}^{N} -\log\sigma\left(v_i \mathbf{h}_i^T\mathbf{x}\right), & \text{if } b = 1, \\ \sum_{i=1}^{N} -\log\left[\sigma\left(b_i^{\text{up}}\right) - \sigma\left(b_i^{\text{low}}\right)\right], & \text{if } b \geq 2, \end{cases} \tag{11}$$

with $b_i^{\text{up}} = q_i^{\text{up}} - \mathbf{h}_i^T\mathbf{x}$ and $b_i^{\text{low}} = q_i^{\text{low}} - \mathbf{h}_i^T\mathbf{x}$. While $f(\mathbf{x})$ incurs a slight performance loss relative to $h(\mathbf{x})$, it becomes nearly equivalent at high SNR [31]. More importantly, its gradient is independent of ρ, enabling more efficient step-size selection. Therefore, we adopt the SNR-normalized form $f(\mathbf{x})$ for subsequent development and analysis.

To handle the discrete feasible set in quantized massive MIMO detection, prior works in [29]–[31] try to relax discrete constraint $\mathcal{X}^K$ and apply gradient-based methods to solve the resulting continuous optimization problem. However, such a relaxation ignores the inherent discrete nature of the constellation set, resulting in significant performance degradation in low-resolution systems.

III. The Proposed Quantized Detection Scheme

In this section, we first reformulate the signal detection problem using customized penalty functions to handle discrete constraints. Based on this, an efficient QPG algorithm is proposed to solve the relaxed problem. We further accelerate the optimization by designing a low-cost proximal step and analyze the overall computational complexity.

A. Customized Penalty Functions

Since the transmitted symbols belong to the ASK constellation set $\mathcal{X}^K$, each element $x_i \in \mathcal{X}$ for $i = 1, 2, \ldots, K$, satisfies the following constrains

$$\cos(\pi x_i) + 1 = 0, \tag{12}$$

$$-\sqrt{M} + 1 \leq x_i \leq \sqrt{M} - 1, \tag{13}$$

where the latter condition is commonly known as the box constraint [35]. Therefore, the signal detection problem in (11) can be rewritten as

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^K} \quad & f(\mathbf{x}) && \text{(14a)} \\ \text{s.t.} \quad & \cos(\pi x_i) + 1 = 0, \quad \forall i, && \text{(14b)} \\ & -\sqrt{M} + 1 \leq x_i \leq \sqrt{M} - 1. \quad \forall i. && \text{(14c)} \end{aligned}$$

However, constraint (14b) is inherently non-convex, as each x_i is restricted to a discrete and infinite set of odd integers. Consequently, solving this problem requires an exhaustive search over all possible symbol combinations in $\mathcal{X}^K$, leading to an exponential growth in computational complexity as the signal dimension increases. To mitigate this issue, we introduce a term proportional to the violation of constellation constraints and relax the problem in (14) as the following unconstrained form

$$\min_{\mathbf{x}\in\mathbb{R}^K} \quad F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}), \tag{15}$$

with the Cosine penalty function

$$g(\mathbf{x}) = \alpha \sum_{i=1}^{K} \cos(\pi x_i), \tag{16}$$

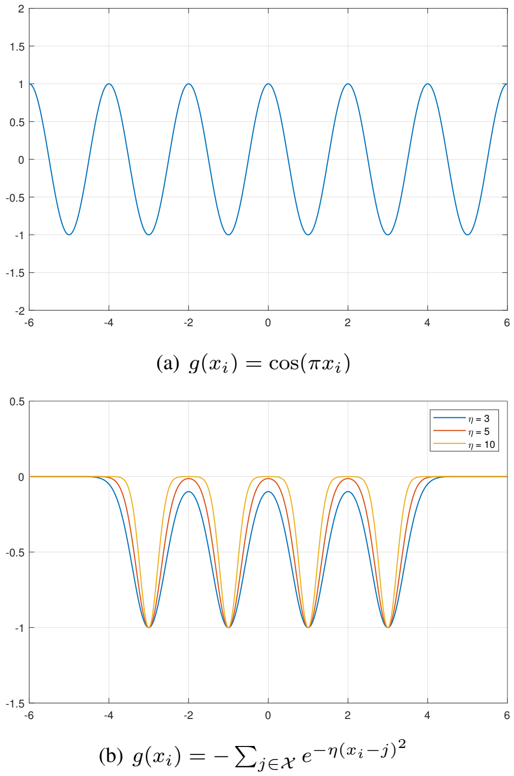
where $\alpha > 0$ controls the influence of penalty term. To further enforce proximity to the discrete constellation points, we design a more complex Gaussian penalty function as follows

$$g(\mathbf{x}) = -\alpha \sum_{i=1}^{K} \sum_{j\in\mathcal{X}} e^{-\eta(x_i-j)^2}, \tag{17}$$

where $\eta > 0$ determines the shape of Gaussian penalty function.

Fig. 2 illustrates the behavior of two different penalty functions under 16-QAM, with various values of η, $\alpha = 1$, and $\mathcal{X} = \{\pm 1, \pm 3\}$. Both penalty functions reach their local minima at the constellation points, thereby encouraging x_i to align with $\mathcal{X}$. Notably, the Gaussian penalty function approaches zero when $|x_i| \geq 4$, which effectively confines the solution within the vicinity of valid constellation points. Moreover, as η increases, the Gaussian penalty function becomes sharper around these points, promoting a steeper descent toward valid constellation values and thus accelerating the convergence.

For $g(\mathbf{x})$, one might choose a very large α to strictly enforce the constraints. However, excessive values of α may lead to numerical instability or ill-conditioning, rendering the optimization infeasible [36], [37]. In Section III-C, we analyze the selection of α to maintain convexity of the proximal objective function, which enables efficient optimization using the Newton method.



(a) $g(x_i) = \cos(\pi x_i)$

(b) $g(x_i) = -\sum_{j\in\mathcal{X}} e^{-\eta(x_i-j)^2}$

Fig. 2. Illustration of the Cosine and Gaussian penalty terms with 16-QAM.

B. Quantized Penalty Gradient (QPG) Algorithm

We now focus on efficiently solving the detection problem in (15). Motivated by the proximal gradient method in [38], we adopt a similar iterative structure to tackle the composite objective $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$. This framework is particularly suitable when f is convex [31], and g is nonconvex but admits a tractable proximal operator, as it allows their updates to be decoupled within the optimization process. In the proposed QPG algorithm, the variable $\mathbf{z}^{(t)}$ is updated based on the vector $\mathbf{s}^{(t)}$ as follows

$$\mathbf{z}^{(t)} = \mathbf{s}^{(t)} - \mu \nabla f(\mathbf{s}^{(t)}), \tag{18}$$

where $t = 1, 2, ..., T_{\text{max}}$ presents the iteration index, and μ denotes the step-size. Gradient computation $\nabla f(\mathbf{s}^{(t)})$ varies according to the quantization resolution. Specifically, for 1-bit quantization case, the gradient is computed as

$$\nabla f(\mathbf{s}^{(t)}) = -\mathbf{G}^T \sigma\left(-\mathbf{G}\mathbf{s}^{(t)}\right), \tag{19}$$

where $\mathbf{G} = \text{diag}\{\text{sign}(\mathbf{y})\}\,\mathbf{H} \in \mathbb{R}^{N\times N}$. For the case of few-bit quantization, the gradient takes the form

$$\nabla f(\mathbf{s}^{(t)}) = -\mathbf{H}^T\left[\mathbf{1} - \sigma(\mathbf{H}\mathbf{s}^{(t)} - \mathbf{q}^{\text{up}}) - \sigma(\mathbf{H}\mathbf{s}^{(t)} - \mathbf{q}^{\text{low}})\right], \tag{20}$$

where $\mathbf{q}^{\text{up}} = [q_1^{\text{up}}, \dots, q_N^{\text{up}}]^T$ and $\mathbf{q}^{\text{low}} = [q_1^{\text{low}}, \dots, q_N^{\text{low}}]^T$ represent the upper and lower quantization thresholds, respectively. Following the gradient update, a proximal operation is applied to $\mathbf{z}^{(t)}$ as

$$\mathbf{x}^{(t)} = \text{prox}_{\mu g}(\mathbf{z}^{(t)}) = \arg\min_{\mathbf{x}} \left\{\frac{1}{2}\|\mathbf{x} - \mathbf{z}^{(t)}\|^2 + \mu g(\mathbf{x})\right\}. \tag{21}$$

Remark 1: The proximal step can be interpreted as a form of *soft symbol projection*, where each update not only minimizes the objective but also naturally guides the solution toward the constellation points. Moreover, this structure flexibly supports various penalty functions, such as Cosine and Gaussian terms, without requiring modifications to the algorithm.

Next, a momentum update is performed by

$$\mathbf{v}^{(t)} = \mathbf{x}^{(t)} + \beta^{(t)}(\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}), \tag{22}$$

where $\beta^{(t)} \in (0, 1]$ denotes the accelerated parameter. After that, the selection of estimate $\mathbf{s}^{(t+1)}$ is based on the cost function evaluation in (15). If the objective function value of $\mathbf{x}^{(t)}$ is lower than that of $\mathbf{v}^t$, i.e., $F(\mathbf{x}^{(t)}) \leq F(\mathbf{v}^{(t)})$, then $\mathbf{x}^{(t)}$

is selected for the next iteration, where acceleration parameter $\beta^{(t)}$ is updated as

$$\mathbf{s}^{(t+1)} = \mathbf{x}^{(t)}, \tag{23}$$

$$\beta^{(t+1)} = \varrho\beta^{(t)}. \tag{24}$$

Otherwise, $\mathbf{v}^{(t)}$ is selected, and the updates become

$$\mathbf{s}^{(t+1)} = \mathbf{v}^{(t)}, \tag{25}$$

$$\beta^{(t+1)} = \min\left\{\frac{\beta^{(t)}}{\varrho}, 1\right\}. \tag{26}$$

Here, $\varrho \in (0,1)$ is a control parameter for the acceleration adjustment.

