AN EFFICIENT QUADRATIC PROGRAMMING RELAXATION BASED ALGORITHM FOR LARGE-SCALE MIMO DETECTION *

4 authors, including:

Qing-Na Li
Beijing Institute of Technology
37 PUBLICATIONS **94** CITATIONS

Chen Weikun
Beijing Institute of Technology
39 PUBLICATIONS **153** CITATIONS

Ya-Feng Liu
Chinese Academy of Sciences
165 PUBLICATIONS **3,196** CITATIONS

AN EFFICIENT QUADRATIC PROGRAMMING RELAXATION BASED ALGORITHM FOR LARGE-SCALE MIMO DETECTION[*]

PING-FAN ZHAO[†], QING-NA LI[‡], WEI-KUN CHEN[‡], AND YA-FENG LIU[§]

Abstract. Multiple-input multiple-output (MIMO) detection is a fundamental problem in wireless communications and it is strongly NP-hard in general. Massive MIMO has been recognized as a key technology in the fifth generation (5G) and beyond communication networks, which on one hand can significantly improve the communication performance, and on the other hand poses new challenges of solving the corresponding optimization problems due to the large problem size. While various efficient algorithms such as semidefinite relaxation (SDR) based approaches have been proposed for solving the small-scale MIMO detection problem, they are not suitable to solve the large-scale MIMO detection problem due to their high computational complexities. In this paper, we propose an efficient sparse quadratic programming (SQP) relaxation based algorithm for solving the large-scale MIMO detection problem. In particular, we first reformulate the MIMO detection problem as an SQP problem. By dropping the sparse constraint, the resulting relaxation problem shares the same global minimizer with the SQP problem. In sharp contrast to the SDRs for the MIMO detection problem, our relaxation does not contain any (positive semidefinite) matrix variable and the numbers of variables and constraints in our relaxation are significantly less than those in the SDRs, which makes it particularly suitable for the large-scale problem. Then we propose a projected Newton based quadratic penalty method to solve the relaxation problem, which is guaranteed to converge to the vector of transmitted signals under reasonable conditions. By extensive numerical experiments, when applied to solve small-scale problems, the proposed algorithm is demonstrated to be competitive with the state-of-the-art approaches in terms of detection accuracy and solution efficiency; when applied to solve large-scale problems, the proposed algorithm achieves better detection performance than a recently proposed generalized power method.

Key words. MIMO Detection, Projected Newton Method, Quadratic Penalty Method, Semidefinite Relaxation, Sparse Quadratic Programming Relaxation

AMS subject classifications. 90C22, 90C20, 90C27

1. Introduction. Multiple-input multiple-output (MIMO) detection is a fundamental problem in modern communications [1, 34]. The input-output relationship of the MIMO channel is

$$(1.1) \qquad\qquad r = Hx^* + v,$$

where $r \in \mathbb{C}^m$ denotes the vector of received signals, $H \in \mathbb{C}^{m \times n}$ denotes an $m \times n$ complex channel matrix (usually $m \geqslant n$), $x^* \in \mathbb{C}^n$ denotes the vector of transmitted signals, and $v \in \mathbb{C}^m$ denotes an additive white circularly symmetric Gaussian noise. The goal of MIMO detection is to recover the transmitted signals x^* from the received signals r based on the channel information H. We refer to [8, 34] for a review

[†]School of Mathematics and Statistics, Beijing Institute of Technology, Beijing (pfzhao@bit.edu.cn).

[‡]Qing-Na Li is the corresponding author. School of Mathematics and Statistics/Beijing Key Laboratory on MCAACI, Beijing Institute of Technology, Beijing (qnl@bit.edu.cn, chenweikun@bit.edu.cn).

[§]State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing (yafliu@lsec.cc.ac.cn).

1

of different formulations and approaches for MIMO detection and [1] for the latest progress in MIMO detection.

In this paper, we assume that x^* in (1.1) is modulated via the M-Phase-Shift Keying (M-PSK) modulation scheme with $M \geqslant 2$. More exactly, each entry x_j^* of x^* belongs to a finite set:

$$(1.2) \qquad x_j^* \in \mathcal{X} \triangleq \left\{ \exp(\mathrm{i}\theta) \ \middle| \ \theta = \frac{2(k-1)\pi}{M}, \ k = 1, \ldots, M \right\}, \ j = 1, \ldots, n,$$

where i is the imaginary unit. The mathematical formulation for the MIMO detection problem is

$$
\begin{aligned}
\text{(P)} \qquad &\min_{x \in \mathbb{C}^n} \quad F(x) \triangleq \|Hx - r\|_2^2 \\
&\text{s.t.} \quad |x_j|^2 = 1, \ j = 1, \ldots, n, \\
&\qquad \arg(x_j) \in \mathcal{A} \triangleq \left\{ 0, \frac{2\pi}{M}, \ldots, \frac{2(M-1)\pi}{M} \right\}, \ j = 1, \ldots, n,
\end{aligned}
$$

where $\| \cdot \|_2$ denotes the Euclidean norm and $\arg(\cdot)$ denotes the argument of the complex number.