C. Efficient Proximal Operation

To solve the proximal operation in (21) efficiently, we first present the following Theorem to ensure the convexity of the proximal subproblem.

Theorem 1. *Given that the penalty parameter* α *in* $g(\mathbf{x})$ *satisfies*

$$\alpha \leq \min\left\{\frac{1}{\pi^2\mu}, \frac{1}{4\mu\eta e^{-\frac{3}{2}}}\right\}, \tag{27}$$

the proximal objective function in (21) *remains convex.*

Proof. To examine the convexity of the proximal objective in (21), we first consider the Cosine penalty in (16), which leads to the following optimization problem

$$x_i^{(t)} = \arg\min_{x_i}\left\{\frac{1}{2}\|x_i - z_i^{(t)}\|^2 + \alpha\mu\cos(\pi x_i)\right\}, \quad \forall i. \tag{28}$$

Defining $w(x_i) = \frac{1}{2}\|x_i - z_i^{(t)}\|^2 + \alpha\mu\cos(\pi x_i)$, its gradient and Hessian are given by

$$\nabla w(x_i) = x_i - z_i^{(t)} - \alpha\mu\pi\sin(\pi x_i), \tag{29}$$

$$\mathcal{H}_w(x_i) = 1 - \alpha\mu\pi^2\cos(\pi x_i). \tag{30}$$

Therefore, to maintain convexity, we require $\mathcal{H}_w(x_i) \geq 0$, which corresponds to the following constraint

$$\alpha \leq \frac{1}{\pi^2\mu}. \tag{31}$$

As for the Gaussian penalty in (17), the proximal objective becomes $k(x_i) = \frac{1}{2}\|x_i - z_i^{(t)}\|^2 - \alpha\mu\sum_{j\in\mathcal{X}} e^{-\eta(x_i-j)^2}$, with its gradient and Hessian expressed as

$$\nabla k(x_i) = x_i - z_i^{(t)} + \alpha\mu\sum_{j\in\mathcal{X}} e^{-\eta(x_i-j)^2}[2\eta(x_i - j)], \tag{32}$$

$$\mathcal{H}_k(x_i) = 1 + \alpha\mu\sum_{j\in\mathcal{X}} e^{-\eta(x_i-j)^2}[2\eta - 4\eta^2(x_i - j)^2]. \tag{33}$$

Thus, convexity requires

$$\alpha \leq \frac{1}{4\mu\eta e^{-\frac{3}{2}}}. \tag{34}$$

The detailed derivation of this bound is omitted for brevity. Finally, by incorporating the results from (31) and (34), the proof is thus completed. □

Algorithm 1: The Proposed QPG Detection Algorithm for Quantized Uplink Massive MIMO Systems

```
Input : $\mathbf{y}$, $\mathbf{H}$, $\mathbf{q}^{\text{up}}$, $\mathbf{q}^{\text{low}}$, $\mu$, $\alpha$, $\eta$, $\beta^{(1)}$, $\varrho$, $T_{\max}$
Output : estimated transmit signal $\widehat{\mathbf{x}}$
Initialization: $\mathbf{s}^{(1)} = \mathbf{x}^{(0)} = \mathbf{0}$, $\mathbf{G} = \operatorname{diag}\{\operatorname{sign}(\mathbf{y})\}\mathbf{H}$
for $t = 1, 2, \ldots, T_{\max}$ do
    // 1-bit quantization:
    $\nabla f(\mathbf{s}^{(t)}) = -\mathbf{G}^T\sigma\left(-\mathbf{G}\mathbf{s}^{(t)}\right)$
    // few-bit quantization:
    $\nabla f(\mathbf{s}^{(t)}) = -\mathbf{H}^T\left[\mathbf{1} - \sigma(\mathbf{H}\mathbf{s}^{(t)} - \mathbf{q}^{\text{up}}) - \sigma(\mathbf{H}\mathbf{s}^{(t)} - \mathbf{q}^{\text{low}})\right]$
    $\mathbf{z}^{(t)} = \mathbf{s}^{(t)} - \mu\nabla f(\mathbf{s}^{(t)})$
    for $i = 1, 2, \ldots, K$ do
        // Cosine penalty function:
        $x_i^{(t)} = z_i^{(t)} + \frac{\alpha\mu\pi\sin(\pi z_i^{(t)})}{1-\alpha\mu\pi^2\cos(\pi z_i^{(t)})}$
        // Gaussian penalty function:
        $x_i^{(t)} = z_i^{(t)} - \frac{\alpha\mu\sum_{j\in\mathcal{X}} e^{-\eta(z_i^{(t)}-j)^2}[2\eta(z_i^{(t)}-j)]}{1+\alpha\mu\sum_{j\in\mathcal{X}} e^{-\eta(z_i^{(t)}-j)^2}[2\eta-4\eta^2(z_i^{(t)}-j)^2]}$
    end for
    $\mathbf{v}^{(t)} = \mathbf{x}^{(t)} + \beta^{(t)}(\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)})$
    if $F(\mathbf{x}^{(t)}) \leq F(\mathbf{v}^{(t)})$ then
        $\mathbf{s}^{(t+1)} = \mathbf{x}^{(t)}$
        $\beta^{(t+1)} = \varrho\beta^t$
    else
        $\mathbf{s}^{(t+1)} = \mathbf{v}^{(t)}$
        $\beta^{(t+1)} = \min\left\{\frac{\beta^{(t)}}{\varrho}, 1\right\}$
    end if
end for
output $\widehat{\mathbf{x}} = \lceil\mathbf{x}^{(T_{\max})}\rfloor_{\mathcal{Q}} \in \mathcal{X}^K$
```

Note that this constraint is derived based on a worst-case bound $\cos(\nu) \geq -1$ for $\forall\nu \in \mathbb{R}$, and thus serves as a sufficient yet conservative condition to ensure the convexity of the proximal objective. Its practical applicability will be further investigated through the simulation results in Section V.

Given the established convexity, the proximal operation in (21) can be efficiently solved using the Newton method [39]. Specifically, for each x_i, we apply a single Newton iteration

$$x_i^{(t)} = z_i^{(t)} + \frac{\alpha\mu\pi\sin(\pi z_i^{(t)})}{1 - \alpha\mu\pi^2\cos(\pi z_i^{(t)})} \tag{35}$$

for the Cosine penalty function, and

$$x_i^{(t)} = z_i^{(t)} - \frac{\alpha\mu\sum_{j\in\mathcal{X}} e^{-\eta(z_i^{(t)}-j)^2}[2\eta(z_i^{(t)} - j)]}{1+\alpha\mu\sum_{j\in\mathcal{X}} e^{-\eta(z_i^{(t)}-j)^2}[2\eta - 4\eta^2(z_i^{(t)} - j)^2]} \tag{36}$$

for the Gaussian penalty function.

The initial choice of $\mathbf{s}^{(1)}$ and $\mathbf{x}^{(0)}$ can be arbitrary vectors and here we set $\mathbf{s}^{(1)} = \mathbf{x}^{(0)} = \mathbf{0}$. Unless otherwise specified, we empirically set $\mu = 0.04$, $\alpha = 1$, $\eta = 10$, $\beta^{(1)} = \frac{1}{3}$, $\varrho = 0.5$, and $T_{\max} = 20$ as default values. After $T_{\max}$ iterations, final estimated vector $\mathbf{x}^{(T_{\max})}$ is obtained, yielding the final detection result. To summarize, the proposed QPG detection algorithm for quantized uplink massive MIMO systems is outlined in Algorithm 1.

TABLE I
COMPUTATIONAL COMPLEXITY COMPARISONS OF THE QUANTIZED DETECTION ALGORITHMS

Algorithm	**Computational complexity**	
	1-bit quantization	Few-bit quantization
Traditional MMSE [12]	$0.5K^3 + K^2N + KN + K^2 + K$	
BMMSE [13]	$0.5N^3 + 2N^2K + N^2 + NK$	
One-stage nML [16]	$NK + (2N+1)KT_{\max}$	–
Two-stage nML [16]	$NK + (2N+1)KT_{\max} + \lvert\mathcal{N}_s\rvert N(K+2)$	–
GD [29]	$NK + (2NK + 3N + K)T_{\max}$	$(2NK + 6N + K)T_{\max}$
Proposed QPG-C	$NK + (4NK + 7N + 12K)T_{\max}$	$(4NK + 10N + 12K)T_{\max}$
Proposed QPG-G	$NK + (4NK + 7N + 5K + 14\lvert\mathcal{X}\rvert K)T_{\max}$	$(4NK + 10N + 5K + 14\lvert\mathcal{X}\rvert K)T_{\max}$

Recent work [23] also employs a Newton-based method with a penalty function for 1-bit massive MIMO orthogonal frequency division multiplexing (OFDM) systems. In contrast, our approach incorporates constellation-aware penalty functions and further establishes theoretical convergence guarantees in Section IV, making it better suited for general few-bit systems with known symbol priors.

D. Computational Complexity Analysis

Here, the computational complexity is evaluated in terms of the required number of real multiplications [40]. For example, the complexity required to invert a matrix of dimension $K \times K$ is quantified as $0.5K^3$ real multiplications [41].