Let

$$(1.3) \qquad\qquad Q = H^\dagger H \text{ and } c = -H^\dagger r,$$

where $(\cdot)^\dagger$ denotes the conjugate transpose. Then problem (P) is equivalent to the following complex quadratic programming problem

$$
\begin{aligned}
\text{(CQP)} \qquad &\min_{x \in \mathbb{C}^n} \quad x^\dagger Q x + 2\mathrm{Re}(c^\dagger x) \\
&\text{s.t.} \quad |x_j|^2 = 1, \ j = 1, \ldots, n, \\
&\qquad \arg(x_j) \in \mathcal{A}, \ j = 1, \ldots, n,
\end{aligned}
$$

where $\mathrm{Re}(\cdot)$ denotes the real part of the complex number.

Various methods to tackle the MIMO detection problem can be summarized into several lines [34, Figure 15], including tree search [7, 26, 32], lattice reduction (LR) [12, 38], and semidefinite relaxation (SDR) [18, 19, 29, 31]. The tree search based methods are the most popular detectors in the era of multi-antenna MIMO systems [34]. Taking the typical tree search based method, the sphere decoder (SD) algorithm [7], as an example, it is regarded as the benchmark for globally solving the MIMO detection problem. However, both the expected and worst-case complexities of the SD algorithm are exponential [9, 30]. The most popular LR algorithm is the Lenstra-Lenstra-Lovász (LLL) algorithm [12], whose worst-case computational complexity can be prohibitively high [10, 35]. Below we mainly review the SDR based approach, which is most related to this work.

The SDR based approach was first proposed for a binary PSK (BPSK) modulated code division multiple access (CDMA) system [29]. Then it was extended to the quadrature PSK (QPSK) scenario [16] and further to the high-order M-PSK scenario [20, 21]. In [22], a quadratic assignment problem formulation was proposed for problem (P), and a near-maximum-likelihood decoding algorithm was designed based on the resulting SDR. Other early SDR based approaches are summarized in [34, Table IX].

SDR based approaches generally perform very well for solving the MIMO detection problem. To understand the reason, various researches have been done and one line of researches is to identify conditions under which the SDRs are *tight* [17, Definition 1]. For the case where $M = 2$, So [27] proposed an SDR of problem (P) and proved its tightness when the following condition

$$(1.4) \qquad \lambda_{\min}(\mathrm{Re}(H^{\dagger}H)) > \|\mathrm{Re}(H^{\dagger}v)\|_{\infty}$$

is satisfied. Here H and v are defined in (1.1), $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a given matrix, and $\| \cdot \|_{\infty}$ denotes the ℓ_{∞}-norm. An open question proposed in [27] is that whether the (conventional) SDR is still tight under condition (1.4) for the case where $M \geqslant 3$. It was negatively answered in [17]. In addition, Lu et al. in [17] proposed an enhanced SDR (see (ERSDR1) further ahead) by adding some valid inequalities and showed that under condition

$$(1.5) \qquad \lambda_{\min}(H^{\dagger}H)\sin\left(\frac{\pi}{M}\right) > \|H^{\dagger}v\|_{\infty},$$

(ERSDR1) is tight. In [15], the relations between different SDRs were further analyzed. In particular, it was proved that (ERSDR1) and the SDR proposed in [22] are equivalent, and as a result, the SDR proposed in [22] is also tight under condition (1.5). Other representative analysis results can be found in [3, 23, 11].

One key advantage of the SDR based approaches, compared to SD and LLL algorithms, is that the SDR admits polynomial-time algorithms. There are well developed solvers for solving the SDR, such as MOSEK [24] and the latest SDPNAL+ [28, 33, 36, 37]. However, the numbers of variables and constraints in the SDRs are much larger than those in problem (CQP), and hence the SDR based approaches cannot be used to solve the *large-scale* MIMO detection problem. On the other hand, it was predicted that the mobile data traffic will grow exponentially in 2017-2022 [5], which calls for higher data rates, larger network capacity, higher spectral efficiency, higher energy efficiency, and better mobility [1]. Massive MIMO is a key and effective technology to meet the above requirements, where the base station (BS) is equipped with tens to hundreds of antennas, in contrast to the current BS equipped with only 4 to 8 antennas. A new challenge coming with the massive MIMO technology is the large problem size in signal processing and optimization. In particular, the MIMO detection problem of our interest in the massive MIMO setup is a *large-scale* strongly NP-hard problem [30]. As far as we know, there are very few works on the large-scale MIMO detection problem. One notable work is [14], which proposes a customized generalized power method (GPM) for solving the large-scale MIMO detection problem. The GPM directly solves problem (P) and at each iteration, the algorithm takes a gradient descent step with an appropriate stepsize and projects the obtained point onto the (discrete) feasible set of problem (P). However, our experiments show that the performance of the GPM heavily depends on the choice of the initial point. Consequently, models and algorithms that can be generalized to the large-scale MIMO detection problem with satisfactory detection performance are still highly in need.