In particular, the computational complexity of QPG can be broken down into several main components. First, the multiplication between $\operatorname{diag}\{\operatorname{sign}(\mathbf{y})\}$ and $\mathbf{H}$ incurs a complexity of NK. At each iteration, computing gradient $\nabla F(\mathbf{s}^{(t)})$ requires $2NK + 3N$ for 1-bit case and $2NK + 6N$ for few-bit case. Updating $\mathbf{z}^{(t)}$ has a computational cost of K. For the proximal operation, updating $\mathbf{x}^{(t)}$ involves computing both the gradient and Hessian matrix of the proximal objective function for each $x_i^{(t)}$, resulting in complexities of $6K$ and $(8|\mathcal{X}| + 3)K$ for the Cosine and Gaussian penalty functions, respectively. After that, calculating $\mathbf{v}^{(t)}$ involves a scalar-vector multiplication and requires K operations. The evaluations of $F(\mathbf{x}^{(t)})$ and $F(\mathbf{v}^{(t)})$ need complexities of $2(K+2)N + 2K$ and $2(K+2)N + 3|\mathcal{X}|K$ for the Cosine and Gaussian cases, respectively. In the following iterations, the cost of updating $\mathbf{s}^{(t+1)}$ and $\beta^{(t+1)}$ is negligible. To summarize, the overall complexity of the QPG detection algorithm is illustrated in Table I.

Throughout the context, $|\mathcal{X}|$ represents the size of constellation set, while $|\mathcal{N}_s|$ refers to the size of the search space for nML. $T_{\max}$ denotes the numbers of iterations. The computational complexity of the considered detection schemes is analyzed separately for different quantization scenarios. Specifically, we refer to the QPG algorithm with the Cosine and Gaussian penalty functions as QPG-C and QPG-G, respectively. As summarized in Table I, both QPG-C and QPG-G exhibit a complexity of only $O(NK)$, which is significantly lower than that of both traditional MMSE and BMMSE methods. Although the complexity of QPG-C is approximately twice that of the GD algorithm, it will be shown in Section V that QPG-C converges substantially faster than GD, resulting in a comparable overall complexity. Furthermore, while QPG-G introduces additional computational overhead, it achieves considerable performance improvements. This trade-off provides a flexible and practical solution for quantized massive MIMO detection.

IV. DETECTION PERFORMANCE ANALYSIS

This section provides a comprehensive analysis of the proposed QPG algorithm. We first examine the Lipschitz continuity of $f(\mathbf{x})$. Based on this, we then prove the monotone descent property to ensure convergence to a stationary point. Finally, assuming convergence to a local minimum, we establish the linear convergence rate of QPG.

A. Lipschitz Continuity

To ensure the stability and convergence of the QPG algorithm, we first establish the Lipschitz continuity of the objective function $f(\mathbf{x})$ defined in (11).

Theorem 2. *The gradient of the objective function* $f(\mathbf{x})$ *in* (11) *is Lipschitz continuous, where the Lipschitz constant* L_f *is given by*

$$L_f = \frac{1}{2}\|\mathbf{H}\|_2^2, \tag{37}$$

where $\|\mathbf{H}\|_2$ *is the* ℓ_2*-norm of the matrix* $\mathbf{H}$.

The proof of Theorem 2 is provided in Appendix A.

This Lipschitz constant plays a fundamental role in determining the step-size constraint $\mu \leq 1/L_f$ and directly impacts the convergence behavior of the QPG algorithm. Specifically, $\|\mathbf{H}\|_2^2$ reflects the system dimensionality, such as the number of users and antennas in massive MIMO systems. Larger system sizes typically lead to higher spectral norms, and hence larger L_f, which reduces the allowable step-size and slows down convergence. This indicates that more careful step-size selection is required in large-scale systems to maintain stable and efficient optimization.

B. Monotone Descent Property

Based on the Lipschitz continuity of $\nabla f(\mathbf{x})$, we now establish the monotone descent property of the proposed QPG algorithm and analyze the behavior of its stationary points.

Theorem 3. *Given the objective function $F(\mathbf{x})$ in* (15) *and a step-size $\mu \leq \frac{1}{L_f}$, the sequence $\{\mathbf{x}^{(t)}\}$ generated by the QPG algorithm satisfies the following descent property*

$$F(\mathbf{x}^{(t+1)}) \leq F(\mathbf{x}^{(t)}), \tag{38}$$

and eventually converges to the stationary point, where

$$\lim_{t\to\infty} \nabla F(\mathbf{x}^{(t)}) = 0. \tag{39}$$

Proof. The key step in analyzing the descent property of QPG lies in the proximal operation in (21), which can be rewritten as

$$\begin{aligned}\mathbf{x}^{(t)} &= \operatorname{prox}_{\mu g}\left\{\mathbf{s}^{(t)} - \mu\nabla f(\mathbf{s}^{(t)})\right\}\\ &= \arg\min_{\mathbf{x}}\left\{g(\mathbf{x}) + \frac{1}{2\mu}\left\|\mathbf{x} - \mathbf{s}^{(t)} + \mu\nabla f(\mathbf{s}^{(t)})\right\|^2\right\}\\ &= \arg\min_{\mathbf{x}}\left\{g(\mathbf{x}) + \langle\nabla f(\mathbf{s}^{(t)}), \mathbf{x} - \mathbf{s}^{(t)}\rangle + \frac{1}{2\mu}\left\|\mathbf{x} - \mathbf{s}^{(t)}\right\|^2\right\},\end{aligned} \tag{40}$$

where $\langle\cdot,\cdot\rangle$ denotes the vector inner product, and $\left\|\nabla f(\mathbf{s}^{(t)})\right\|$ is omitted in the optimization progress for $\mathbf{x}$. Notably, when $\mathbf{x} = \mathbf{s}^{(t)}$, the above function equals to $g(\mathbf{s}^{(t)})$. Since $\mathbf{x}^{(t)}$ is the optimal solution to this problem, the following inequality holds

$$g(\mathbf{x}^{(t)}) + \langle\nabla f(\mathbf{s}^{(t)}), \mathbf{x}^{(t)} - \mathbf{s}^{(t)}\rangle + \frac{1}{2\mu}\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\|^2 \leq g(\mathbf{s}^{(t)}). \tag{41}$$

Furthermore, using the Lipschitz continuity of $\nabla f(\mathbf{x})$ established in Theorem 2, we obtain

$$\begin{aligned}F(\mathbf{x}^{(t)}) &= f(\mathbf{x}^{(t)}) + g(\mathbf{x}^{(t)})\\ &\leq f(\mathbf{s}^{(t)}) + \langle\nabla f(\mathbf{s}^{(t)}), \mathbf{x}^{(t)} - \mathbf{s}^{(t)}\rangle + \frac{L_f}{2}\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\|^2\\ &\quad + g(\mathbf{s}^{(t)}) - \langle\nabla f(\mathbf{s}^{(t)}), \mathbf{x}^{(t)} - \mathbf{s}^{(t)}\rangle - \frac{1}{2\mu}\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\|^2\\ &= F(\mathbf{s}^{(t)}) - \left(\frac{1}{2\mu} - \frac{L_f}{2}\right)\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\|^2.\end{aligned} \tag{42}$$

Since the step-size satisfies $\mu \leq \frac{1}{L_f}$, it follows that $F(\mathbf{x}^{(t)}) \leq F(\mathbf{s}^{(t)})$. Additionally, the update rule of QPG, from (23) to (26), ensures that $F(\mathbf{s}^{(t+1)}) \leq F(\mathbf{x}^{(t)})$. Thus, for all iteration index t, we conclude

$$F(\mathbf{s}^{(t+1)}) \leq F(\mathbf{x}^{(t)}) \leq F(\mathbf{s}^{(t)}) \leq F(\mathbf{x}^{(t-1)}). \tag{43}$$

This establishes the monotone descent property of the sequence $\{\mathbf{x}^{(t)}\}$ generated by QPG.

Combing with the fact that $F(\mathbf{x}^{(t)}), F(\mathbf{s}^{(t)}) \geq \inf F > -\infty$ for all t, we conclude that both sequences $\{F(\mathbf{x}^{(t)})\}$ and $\{F(\mathbf{s}^{(t)})\}$ converge to the same limit, denoted as F^*, namely,

$$\lim_{t\to\infty} F(\mathbf{x}^{(t)}) = \lim_{t\to\infty} F(\mathbf{s}^{(t)}) = F^*. \tag{44}$$

Now combining (42) and (43) yields

$$\begin{aligned}\left(\frac{1}{2\mu} - \frac{L_f}{2}\right)\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\|^2 &\leq F(\mathbf{s}^{(t)}) - F(\mathbf{x}^{(t)})\\ &\leq F(\mathbf{s}^{(t)}) - F(\mathbf{s}^{(t+1)}).\end{aligned} \tag{45}$$

Summing over $t = 1, 2, \ldots, \infty$, we obtain

$$\left(\frac{1}{2\mu} - \frac{L_f}{2}\right)\sum_{t=1}^{\infty}\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\|^2 \leq F(\mathbf{s}^{(1)}) - F^* < \infty. \tag{46}$$

Since $\mu \leq \frac{1}{L_f}$, it follows that

$$\lim_{t\to\infty}\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\| = 0. \tag{47}$$

This implies that the QPG algorithm converges to a stationary point.