Contributions. The contributions of the paper are twofold. Firstly, we propose a sparse quadratic programming (SQP) formulation for the MIMO detection problem. We prove that, somewhat surprisingly, its relaxation obtained by dropping the sparse constraint is equivalent to the original formulation. Moreover, the relaxation formulation is able to recover the vector of transmitted signals under condition (1.5). Secondly, we present a projected Newton based quadratic penalty (PN-QP) method to

solve the proposed (relaxation) formulation, which is demonstrated to be quite efficient in terms of detection accuracy and solution efficiency. Under reasonable assumptions on the channel matrix and noise, the sequence generated by PN-QP is guaranteed to converge to the vector of transmitted signals. In particular, our extensive numerical results show that (i) compared to SD and MOSEK (for solving (ERSDR1)), PN-QP is more efficient on massive MIMO detection; (ii) compared to GPM, PN-QP achieves significantly better detection performance than a recently proposed generalized power method.

Two key features of our proposed approach are highlighted as follows. Firstly, in sharp contrast to the matrix based SDRs, due to the vector based formulation for the MIMO detection problem, our relaxation is particularly suitable to deal with the large-scale MIMO detection problem. Secondly, by exploring the sparse structure of the optimal solution, the computational cost of PN-QP is significantly reduced. In particular, PN-QP is designed to identify the support set of the optimal solution rather than to find the solution itself, leading to a low computational cost.

The rest of this paper is organized as follows. In section 2, we introduce different formulations for the MIMO detection problem, including the SQP formulation. In section 3, we discuss the relaxation problem and its properties. In section 4, we present the PN-QP method and its convergence result. In section 5, we perform extensive numerical experiments to compare different algorithms for solving the MIMO detection problem. Finally, we conclude the paper in section 6.

We adopt the following standard notations in this paper. Let i denote the imaginary unit (satisfying $i^2 = -1$). For a given complex vector x, we use x_j to denote its j-th entry, and $|x_j|$ to denote the modulus of its j-th entry. Let $\|\cdot\|_2$ denote the ℓ_2-norm for vectors and Frobenius norm for matrices. We use $\|x\|_0$ to denote the number of nonzero entries in vector x. Let $\mathrm{diag}(X)$ denote the vector formed by the diagonal elements in matrix X, and $\mathrm{Diag}(x)$ denote the diagonal matrix with the diagonal entries being vector x. For matrices $X_{11}, \ldots, X_{nn} \in \mathbb{R}^{M \times M}$, we also use $\mathrm{Diag}(X_{11}, \ldots, X_{nn}) \in \mathbb{R}^{nM \times nM}$ to denote the block-diagonal matrix whose (j, j)-th block is X_{jj}. For a complex matrix C, let $\mathrm{Re}(C)$ and $\mathrm{Im}(C)$ denote the real and imaginary parts of C, respectively, and $C^\dagger$ and $C^\top$ denote the conjugate transpose and transpose of C, respectively. $C \succeq 0$ means C is positive semidefinite, and $\mathrm{Tr}(C)$ denotes the trace of C. Define the inner product for $x, v \in \mathbb{C}^n$ as $\langle x, v \rangle = \mathrm{Re}(x^\dagger v)$. For two Hermitian matrices A and B, the inner product is defined similarly as $\langle A, B \rangle = \mathrm{Re}(\mathrm{Tr}(A^\dagger B))$. Let e be a vector of an appropriate length with all elements being one. For a sequence $\{x^k\}$, $x^k \uparrow c$ and $x^k \downarrow c$ mean that x^k tends to increasingly and decreasingly to a certain value c, respectively. We use $\otimes$ to denote the Kronecker product. For $t \in \mathbb{R}^{nM}$, we assume that t has the partition as $t = (\bar{t}_1^\top, \ldots, \bar{t}_n^\top)^\top$, where $\bar{t}_j \in \mathbb{R}^M$ is the j-th block of t. Finally, the k-th entry in block $\bar{t}_j$ is denoted as $(\bar{t}_j)_k$.

2. Different Formulations for MIMO Detection. In this section, we introduce some formulations for the MIMO detection problem and discuss their properties.

Define

$$(2.1) \qquad \mathcal{Y} = \left\{ (\cos\theta_k, \sin\theta_k) \;\middle|\; \theta_k = \frac{2(k-1)\pi}{M}, \; k = 1, \ldots, M \right\}.$$

Then, for each $j = 1, \ldots, n$, it is easy to see that $x_j \in \mathcal{X}$ (defined in (1.2)) if and only if $(\mathrm{Re}(x_j), \mathrm{Im}(x_j)) \in \mathcal{Y}$. The feasible points of $\mathcal{Y}$ for $M = 4$ and $M = 8$ are illustrated in Figure 1.

(a) $M = 4$. (b) $M = 8$.
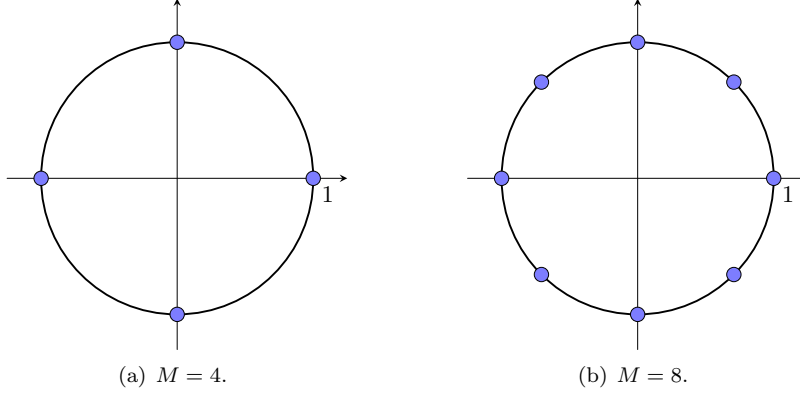
FIG. 1. *An illustration of* $\mathcal{Y}$.

Let

$$(2.2) \quad \widehat{Q} = \begin{bmatrix} \operatorname{Re}(Q) & -\operatorname{Im}(Q) \\ \operatorname{Im}(Q) & \operatorname{Re}(Q) \end{bmatrix} = (\hat{q}_{jk})_{2n \times 2n}, \ \hat{c} = \begin{bmatrix} \operatorname{Re}(c) \\ \operatorname{Im}(c) \end{bmatrix}, \text{ and } y = \begin{bmatrix} \operatorname{Re}(x) \\ \operatorname{Im}(x) \end{bmatrix} \in \mathbb{R}^{2n}.$$

Problem (CQP) can be equivalently written as the following real form:

(RQP)
$$\min_{y \in \mathbb{R}^{2n}} \quad y^{\top} \widehat{Q} y + 2\hat{c}^{\top} y$$
$$\text{s.t.} \quad (y_j, y_{n+j}) \in \mathcal{Y}, \ j = 1, \ldots, n.$$

Let $t = (\bar{t}_1^{\top}, \ldots, \bar{t}_n^{\top})^{\top} \in \mathbb{R}^{nM}$ where $\bar{t}_j \in \mathbb{R}^M$ is the assignment variable corresponding to (y_j, y_{n+j}), i.e.,

$$(\bar{t}_j)_k = \begin{cases} 1, & \text{if } (y_j, y_{n+j}) = (\cos \theta_k, \sin \theta_k); \\ 0, & \text{otherwise.} \end{cases}$$

By the above definition, the constraints in problem (RQP) can be equivalently written as

$$\begin{bmatrix} y_j \\ y_{n+j} \end{bmatrix} = \sum_{k=1}^{M} (\bar{t}_j)_k \begin{bmatrix} \cos \theta_k \\ \sin \theta_k \end{bmatrix}, \ \bar{t}_j \in \{0, 1\}^M, \ \boldsymbol{e}^{\top} \bar{t}_j = 1, \ j = 1, \ldots, n.$$

Then problem (RQP) can be equivalently written as

(2.3)
$$\min_{\substack{y \in \mathbb{R}^{2n}, \\ t \in \mathbb{R}^{nM}}} \quad y^{\top} \widehat{Q} y + 2\hat{c}^{\top} y$$
$$\text{s.t.} \quad y_j = \alpha^{\top} \bar{t}_j, \ j = 1, \ldots, n,$$
$$y_{n+j} = \beta^{\top} \bar{t}_j, \ j = 1, \ldots, n,$$
$$\boldsymbol{e}^{\top} \bar{t}_j = 1, \ j = 1, \ldots, n,$$
$$t \in \{0, 1\}^{nM},$$

where

(2.4)
$$\alpha = \left(\cos 0, \cos \left(\frac{2\pi}{M} \right), \ldots, \cos \left(\frac{2(M-1)\pi}{M} \right) \right)^{\top} \in \mathbb{R}^M \text{ and}$$
$$\beta = \left(\sin 0, \sin \left(\frac{2\pi}{M} \right), \ldots, \sin \left(\frac{2(M-1)\pi}{M} \right) \right)^{\top} \in \mathbb{R}^M.$$