Finally, by the optimality condition of the proximal operation in (40), which is shown to be convex in Theorem 1, we have

$$\nabla g(\mathbf{x}^{(t)}) + \nabla f(\mathbf{s}^{(t)}) + \frac{1}{\mu}(\mathbf{x}^{(t)} - \mathbf{s}^{(t)}) = 0. \tag{48}$$

Thus, we can obtain the gradient of $F(\mathbf{x}^{(t)})$ as

$$\begin{aligned}\nabla F(\mathbf{x}^{(t)}) &= \nabla f(\mathbf{x}^{(t)}) + \nabla g(\mathbf{x}^{(t)})\\ &= \nabla f(\mathbf{x}^{(t)}) - \nabla f(\mathbf{s}^{(t)}) - \frac{1}{\mu}(\mathbf{x}^{(t)} - \mathbf{s}^{(t)}),\end{aligned} \tag{49}$$

whose norm can be further upper bounded by

$$\begin{aligned}\|\nabla F(\mathbf{x}^{(t)})\| &= \left\|\nabla f(\mathbf{x}^{(t)}) - \nabla f(\mathbf{s}^{(t)}) - \frac{1}{\mu}(\mathbf{x}^{(t)} - \mathbf{s}^{(t)})\right\|\\ &\leq \left\|\nabla f(\mathbf{x}^{(t)}) - \nabla f(\mathbf{s}^{(t)})\right\| + \frac{1}{\mu}\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\|\\ &\leq \left(L_f + \frac{1}{\mu}\right)\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\|.\end{aligned} \tag{50}$$

Taking the limit as $t \to \infty$, and using the fact that $\lim_{t\to\infty}\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\| = 0$, we obtain $\lim_{t\to\infty}\nabla F(\mathbf{x}^{(t)}) = 0$, which completes the proof. □

Clearly, Theorem 3 ensures that sequence $\{\mathbf{x}^{(t)}\}$ converges to a stationary point where the gradient vanishes. Given the monotone descent property, the stationary point may be a saddle point or a local minimum. In the next subsection, we focus on the case where QPG converges to a local minimum and demonstrate the resulting linear convergence behavior.

C. Linear Convergence Rate of QPG

Building on the monotone descent property, we now formalize the linear convergence rate of the QPG algorithm, under the condition that stationary point is a local minimum.

Theorem 4. *Consider the objective function $F(\mathbf{x})$ in* (15) *and a step-size $\mu \leq \frac{1}{L_f}$, then the QPG algorithm exhibits linear convergence rate, characterized by*

$$\varepsilon^{(t)} \leq \left(\frac{dC^2}{1 + dC^2}\right)^t \varepsilon^{(0)}, \tag{51}$$

with constants

$$d = \frac{2(\mu L_f + 1)^2}{\mu(1-\mu L_f)}, \quad C = \frac{1}{\sqrt{2\lambda_{\min}\left(\nabla^2 F(\mathbf{x}^*)\right)}}, \tag{52}$$

where $\varepsilon^{(t)} = F(\mathbf{x}^{(t)}) - F(\mathbf{x}^*)$ represents the residual, and $\mathbf{x}^*$ denotes the local minimum of $F(\mathbf{x})$.

Proof. To analyze the convergence rate, we first consider the local minimum $\mathbf{x}^*$ of $F(\mathbf{x})$, which satisfies

$$\nabla F(\mathbf{x}^*) = \mathbf{0}, \quad \nabla^2 F(\mathbf{x}^*) \succeq \mathbf{0}. \tag{53}$$

Using the first-order Taylor expansion around $\mathbf{x}^*$, we expend the gradient of $F(\mathbf{x}^{(t)})$ as follows

$$\nabla F(\mathbf{x}^{(t)}) \approx \nabla F(\mathbf{x}^*) + \nabla^2 F(\mathbf{x}^*)(\mathbf{x}^{(t)} - \mathbf{x}^*). \tag{54}$$

Taking the norm on both sides, we obtain the following bound

$$\begin{aligned}\|\nabla F(\mathbf{x}^{(t)})\| &= \|\nabla^2 F(\mathbf{x}^*)(\mathbf{x}^{(t)} - \mathbf{x}^*)\| \\ &\geq \sigma_{\min}\left(\nabla^2 F(\mathbf{x}^*)\right)\|\mathbf{x}^{(t)} - \mathbf{x}^*\|, \end{aligned} \tag{55}$$

where $\sigma_{\min}\left(\nabla^2 F(\mathbf{x}^*)\right) \geq 0$ denotes the smallest singular value of the Hessian matrix. Similarly, expending $F(\mathbf{x}^{(t)})$ using the second-order Taylor expansion around $\mathbf{x}^*$, we have

$$\begin{aligned}F(\mathbf{x}^{(t)}) &\approx F(\mathbf{x}^*) + \nabla F(\mathbf{x}^*)^T(\mathbf{x}^{(t)} - \mathbf{x}^*) \\ &\quad + \frac{1}{2}(\mathbf{x}^{(t)} - \mathbf{x}^*)^T \nabla^2 F(\mathbf{x}^*)(\mathbf{x}^{(t)} - \mathbf{x}^*). \end{aligned} \tag{56}$$

Thus, the residual $F(\mathbf{x}^{(t)}) - F(\mathbf{x}^*)$ can be lower bounded as

$$\begin{aligned}F(\mathbf{x}^{(t)}) - F(\mathbf{x}^*) &= \frac{1}{2}(\mathbf{x}^{(t)} - \mathbf{x}^*)^T \nabla^2 F(\mathbf{x}^*)(\mathbf{x}^{(t)} - \mathbf{x}^*) \\ &\geq \frac{1}{2}\lambda_{\min}\left(\nabla^2 F(\mathbf{x}^*)\right)\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2, \end{aligned} \tag{57}$$

where $\lambda_{\min}\left(\nabla^2 F(\mathbf{x}^*)\right) \geq 0$ refers to the smallest eigenvalue of $\nabla^2 F(\mathbf{x}^*)$. Given that $\nabla^2 F(\mathbf{x}^*)$ is symmetric and positive semi-definite, its eigenvalues coincide with its singular values [42]. Thus, we have $\lambda_{\min}\left(\nabla^2 F(\mathbf{x}^*)\right) = \sigma_{\min}\left(\nabla^2 F(\mathbf{x}^*)\right)$.

To analyze the convergence rate, we introduce the desingularizing function $\varphi(x) = \frac{C}{\phi}x^{\phi}$, where $\phi = \frac{1}{2}$ and $C = 1/\sqrt{2\lambda_{\min}\left(\nabla^2 F(\mathbf{x}^*)\right)}$. Utilizing (55) and (57), we derive the following key inequality

$$\begin{aligned}&\varphi'\left(F(\mathbf{x}^{(t)}) - F(\mathbf{x}^*)\right)\|\nabla F(\mathbf{x}^{(t)})\| \\ &= C\left(F(\mathbf{x}^{(t)}) - F(\mathbf{x}^*)\right)^{\phi-1}\|\nabla F(\mathbf{x}^{(t)})\| \\ &\geq \frac{C}{2^{\phi-1}}\lambda_{\min}^{\phi}\left(\nabla^2 F(\mathbf{x}^*)\right)\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^{2\phi-1} \\ &\geq 1. \end{aligned} \tag{58}$$

Defining the residual as $\varepsilon^{(t)} = F(\mathbf{x}^{(t)}) - F(\mathbf{x}^*)$, and squaring both sides of (58), we obtain

$$\begin{aligned}1 &\leq C^2(\varepsilon^{(t)})^{-1}\|\nabla F(\mathbf{x}^{(t)})\|^2 \\ &\overset{(a)}{\leq} C^2(\varepsilon^{(t)})^{-1}\left(L_f + \frac{1}{\mu}\right)^2\|\mathbf{x}^{(t)} - \mathbf{s}^{(t)}\|^2 \\ &\overset{(b)}{\leq} C^2(\varepsilon^{(t)})^{-1}\left(L_f + \frac{1}{\mu}\right)^2 \frac{1}{\left(\frac{1}{2\mu} - \frac{L_f}{2}\right)}\left(F(\mathbf{x}^{(t-1)}) - F(\mathbf{x}^{(t)})\right) \\ &= \frac{2(\mu L_f + 1)^2}{\mu(1-\mu L_f)}C^2(\varepsilon^{(t)})^{-1}(\varepsilon^{(t-1)} - \varepsilon^{(t)}) \\ &= dC^2\left(\frac{\varepsilon^{(t-1)}}{\varepsilon^{(t)}} - 1\right). \end{aligned} \tag{59}$$

Here, step (a) follows from (50), while step (b) is derived from (43) and (45), with the constant $d = \frac{2(\mu L_f + 1)^2}{\mu(1-\mu L_f)}$. Rearranging the inequality in (59), we obtain

$$\varepsilon^{(t)} \leq \frac{dC^2}{1 + dC^2} \cdot \varepsilon^{(t-1)}, \tag{60}$$

which confirms the linear convergence rate of QPG, thus completing the proof. □

TABLE II
Optimal Quantization for a Standard Gaussian Input [34]

b	**Quantization Alphabet $\mathcal{A}$**	**Quantization Thresholds $\mathcal{C}$**
1	$\{-0.798, 0.798\}$	$\{-\infty, 0, \infty\}$
2	$\{-1.510, -0.453, 0.453, 1.510\}$	$\{-\infty, -0.982, 0, 0.982, \infty\}$
3	$\{-2.152, -1.344, -0.756, -0.242,$ $0.242, 0.756, 1.344, 2.152\}$	$\{-\infty, -1.748, -1.050, -0.500,$ $0, 0.500, 1.050, 1.748, \infty\}$

The above Theorem demonstrates that the proposed QPG detection algorithm linearly converges to a local minimum, which guarantees its superior performance within only a few iterations. To further characterize the trade-off between detection performance and computational complexity, we provide the following Corollary.

Corollary 1. *Given a target error tolerance δ, i.e., $F(\mathbf{x}^{(t)}) - F(\mathbf{x}^*) \leq \delta$, the required number of iterations t for QPG must satisfy*

$$t \geq \frac{\ln\delta - \ln\varepsilon^{(0)}}{\ln dC^2 - \ln(1 + dC^2)}. \tag{61}$$

The proof of Corollary 1 is provided in Appendix B.

This Corollary provides an explicit guideline for determining the number of iterations to meet a target error tolerance δ. Specifically, it shows that only $O(\ln\delta)$ iterations are required, highlighting the efficiency of the QPG algorithm for practical implementations.