We now eliminate the variables y_j for $j = 1, \ldots, 2n$, based on the constraints in problem (2.3). Let

$$(2.5) \quad A = I \otimes \alpha^\top \in \mathbb{R}^{n \times nM}, \ B = I \otimes \beta^\top \in \mathbb{R}^{n \times nM}, \ \text{and} \ P = \begin{bmatrix} A \\ B \end{bmatrix} \in \mathbb{R}^{2n \times nM}.$$

We obtain the following quadratic assignment problem:

$$
\begin{aligned}
\min_{t \in \mathbb{R}^{nM}} \quad & h(t) \triangleq t^\top G t + 2 w^\top t \\
\text{s.t.} \quad & e^\top \bar{t}_j = 1, \ j = 1, \ldots, n, \\
& t \in \{0, 1\}^{nM},
\end{aligned}
$$

(QAP)

where

$$(2.6) \quad G = P^\top \widehat{Q} P \in \mathbb{R}^{nM \times nM} \ \text{and} \ w = P^\top \hat{c} \in \mathbb{R}^{nM}.$$

Inspired by the sparse formulation in [6, (2.10)], we define the following SQP problem:

$$
\begin{aligned}
\min_{t \in \mathbb{R}^{nM}} \quad & h(t) \\
\text{s.t.} \quad & e^\top \bar{t}_j = 1, \ j = 1, \ldots, n, \\
& t \geqslant 0, \\
& \|t\|_0 \leqslant n,
\end{aligned}
$$

(SQP1)

where $\|t\|_0 \leqslant n$ is the *sparse constraint*, denoting that the sparsity (the number of nonzeros) of t is not greater than n. We have the following result addressing the connection between problems (QAP) and (SQP1).

PROPOSITION 2.1. *Problems* (QAP) *and* (SQP1) *are equivalent.*

Proof. For each feasible point t of problem (QAP), there is $\|\bar{t}_j\|_0 = 1$, implying that $\|t\|_0 \leqslant n$. Consequently, point t is feasible for problem (SQP1). On the other hand, for each feasible point t of problem (SQP1), it follows that $e^\top \bar{t}_j = 1$ for $j = 1, \ldots, n$, and $\|t\|_0 \leqslant n$, implying that $\|\bar{t}_j\|_0 = 1$ for $j = 1, \ldots, n$. Therefore, each entry in $\bar{t}_j$ must be either zero or one, i.e., $t \in \{0, 1\}^{nM}$. This shows that point t is also feasible for problem (QAP). Therefore, problems (QAP) and (SQP1) are equivalent. □

By Proposition 2.1, problem (SQP1) is equivalent to the original problem (P). Specifically, if x^* is a global minimizer of problem (P), then t^* obtained by the following

$$(2.7) \quad \begin{bmatrix} \mathrm{Re}(x^*) \\ \mathrm{Im}(x^*) \end{bmatrix} = P t^* = \begin{bmatrix} A \\ B \end{bmatrix} t^* = \begin{bmatrix} A t^* \\ B t^* \end{bmatrix}$$

is a global minimizer of problem (SQP1). Conversely, for a global minimizer t^* of problem (SQP1), one can get a global minimizer x^* of problem (P) by

$$x^* = A t^* + \mathrm{i} B t^*.$$

This reveals that there is a one-to-one correspondence between the global minimizers of problem (SQP1) and those of problem (P).

Next, we partition the matrix G (defined in (2.6)) as follows:

$$(2.8) \qquad G = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n1} & S_{n2} & \cdots & S_{nn} \end{bmatrix},$$

where $S_{jk} \in \mathbb{R}^{M \times M}$ for $j, \; k = 1, \ldots, n$. Define a new matrix $\widetilde{G} \in \mathbb{R}^{nM \times nM}$ obtained by removing the diagonal blocks in G, i.e., $\widetilde{G} \triangleq G - \widetilde{D}$. Here $\widetilde{D}$ is the matrix with diagonal blocks $S_{11}, \ldots, S_{nn}$, denoted as

$$(2.9) \qquad \widetilde{D} = \mathrm{Diag}(S_{11}, \ldots, S_{nn}).$$

We have the following result.