V. Simulation Results

In this section, we present a comprehensive simulation study to evaluate the performance of the proposed QPG detection algorithm in uplink quantized massive MIMO systems. Our investigation focuses on the algorithms' convergence behavior, bit error rate (BER) performance, and computational complexity under a variety of realistic channel conditions, including Rayleigh, Rician, correlated fading, and channel estimation error. All simulations employ the non-uniform quantizers introduced in [34], with quantization alphabets $\mathcal{A}$ and thresholds $\mathcal{C}$ for different bit resolutions summarized in Table II. These values are scaled by σ_r defined in (6) to match the dynamic range of the received signal.

Fig. 3 illustrates the BER performance of QPG algorithm with the Cosine penalty function (denoted as QPG-C) under

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2025.3646813

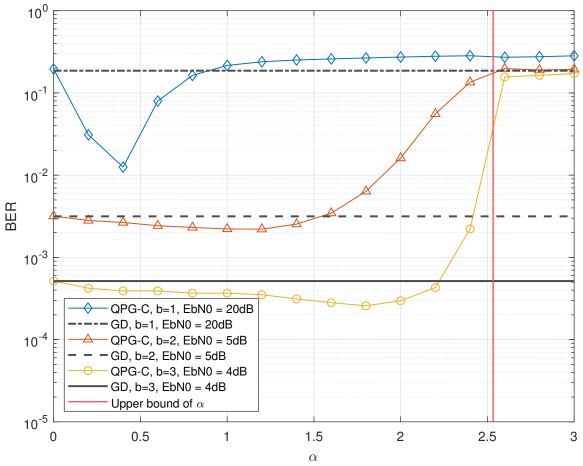


Fig. 3. The impact of penalty parameter α on the performance of the QPG-C detector in the uncoded 128×16 MIMO system with 16-QAM modulation.

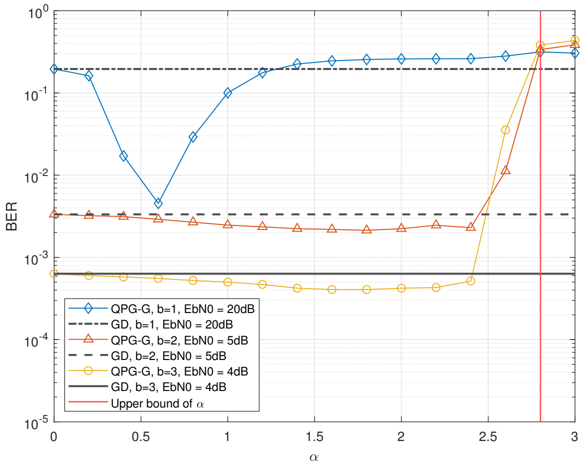


Fig. 4. The impact of penalty parameter α on the performance of the QPG-G detector in the uncoded 128×16 MIMO system with 16-QAM modulation.

varying penalty parameters α in an uncoded 128×16 MIMO system using 1-3 bits ADCs. The channel elements are assumed to be i.i.d. with distribution $\mathcal{CN}(0,1)$. Simulations are conducted under three average per-bit SNR scenarios: 20 dB for $b=1$, 5 dB for $b=2$, and 4 dB for $b=3$. The QPG-C algorithm is executed with fixed parameters $\mu = 0.04$ and $T_{\text{max}} = 20$. For reference, the BER performance of the GD-based detection algorithm in [29] is also shown, using the same step-size and number of iterations as QPG-C. Notably, for small values of α (i.e., 0.2-1), QPG-C consistently outperforms the GD method, highlighting significant nonlinear performance gains. However, as α increases (i.e., 2.2-3), the performance of QPG-C begins to degrade. This degradation occurs because larger α values introduce greater non-convexity into the detection problem formulated in (15), effectively rendering it NP-hard again. The figure also marks the upper bound on α derived in Theorem 1. While this theoretical bound is not tight, since the best performance is observed at values well below it. Nevertheless, it still provides a useful sufficient condition for stable optimization across various SNR and quantization levels.

Fig. 4 further extends the analysis to the QPG algorithm employing a Gaussian penalty function (denoted as QPG-G), evaluated under the same system setup and simulation conditions as in Fig. 3, with fixed $\eta = 10$. Similar to QPG-C, the QPG-G detector achieves noticeable nonlinear gains for small α values (i.e., 0.2-1.2), with performance degrading when α increases further (i.e., 2.6-3). Compared to QPG-C, however, QPG-G demonstrates a wider effective range for α, indicating better robustness and adaptability to different settings. Also, the corresponding upper bound on α is also shown, serving as a sufficient condition to ensure reliable performance improvements across diverse SNR and quantization configurations.

Fig. 5 illustrates the sensitivity of the QPG-G detector's performance to the shape parameter η in an uncoded 128×16 MIMO system employing 16-QAM modulation. The simulation settings are consistent with those used in Fig. 4, with the penalty parameter fixed at $\alpha = 0.5$. As η increases, the penalty function becomes sharper and more peaked, enabling the QPG-G detector to achieve greater performance gains over the GD baseline, particularly in the low η range (i.e., 0-5). However, when η becomes excessively large (e.g., beyond 50), the detector's performance begins to degrade. This degradation is attributed to the excessive non-convexity introduced into the detection problem by large η values, effectively turning the formulation in (15) back into a hard discrete optimization problem.

Fig. 6 compares the detection performance of the proposed QPG algorithm with several existing quantized detection schemes in an uncoded 128×16 MIMO system using 4-QAM modulation under 1-bit quantization. As baselines, we include the traditional MMSE detector for unquantized systems [12] (denoted as traMMSE), the BMMSE algorithm [13], the GD method [29], and both the one-stage and two-stage nML detectors from [16]. Step sizes are set to 0.01 for nML and 0.04 for GD. For the two-stage nML method, the constant c is chosen as 1.3. For the proposed methods, QPG-C uses parameters $\alpha = 0.5$ and $\mu = 0.04$, while QPG-G employs $\alpha = 2$, $\mu = 0.04$, and $\eta = 10$. All iterative detectors are run for a fixed number of iterations, $T_{\text{max}} = 5$. Due to the severe nonlinear distortion introduced by 1-bit quantization, traMMSE, BMMSE, one-stage nML, and GD all exhibit high error floors. The two-stage nML method alleviates this issue via a refined search near the initial estimate, achieving better accuracy. However, this performance gain comes from the computational complexity, which increases exponentially with the search space, rendering it impractical for massive MIMO systems. In contrast, the proposed QPG methods strike a better performance and complexity trade-off. Specifically, QPG-C achieves substantial gains over GD with similar complexity. For example, at $\text{Eb}/\text{N}_0 = 20$ dB, QPG-C attains a BER of 7.72×10^{-6}, significantly outperforming GD, which reaches only 2.49×10^{-5} under the same conditions. Furthermore, QPG-G, which incorporates a Gaussian penalty

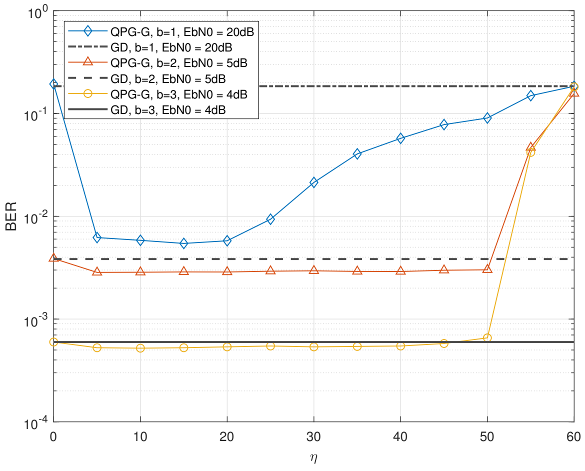


Fig. 5. The impact of shape parameter η on the performance of the QPG-G detector in the uncoded 128×16 MIMO system with 16-QAM modulation.

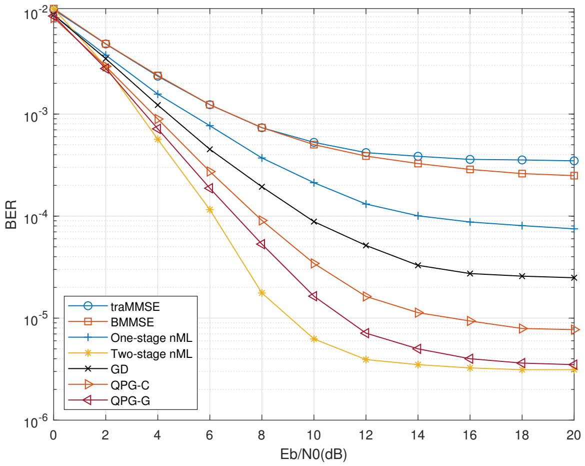


Fig. 6. BER performance comparison of different methods for the uncoded 128×16 MIMO system with 4-QAM modulation and 1-bit ADCs.

function, achieves further improvements in BER and closely approaches the performance of the two-stage nML detector while maintaining significantly lower computational overhead.

Fig. 7 compares the BER performance of various detectors in an uncoded 256×32 MIMO system using 16-QAM modulation under 2-bit and 3-bit quantization. In addition to standard Rayleigh fading, we also evaluate the robustness of the proposed QPG algorithm under Rician fading channels. Following the Rician channel model in [43], the channel matrix is constructed as

$$\widehat{\mathbf{H}} = \sqrt{\frac{\mathcal{K}}{\mathcal{K}+1}}\mathbf{H}_{\text{LOS}} + \sqrt{\frac{1}{\mathcal{K}+1}}\mathbf{H}_{\text{NLOS}}, \tag{62}$$

where $\mathcal{K}$ controls the power ratio between the line-of-sight (LoS) and non-line-of-sight (NLoS) components. The LoS component is defined as $\mathbf{H}_{\text{LOS}_{n,k}} = e^{-j(n-1)\pi \sin(\theta_k)}$, where θ_k denotes the angle of arrival (AoA) for k-th user, uniformly distributed over $[0, 2\pi)$. The NLoS component $\mathbf{H}_{\text{NLOS}_{n,k}}$ is modeled as i.i.d. $\mathcal{CN}(0,1)$ distribution. Consistent with [44], we assume a common Rician factor of $\mathcal{K} = 5$ for all users, simulating a strongly LoS-dominant environment. The number of iterations is set to 20, and other simulation parameters follow those in Fig. 6. While traMMSE and BMMSE benefit from higher quantization levels in terms of convergence, their BER performance remains unsatisfactory under both Rayleigh and Rician fading. In contrast, QPG-C outperforms all other quantized detection baselines, achieving a substantially lower error floor. Remarkably, QPG-G further improves upon QPG-C by leveraging an enhanced penalty function, delivering the best BER performance across all tested configurations. With only second-order computational complexity, QPG-G establishes itself as a state-of-the-art solution for detection in quantized massive MIMO systems.