THEOREM 2.2. *Problem* (SQP1) *is equivalent to the following problem:*

$$\min_{t \in \mathbb{R}^{nM}} \quad f(t) \triangleq t^\top \widetilde{G} t + 2w^\top t$$

(SQP2)
$$\text{s.t.} \quad e^\top \bar{t}_j = 1, \; j = 1, \ldots, n,$$
$$t \geqslant 0,$$
$$\|t\|_0 \leqslant n.$$

Proof. The proof is relegated to Appendix A.   □

Below, we give a property of the objective function $f(t)$ in problem (SQP2) stating that $f(t)$ is a linear function with respect to $\bar{t}_j$, which follows from the fact that the diagonal block in $\widetilde{G}$ is zero. Such a property is similar to that in [6, Proposition 3] for hypergraph matching.

PROPOSITION 2.3. *For each block* $\bar{t}_j, \; j = 1, \ldots, n,$ $f(t)$ *in problem* (SQP2) *is a linear function of* $\bar{t}_j$, *i.e.,* $\nabla_{\bar{t}_j} f(t)$ *is independent of* $\bar{t}_j$.

Due to Proposition 2.3, given a particular block $\bar{t}_j$, the function $f(\cdot)$ can be written as

$$(2.10) \quad f(t) = \nabla_{\bar{t}_j} f(t)^\top \bar{t}_j + f^{-j}(t_{-j}), \;\; t_{-j} \triangleq (\bar{t}_1, \cdots, \bar{t}_{j-1}, \bar{t}_{j+1}, \cdots, \bar{t}_n)^\top \in \mathbb{R}^{(n-1)M}.$$

Here $\nabla_{\bar{t}_j} f(t)$ is only related to t_{-j}, and $f^{-j}(t_{-j})$ represents the part in $f(\cdot)$ which is only related to t_{-j}.

3. Relaxation for MIMO Detection. In this section, we first show the equivalence between problem (SQP2) and its relaxation problem obtained by dropping the sparse constraint. Then we present the properties of the relaxation problem as well as its relations to SDRs.

3.1. Relaxation of Problem (SQP2)**.** By dropping the sparse constraint in problem (SQP2), i.e., $\|t\|_0 \leqslant n$, we get the following relaxation problem:

$$\min_{t \in \mathbb{R}^{nM}} \quad f(t)$$

(RSQP)
$$\text{s.t.} \quad e^\top \bar{t}_j = 1, \; j = 1, \ldots, n,$$
$$t \geqslant 0.$$

The following shows that problem (RSQP) is actually equivalent to problem (SQP2).

THEOREM 3.1. *There exists a global minimizer t^* of problem* (RSQP) *such that* $\|t^*\|_0 = n$. *As a result, t^* is a global minimizer of problem* (SQP2).

Proof. We proceed the proof by showing that for problem (RSQP), there exists a global optimal solution t^* such that each block $\bar{t}_j^*$ is an extreme point of the simplex set Ω defined by $\Omega = \{\nu \in \mathbb{R}^M \mid \nu^\top e = 1, \ \nu \geqslant 0\}$. Let t° be a global optimal solution of problem (RSQP). Suppose that there exists one block as $\bar{t}_{j'}^\circ$, such that $\bar{t}_{j'}^\circ$ is not an extreme point of Ω (i.e., $\|\bar{t}_{j'}^\circ\|_0 > 1$). Clearly, point $\bar{t}_{j'}^\circ$ is an optimal solution of the linear programming problem with simplex constraint

$$(3.1) \qquad \min_{\nu \in \Omega} \nabla_{\bar{t}_{j'}} f(t^\circ)^\top \nu + f^{-j'}(t^\circ_{-j'}),$$

where $\bar{t}^\circ_{-j'}$ is defined similarly as in (2.10). From the basic linear programming theory, there must exist an extreme point $\nu^1 \in \Omega$, such that ν^1 is an optimal solution of problem (3.1). Then we must have $\nabla_{\bar{t}_{j'}} f(t^\circ)^\top \nu^1 + f^{-j'}(t^\circ_{-j'}) = f(t^\circ)$. Define a new point $t^1 \in \mathbb{R}^{nM}$ by $\bar{t}_{j'}^1 = \nu^1$, $\bar{t}_j^1 = \bar{t}_j^\circ$, $j \neq j'$, $j = 1, \ldots, n$. We have $f(t^1) = f(t^\circ)$ and, hence t^1 is a global minimizer of problem (RSQP). If all of the blocks in t^1, i.e., $\bar{t}_j^1$, $j = 1, \ldots, n$, are extreme points of the set Ω, let $t^* = t^1$. The proof is finished. Otherwise, repeat the above process. After at most k steps ($k \leqslant n$), one will reach a global minimizer $t^* \triangleq t^k$, such that all of the blocks in t^* are extreme points of the set Ω. This completes the proof. $\qquad\square$

Remark 3.2. The proof of Theorem 3.1 makes use of the properties of linear programming. Another way to prove the result is to apply Corollary 2 in [6] as well as Proposition 2.3.