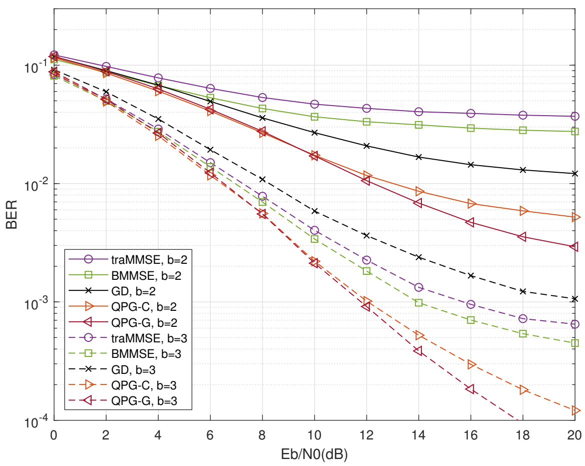


Fig. 7. BER performance comparison of different methods for the uncoded 256×32 MIMO system with 16-QAM modulation and few-bit ADCs over Rician channels.

Fig. 8 evaluates the performance of the proposed QPG algorithms under correlated channels in a quantized massive MIMO system with 2-bit quantization and 16-QAM modulation. Following the channel correlation model in [45], the correlated channel matrix is constructed as

$$\widehat{\mathbf{H}} = \mathbf{R}_{\text{cor}}^{\frac{1}{2}}\mathbf{H}\mathbf{T}_{\text{cor}}^{\frac{1}{2}}, \tag{63}$$

where $\mathbf{R}_{\text{cor}} \in \mathbb{C}^{N\times N}$ and $\mathbf{T}_{\text{cor}} \in \mathbb{C}^{K\times K}$ denote the receive and transmit correlation matrices, respectively. The degree of spatial correlation is controlled by the normalized correlation coefficient $\psi \in [0,1]$, where $\psi = 0$ corresponds to an uncorrelated scenario and $\psi = 1$ indicates full correlation. In this experiment, we set $\psi = 0.1$ to emulate a highly correlated channel environment. We set $\mu = 0.01$, $\alpha = 0.8$, and $\eta = 10$. Due to strong channel correlation, the GD method exhibits slow convergence and poor BER performance, even with 20 iterations. In contrast, both QPG-C and QPG-G demonstrate clear decoding gains over GD. Meanwhile, as expected, QPG-G with an additional computational complexity converges faster and achieves lower BER than QPG-C under the same number of iterations.

Conversely, Fig. 9 investigates the BER performance of the

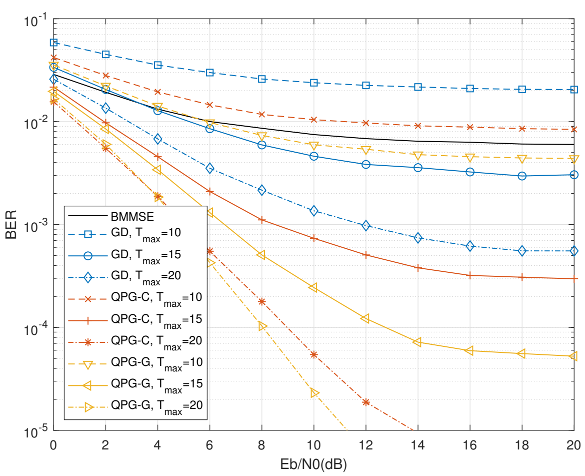


Fig. 8. BER performance comparison of different methods for the uncoded 256×16 MIMO system with 16-QAM modulation and 2-bit ADCs over correlated channels.

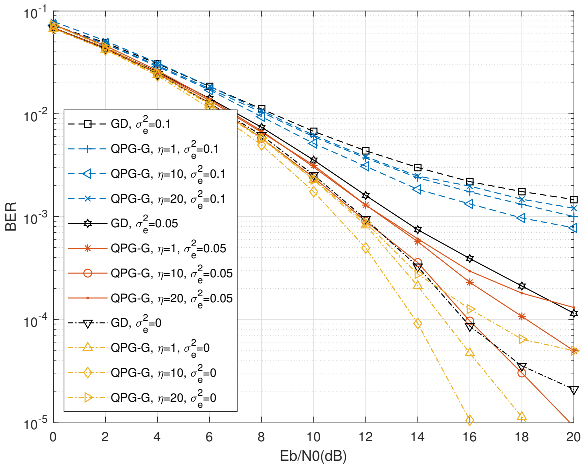


Fig. 9. BER performance comparison of different methods for the uncoded 256×32 MIMO system with 64-QAM modulation, 3-bit ADCs and imperfect CSI.

proposed QPG-G detection algorithm under imperfect channel state information (CSI) in an uncoded 256×32 massive MIMO system using 64-QAM modulation and 3-bit quantization. To model CSI uncertainty at the receiver, we adopt the standard additive error model

$$\widehat{\mathbf{H}} = \mathbf{H} + \triangle\mathbf{H}, \tag{64}$$

where $\triangle\mathbf{H} \sim \mathcal{CN}(\mathbf{0}, \sigma_e^2\mathbf{I})$ denotes the channel estimation error. Following [46], the error variance is set as $\sigma_e^2 = \frac{K}{n_p E_p}$, where n_p and E_p denote the number and power of pilot symbols, respectively. We simulate three representative scenarios: $\sigma_e^2 = 0$ (perfect CSI), $\sigma_e^2 = 0.05$, and $\sigma_e^2 = 0.1$ (severe estimation error). The QPG-G algorithm is configured with $\mu = 0.02$, $\alpha = 2$, and various values of η. The number of iterations is set to 20 for QPG-G and 30 for GD. Compared to the results of perfect CSI, the BER performance of all the detection schemes under imperfect CSI degrade gradually with the increase of σ_e^2. Nevertheless, the proposed QPG-G algorithm maintains robust performance under imperfect CSI, even in the presence of high-order modulation and coarse quantization. Moreover, by tuning η between 1 and 10, QPG-G benefits from a sharper penalty shape, leading to faster convergence. However, when η becomes too large (e.g., $\eta = 20$), performance degradation is observed due to increased non-convexity, which is in line with the analysis of Fig. 5. In practice, we find that moderate values of η (e.g., 10) strike a good balance between convergence speed and detection accuracy, especially for higher-order constellations, which maintains sufficient penalty sharpness without introducing excessive non-convexity.

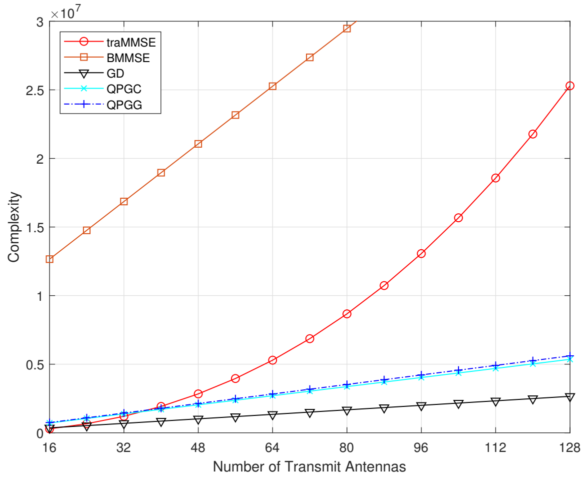


Fig. 10. Comparisons of Computational Complexity for $128 \times N_t$ Quantized Massive MIMO Systems.

Fig. 10 presents a comparative analysis of the computational complexity of various quantized detection algorithms in an uncoded $128 \times N_t$ massive MIMO system employing 16-QAM modulation. Complexity is measured in terms of the number of real-valued multiplications required per detection. For fair comparison, the complexity of GD, QPG-C, and QPG-G detectors is evaluated using a fixed number of iterations, $T_{\text{max}} = 20$. As expected, the computational cost of all schemes increases with the number of transmit antennas N_t. However, unlike the traMMSE and BMMSE algorithms, whose complexity increases rapidly with N_t, the proposed QPG-C and QPG-G algorithms demonstrate significantly slower complexity growth, scaling as $O(N_t N_r)$. In summary, the proposed QPG framework achieves an excellent trade-off between detection performance and computational complexity. QPG-C is particularly attractive for resource-constrained environments and high-order modulation scenarios, offering efficient and practical detection. On the other hand, QPG-G, with slightly higher computational overhead, generally achieves faster convergence and higher accuracy under most practical settings.

VI. Conclusion

This paper has proposed the QPG detection algorithm for massive MIMO systems equipped with low-resolution

ADCs, addressing key limitations of conventional quantized detection approaches. The QPG framework formulates an unconstrained relaxation of the signal detection problem by introducing tailored penalty functions that promote alignment between the estimated symbols and the discrete constellation set. The QPG algorithm is developed to solve this relaxed problem efficiently, achieving competitive detection accuracy with only second-order computational cost. A rigorous theoretical analysis establishes the Lipschitz continuity of the objective function, leading to guaranteed monotonic descent and linear convergence of the proposed algorithm. Extensive simulation results validate the effectiveness of QPG, showing substantial improvements in BER performance and significant reductions in complexity, outperforming existing quantized detection methods under various practical channel conditions.