Remark 3.3. Suppose that one gets a global minimizer of problem (RSQP), denoted as $t^\circ \in \mathbb{R}^{nM}$. As pointed out in [6, Remark 3], we can get a global minimizer t^* of problem (SQP2) in the following way. For each block $\bar{t}_j^*$, pick up any nonzero entry in $\bar{t}_j^\circ$, say p_j, and set

$$(3.2) \qquad \left(\bar{t}_j^*\right)_{p_j} = 1, \text{ and } \left(\bar{t}_j^*\right)_l = 0, \ l \in \{1, \ldots, M\} \backslash p_j.$$

Notice that $\bar{t}_j^*$ corresponds to an extreme point of the simplex Ω, which is an optimal solution of problem (3.1) with j' replaced by j. Repeatedly applying the above rounding procedure, we will obtain a global minimizer of problem (SQP2).

Remark 3.4. From the strong NP-hardness of problem (P) and the equivalence between problems (SQP2) and (RSQP) (cf. Theorem 3.1), problem (RSQP) is also strongly NP-hard. However, problem (RSQP) enjoys more advantages than problem (P). Firstly, problem (RSQP) is a continuous optimization problem so that the local information can be used to design efficient algorithms whereas problem (P) is a discrete optimization problem. Moreover, the feasible region of problem (RSQP) is described by simplex constraints, which are relatively simple. In terms of the objective function, although it is nonconvex, it is a quadratic function. In particular, we have shown in Proposition 2.3 that it is a linear function for each block $\bar{t}_j$ (with all the others being fixed). Consequently, such a special structured nonlinear programming problem with simplex constraints provides us more freedom to explore various numerical algorithms to solve the problem. In other words, by transforming the discrete problem (P) into the continuous optimization problem (RSQP), we can make full use of various techniques and algorithms in nonlinear optimization.

3.2. Properties of Relaxation Problem (RSQP). An interesting question is that under which condition, the relaxation problem (RSQP) admits a unique global minimizer, which corresponds to the vector of transmitted signals x^* in (1.1) by (2.7). To answer this question, we first characterize the condition under which, problem (RSQP) admits a unique global minimizer.

THEOREM 3.5. *Suppose that t^* is the unique global optimal solution of problem* (SQP2). *Then, t^* is the unique global minimizer of problem* (RSQP).

Proof. We use the contradiction argument. Assume that t^* is not the unique global minimizer of problem (RSQP), then there must exist another global minimizer t° of problem (RSQP) such that $t^\circ \neq t^*$. This, together with the assumption that t^* is the unique global minimizer of problem (SQP2) and $\boldsymbol{e}^T \bar{t}_j^\circ = 1$, implies that $\|t^\circ\|_0 \geqslant n+1$ must hold, and hence there must exist a block $\bar{t}_j^\circ$ such that $\|\bar{t}_j^\circ\|_0 \geqslant 2$. Without loss of generality, let $\|\bar{t}_1^\circ\|_0 \geqslant 2$, $(\bar{t}_1^\circ)_1 > 0$, and $(\bar{t}_1^\circ)_2 > 0$. Applying the rounding procedure in (3.2) by setting

$$(\bar{t}_1^\circ)_j = \left\{ \begin{array}{ll} 1, & \text{if } j = 1; \\ 0, & \text{if } j = 2,\ldots,M, \end{array} \right. \quad \text{and } (\bar{t}_1^\circ)_j = \left\{ \begin{array}{ll} 1, & \text{if } j = 2; \\ 0, & \text{if } j = 1,3,\ldots,M, \end{array} \right.$$

respectively, we will obtain two different global minimizers of problem (RSQP). Repeatedly applying the rounding procedure in (3.2) to other blocks of these two points, we can obtain two *different* global minimizers of problem (SQP2), which contradicts with the assumption that t^* is the unique global minimizer of problem (SQP2). Consequently, t^* is the unique global minimizer of problem (RSQP). ∎

Theorem 3.5 implies that if the vector of transmitted signals x^* is the unique global minimizer of problem (P), then the corresponding t^* obtained via (2.7) is a unique global minimizer of problem (RSQP). The remaining question is under which condition, t^* is the unique global minimizer of problem (SQP2). To address this question, we need the definition of tightness and the enhanced SDR in [17].