The proposed QPG algorithm readily extends to the SNR-explicit signal detection problem in (10), and an adaptive step-size scheme may further improve performance. The framework is also structurally compatible with MIMO-OFDM systems by applying frequency-domain transforms, as in [23], and this extension will be explored in future work. Moreover, delay and Doppler effects are important considerations in practical deployment scenarios, particularly in mobile and wideband communication systems, which presents a promising direction for further research.

Appendix

A. Proofs of Theorem 2

To prove that $\nabla f(\mathbf{x})$ is Lipschitz continuous, we first consider the Hessian matrix $\mathcal{H}_f$ of the objective function $f(\mathbf{x})$ in its general form, which is given by

$$\begin{aligned}\mathcal{H}_f(\mathbf{x}) &= \mathbf{H}^T\left[\frac{\partial\sigma(\mathbf{H}\mathbf{x}-\mathbf{q}^{\mathrm{up}})}{\partial\mathbf{x}}+\frac{\partial\sigma(\mathbf{H}\mathbf{x}-\mathbf{q}^{\mathrm{low}})}{\partial\mathbf{x}}\right]\\ &= \mathbf{H}^T\mathrm{diag}\left\{\sigma(\mathbf{u}^{\mathrm{up}})\odot[1-\sigma(\mathbf{u}^{\mathrm{up}})]\right\}\frac{\partial\mathbf{u}^{\mathrm{up}}}{\partial\mathbf{x}}\\ &\quad+\mathbf{H}^T\mathrm{diag}\left\{\sigma(\mathbf{u}^{\mathrm{low}})\odot\left[1-\sigma(\mathbf{u}^{\mathrm{low}})\right]\right\}\frac{\partial\mathbf{u}^{\mathrm{low}}}{\partial\mathbf{x}}\\ &= \mathbf{H}^T\mathrm{diag}\left\{\sigma(\mathbf{u}^{\mathrm{up}})\odot[1-\sigma(\mathbf{u}^{\mathrm{up}})]\right\}\mathbf{H}\\ &\quad+\mathbf{H}^T\mathrm{diag}\left\{\sigma(\mathbf{u}^{\mathrm{low}})\odot\left[1-\sigma(\mathbf{u}^{\mathrm{low}})\right]\right\}\mathbf{H},\end{aligned}\tag{65}$$

with $\mathbf{u}^{\mathrm{up}}=\mathbf{H}\mathbf{x}-\mathbf{q}^{\mathrm{up}}$ and $\mathbf{u}^{\mathrm{low}}=\mathbf{H}\mathbf{x}-\mathbf{q}^{\mathrm{low}}$. Here, the gradient of the sigmoid function is computed as $\frac{\partial\sigma(\boldsymbol{\nu})}{\partial\boldsymbol{\nu}}=\mathrm{diag}\left\{\sigma(\boldsymbol{\nu})\odot[\mathbf{1}-\sigma(\boldsymbol{\nu})]\right\}$ [29]. This general formulation is adopted because the 1-bit quantization is a special case of the multi-bit setting, and thus analyzing the general form ensures broader applicability. Based on (65), for any non-zero vector $\mathbf{z}\in\mathbb{R}^K$, the quadratic form of the Hessian satisfies

$$\begin{aligned}\mathbf{z}^T\mathcal{H}_f(\mathbf{x})\mathbf{z} &= \mathbf{z}^T\mathbf{H}^T\mathrm{diag}\left\{\sigma(\mathbf{u}^{\mathrm{up}})\odot[1-\sigma(\mathbf{u}^{\mathrm{up}})]\right\}\mathbf{H}\mathbf{z}\\ &\quad+\mathbf{z}^T\mathbf{H}^T\mathrm{diag}\left\{\sigma(\mathbf{u}^{\mathrm{low}})\odot\left[1-\sigma(\mathbf{u}^{\mathrm{low}})\right]\right\}\mathbf{H}\mathbf{z}\\ &= \|\sqrt{\mathrm{diag}\left\{\sigma(\mathbf{u}^{\mathrm{up}})\odot[1-\sigma(\mathbf{u}^{\mathrm{up}})]\right\}}\mathbf{H}\mathbf{z}\|^2\\ &\quad+\|\sqrt{\mathrm{diag}\left\{\sigma(\mathbf{u}^{\mathrm{low}})\odot[1-\sigma(\mathbf{u}^{\mathrm{low}})]\right\}}\mathbf{H}\mathbf{z}\|^2\\ &\le\frac{1}{2}\|\mathbf{H}\mathbf{z}\|^2\\ &\le\frac{1}{2}\|\mathbf{H}\|_2^2\|\mathbf{z}\|^2,\end{aligned}\tag{66}$$

where we used the fact that $\sigma(\nu)\cdot[1-\sigma(\nu)]\in[0,\frac{1}{4}]$, for all $\nu\in\mathbb{R}$ [31]. This result leads to the upper bound of $\mathcal{H}_f$ as

$$\mathcal{H}_f(\mathbf{x})\preceq\frac{1}{2}\|\mathbf{H}\|_2^2\mathbf{I}_K,\tag{67}$$

where $\preceq$ denotes the Loewner partial order, i.e., $\mathcal{H}_f$ is upper bounded by $\frac{1}{2}\|\mathbf{H}\|_2^2\mathbf{I}_K$ in the sense of positive semi-definite matrices. Consequently, the gradient $\nabla f(\mathbf{x})$ is Lipschitz continuous with constant $L_f=\frac{1}{2}\|\mathbf{H}\|_2^2$, completing the proof.

B. Proof of Corollary 1

From the convergence result in Theorem 4, we have

$$\varepsilon^{(t)}=F(\mathbf{x}^{(t)})-F(\mathbf{x}^*)\le\left(\frac{dC^2}{1+dC^2}\right)^t\varepsilon^{(0)}.\tag{68}$$

To ensure the residual is below the desired tolerance, i.e., $\varepsilon^{(t)}\le\delta$, it suffices to find the minimum number of iterations t satisfying

$$\left(\frac{dC^2}{1+dC^2}\right)^t\varepsilon^{(0)}\le\delta.\tag{69}$$

Taking the natural logarithm on both sides yields

$$t\cdot\ln\left(\frac{dC^2}{1+dC^2}\right)+\ln\varepsilon^{(0)}\le\ln\delta,\tag{70}$$

which leads directly to the expression stated in Corollary 1, thus completing the proof.

References

[1] M. A. Albreem, M. Juntti, and S. Shahabuddin, "Massive MIMO Detection Techniques: A Survey," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 4, pp. 3109-3132, Aug. 2019.

[2] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 742-758, Oct. 2014.

[3] X. Lin, et al., "5G New Radio: Unveiling the essentials of the next generation wireless access technology," *IEEE Commun. Standards Mag.*, vol. 3, no. 3, pp. 30-37, Sep. 2019.

[4] J. Tian, Y. Han, S. Jin, J. Zhang, and J. Wang, "Analytical channel modeling: From MIMO to extra large-scale MIMO," *Chin. J. Electron.*, vol. 34, no. 1, pp. 1-15, Jan. 2025.

[5] Z. Qin, G. Y. Li, and H. Ye, "Federated learning and wireless communications," *IEEE Wireless Commun.*, vol. 28, no. 5, pp. 134-140, Oct. 2021.

[6] J. Zhang, L. Dai, X. Li, Y. Liu and L. Hanzo, "On Low-Resolution ADCs in Practical 5G Millimeter-Wave Massive MIMO Systems," in *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 205-211, July. 2018.

[7] J. Liu, Z. Luo and X. Xiong, "Low-Resolution ADCs for Wireless Communication: A Comprehensive Survey," in *IEEE Access*, vol. 7, pp. 91291-91324, 2019.

[8] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 4, pp. 539-550, Apr. 1999.

[9] L. Liu, Y. Ma, and N. Yi, "Hermite expansion model and LMMSE analysis for low-resolution quantized MIMO detection," *IEEE Trans. Signal Process.*, vol. 69, pp. 5313-5328, Sept. 2021.

[10] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput analysis of massive MIMO uplink with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4038-4051, Jun. 2017.

[11] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075-4089, Aug. 2017.

[12] L. V. Nguyen and D. H. N. Nguyen, "Linear receivers for massive MIMO systems with one-bit ADCs," *arXiv preprint*, 2019. [Online]. Available: https://arxiv.org/abs/1907.06664.

[13] A. Mezghani and J. A. Nossek, “Capacity lower bound of MIMO channels with output quantization and correlated noise,” in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp.1-5.

[14] T. Liu, J. Tong, Q. Guo, J. Xi, Y. Yu, and Z. Xiao, “Energy efficiency of massive MIMO systems with low-resolution ADCs and successive interference cancellation,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 3987-4002, 2019.

[15] A. Mezghani, M. Khoufi, and J. A. Nossek, “Maximum likelihood detection for quantized MIMO systems,” in *Proc. Int. ITG Work. Smart Antennas (WSA)*, Feb. 2008, pp. 278-284.

[16] J. Choi, J. Mo, and R. W. Heath, “Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs,” *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2005-2018, May. 2016.

[17] K. Safa, R. Combes, R. de Lacerda and S. Yang, “Data Detection in 1-bit Quantized MIMO Systems,” *IEEE Trans. Commun.*, vol. 72, no. 9, pp. 5396-5410, Sept. 2024.

[18] Y. Jeon, N. Lee, S. Hong, and R. W. Heath, “One-bit sphere decoding for uplink massive MIMO systems with one-bit ADCs,” in *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4509-4521, July. 2018.

[19] S.-N. Hong, S. Kim, and N. Lee, “A weighted minimum distance decoding for uplink multiuser MIMO systems with low-resolution ADCs,” *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 1912-1924, May. 2018.

[20] F. Mousavi and A. Tadaion, “A simple two-stage detector for massive MIMO systems with one-bit ADCs,” in *Proc. 27th Iranian Conf. Electr. Eng. (ICEE)*, May. 2019, pp. 1674-1678.

[21] L. V. Nguyen, A. L. Swindlehurst, and D. H. N. Nguyen, “SVM based channel estimation and data detection for one-bit massive MIMO systems,” *IEEE Trans. Signal Process.*, vol. 69, pp. 2086-2099, Mar. 2021.

[22] G. Yılmaz and A. Ö. Yılmaz, “Pseudo-Random Quantization Based Two-Stage Detection in One-Bit Massive MIMO Systems,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 5, pp. 4397-4410, May 2024.

[23] G. Yılmaz and A. Ö. Yılmaz, “Quasi-Newton FDE in one-bit pseudo-randomly quantized massive MIMO-OFDM systems,” *IEEE Commun. Lett.*, vol. 28, no. 4, pp. 917–921, Apr. 2024.

[24] O. T. Demir and E. Björnson, “ADMM-based one-bit quantized signal detection for massive MIMO systems with hardware impairments,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May. 2020, pp. 9120-9124.

[25] C. Studer and G. Durisi, “Quantized massive MU-MIMO-OFDM uplink,” *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2387-2399, Jun. 2016.

[26] C. -K. Wen, C. -J. Wang, S. Jin, K. -K. Wong, and P. Ting, “Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs,” *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541-2556, May. 2016.

[27] Y. Xiong, N. Wei and Z. Zhang, “A Low-Complexity Iterative GAMP-Based Detection for Massive MIMO with Low-Resolution ADCs,” in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1-6.

[28] D. Ho, “*Channel estimation and data detection methods for 1-bit massive MIMO systems*,” Ph.D. dissertation, University of California, 2022.

[29] L. V. Nguyen, D. H. N. Nguyen and A. L. Swindlehurst, “DNN-based Detectors for Massive MIMO Systems with Low-Resolution ADCs,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1-6.

[30] L. V. Nguyen, A. L. Swindlehurst, and D. H. Nguyen, “Linear and deep neural network-based receivers for massive MIMO systems with one-bit ADCs,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7333-7345, 2021.

[31] A. Sant and B. D. Rao, “Insights Into Maximum Likelihood Detection for 1-bit Massive MIMO Communications,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 11, pp. 16275-16289, Nov. 2024.

[32] L. Liu, S. Xue, Y. Ma, N. Yi, and R. Tafazolli, “On the design of quantization functions for uplink massive MIMO with low-resolution ADCs,” in *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, Apr. 2021, pp. 1-5.

[33] Z. Liu, C. Yu, Y. Wang, and S. Liu, “Graph signal reconstruction from low-resolution multi-bit observations,” *Chin. J. Electron.*, vol. 33, no. 1, pp. 153-160, 2024.

[34] S. Lloyd, “Least squares quantization in PCM,” in *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129-137, Mar. 1982.

[35] S. Shahabuddin, I. Hautala, M. Juntti and C. Studer, “ADMM-Based Infinity-Norm Detection for Massive MIMO: Algorithm and VLSI Architecture,” *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 29, no. 4, pp. 747-759, Apr. 2021.

[36] T. Cui and C. Tellambura, “Polynomial-constrained detection using a penalty function and a differential-equation algorithm for MIMO systems,” *IEEE Signal Process. Lett.*, vol. 13, no. 3, pp. 133-136, Mar. 2006.

[37] B. Yan, Z. Wang, J. Zhang, and Y. Huang, “Joint antenna activation and power allocation for energy-efficient cell-free massive MIMO systems,” *IEEE Wireless Commun. Lett.*, vol. 14, no. 1, pp. 243-247, Jan. 2025.

[38] H. Li and Z. Lin, “Accelerated proximal gradient methods for nonconvex programming,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[40] Q. Chen, Z. Wang, C. Qi, Z. Gao, Y. Huang and D. Niyato, “Decentralized Likelihood Ascent Search-Aided Detection For Distributed Large-Scale MIMO Systems,” *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 4160-4173, May. 2025.

[41] A. Krishnamoorthy and D. Menon, “Matrix inversion using Cholesky decomposition,” in *Proc. Signal Process. Algorithms, Architect., Arrange., Appl. Conf. (SPA)*, Sept. 2013, pp. 70-72.

[42] G. Strang, *Introduction to Linear Algebra*, 5th ed. Wellesley-Cambridge Press, 2016.

[43] T. Liu, J. Tong, Q. Guo, J. Xi, Y. Yu and Z. Xiao, “On the Performance of Massive MIMO Systems With Low-Resolution ADCs and MRC Receivers Over Rician Fading Channels,” in *IEEE Syst. J.*, vol. 15, no. 3, pp. 4514-4524, Sept. 2021.

[44] P. Liu, D. Kong, J. Ding, Y. Zhang, K. Wang and J. Choi, “Channel Estimation Aware Performance Analysis for Massive MIMO With Rician Fading,” in *IEEE Trans. on Commun.*, vol. 69, no. 7, pp. 4373-4386, Jul. 2021.

[45] B. Costa, A. Mussi, and T. Abrao, “MIMO detectors under correlated channels,” *Semin. Cienc. Exatas Tecnol.*, vol. 37, no. 1, pp. 3-12, 2016.

[46] Q. Chen, S. Zhang, S. Xu, and S. Cao, “Efficient MIMO detection with imperfect channel knowledge: A deep learning approach,” in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2019, pp. 1-6.

Qiqiang Chen received the B.S. degree in information engineering from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2023. He is currently pursuing a Ph.D. degree in signal and information processing at Southeast University. His research interests include massive MIMO systems and decentralized signal processing.

Zheng Wang (Senior Member, IEEE) received the B.S. degree in electronic and information engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2009, and the M.S. degree in communications from University of Manchester, Manchester, U.K., in 2010. He received the Ph.D degree in communication engineering from Imperial College London, UK, in 2015.

Since 2021, he has been an Associate Professor in the School of Information and Engineering, Southeast University (SEU), Nanjing, China. From 2015 to 2016, he served as a Research Associate at Imperial College London, UK. From 2016 to 2017, he was a senior engineer with the Radio Access Network R&D division, Huawei Technologies Co. From 2017 to 2020, he was an Associate Professor at the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include massive MIMO systems, machine learning and data analytics over wireless networks, and lattice theory for wireless communications.

Chenhao Qi (Senior Member, IEEE) received the B.S. degree (Hons.) in information engineering from the Chien-Shiung Wu Honored College, Southeast University, China, in 2004, and the Ph.D. degree in signal and information processing from Southeast University in 2010.

From 2008 to 2010, he visited the Department of Electrical Engineering, Columbia University, New York, USA. Since 2010, he has been a Faculty Member with the School of Information Science and Engineering, Southeast University, where he is currently a Professor and the Head of Jiangsu Multimedia Communication and Sensing Technology Research Center. He received Best Paper Awards from IEEE GLOBECOM in 2019, IEEE/CIC ICCC in 2022, and the 11th International Conference on Wireless Communications and Signal Processing (WCSP) in 2019. He has served as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY, and China Communications.

Feng Shu (Senior Member, IEEE) received the B.S. degree from Fuyang Teaching College, Fuyang, China, in 1994, the M.S. degree from Xidian University, Xi’an, China, in 1997, and the Ph.D. degree from Southeast University, Nanjing, China, in 2002. He was a Visiting Postdoctoral Fellow with the University of Texas at Dallas, Richardson, TX, USA and a Visiting Scholar with the Royal Melbourne Institute of Technology, Melbourne VIC, Australia. From 2005 to 2020, he was at Nanjin University of Science and Technology, being promoted to Full Professor in 2013. Since 2020, he has been with the School of Information and Communication Engineering, Hainan University, where he is currently a Professor and a Supervisor of Ph.D. and graduate students. He has authored or co-authored over 400 archival journal papers, including over 150 on IEEE journals and 300 SCI - indexed papers with over 11000 citations. His research interests include machine learning, wireless networks, wireless location, and array signal processing. He was awarded with the Fujian Prize for Natural Sciences in 2024, the Leading-Talent Plan of Hainan Province in 2020, the Fujian Hundred-Talent Plan of Fujian Province in 2018, and the Minjiang Scholar Chair Professor in 2015. He holds one US patent and about 50 Chinese patents. From 2019 to 2024, he is ranked in the 2% Top Scientists by Stanford/Elsevier, and enters the list of 1% Top scientists in 2024. He is editor of IEEE WCL, IEEE Syst J and IEEE Access. He is also guest editor of CJA, J ELECTRON INF TECHN and IET COMMUN.

Yongming Huang (Fellow, IEEE) received the B.S. and M.S. degrees from Nanjing University, Nanjing, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from Southeast University, Nanjing, in 2007.

Since March 2007 he has been a faculty in the School of Information Science and Engineering, Southeast University, China, where he is currently a full professor. He has also been the Director of the Pervasive Communication Research Center, Purple Mountain Laboratories, since 2019. During 2008-2009, Dr. Huang visited the Signal Processing Lab, Royal Institute of Technology (KTH), Stockholm, Sweden. His current research interests include intelligent 5G/6G mobile communications and millimeter wave wireless communications. He has published over 200 peer-reviewed papers, hold over 80 invention patents. He submitted around 20 technical contributions to IEEE standards, and was awarded a certificate of appreciation for outstanding contribution to the development of IEEE standard 802.11aj. He served as an Associate Editor for the IEEE Transactions on Signal Processing and a Guest Editor for the IEEE Journal Selected Areas in Communications. He is currently an Editor-at-Large for the IEEE Open Journal of the Communications Society and an Associate Editor for the IEEE Wireless Communications Letters.