DEFINITION 3.6. *An SDR of problem* (P) *is called tight if the following two conditions hold: the gap between the SDR and problem* (P) *is zero; and the SDR recovers the true vector of transmitted signals.*

The enhanced SDR in [17] is briefly described as follows:

$$\min_{\substack{y \in \mathbb{R}^{2n}, \ t \in \mathbb{R}^{nM}, \\ Y \in \mathbb{R}^{2n \times 2n}}} \quad \langle \widehat{Q}, Y \rangle + 2\hat{c}^\top y$$

$$\text{s.t.} \qquad \mathbf{Y}(j) = \sum_{k=1}^{M} (\bar{t}_j)_k U_k, \ j = 1,\ldots,n,$$

(ERSDR1)
$$\sum_{k=1}^{M} (\bar{t}_j)_k = 1, \ j = 1,\ldots,n,$$

$$\begin{bmatrix} 1 & y^\top \\ y & Y \end{bmatrix} \succeq 0,$$

$$t \geqslant 0,$$

where $\widehat{Q}$ is defined as in (2.2),

$$\mathbf{Y}(j) = \begin{bmatrix} 1 & y_j & y_{n+j} \\ y_j & Y_{jj} & Y_{j(n+j)} \\ y_{n+j} & Y_{(n+j)j} & Y_{(n+j)(n+j)} \end{bmatrix}, \ j = 1,\ldots,n,$$

and

$$U_k = \begin{bmatrix} 1 \\ \cos \theta_k \\ \sin \theta_k \end{bmatrix} \begin{bmatrix} 1 & \cos \theta_k & \sin \theta_k \end{bmatrix}, \ k = 1, \dots, M.$$

We have the following result.

THEOREM 3.7. *Let $t^* \in \mathbb{R}^{nM}$ be the vector corresponding to the vector of transmitted signals $x^* \in \mathbb{C}^n$ in (1.1). If condition (1.5) holds, t^* is a unique global minimizer of problem (SQP2).*

Proof. Note that under condition (1.5), problem (ERSDR1) is tight [17, Theorem 4.4]. By the proof in [17, Theorem 4.2, Corollary 4.3, Theorem 4.4], problem (ERSDR1) admits a unique optimal solution, which corresponds to the vector of transmitted signals x^* in (1.1). This, together with the tightness of problem (ERSDR1) and the fact that problem (ERSDR1) is a relaxation of problem (P), shows that x^* is also a unique solution of problem (P). Equivalently, under condition (1.5), t^* is also a unique global minimizer of problem (SQP2). ∎

Remark 3.8. Theorems 3.5 and 3.7 imply that under condition (1.5), problem (RSQP) is also tight.

We illustrate several formulations for the MIMO detection problem in Figure 2, which demonstrates the equivalence between problems (P), (CQP), (RQP), (SQP1), (SQP2), as well as (RSQP).

$$(P) \Leftrightarrow (CQP) \xleftrightarrow{\textit{Real form}} (RQP) \xleftrightarrow[\textit{sparse constraint}]{\textit{Introducing}} (SQP1) \Longleftrightarrow (SQP2) \xleftrightarrow{\textit{Relaxation}} (RSQP)$$

FIG. 2. *The map of equivalent formulations.*

It should be emphasized that problem (RSQP) is a vector based formulation and its size is much smaller (than that of SDRs for problem (P)), and thus it is more suitable to be used for designing algorithms for the large-scale problems. More detailed comparisons between problem (RSQP) and various SDRs will be shown in the next subsection.

3.3. Relations to the SDRs. Recall that the enhanced SDR studied in [17] is *tight* under condition (1.5). In fact, we can also show the tightness result of the SDR of our proposed formulation (SQP2) under the same condition. It is easy to check that the following SDR of problem (QAP) proposed in [22]

$$\begin{aligned} \min_{\substack{T \in \mathbb{R}^{nM \times nM}, \\ t \in \mathbb{R}^{nM}}} \quad & \bar{f}_1(T, t) \triangleq \langle G, T \rangle + 2w^\top t \\ \text{(ERSDR2)} \qquad\qquad \text{s.t.} \quad & e^\top \bar{t}_j = 1, \ j = 1, \dots, n, \\ & T_{jj} = \text{Diag}(\bar{t}_j), \ j = 1, \dots, n, \\ & T \succeq tt^\top, \ t \geq 0, \end{aligned}$$

is equivalent to the following SDR of problem (SQP2)

$$\min_{\substack{T \in \mathbb{R}^{nM \times nM}, \\ t \in \mathbb{R}^{nM}}} \quad \bar{f}_2(T, t) \triangleq \langle \widetilde{G}, T \rangle + 2w^\top t$$

(ERSDR3)
$$\text{s.t.} \quad \boldsymbol{e}^\top \bar{t}_j = 1, \ j = 1, \ldots, n,$$
$$T_{jj} = \text{Diag}(\bar{t}_j), \ j = 1, \ldots, n,$$
$$T \succeq tt^\top, \ t \geqslant 0,$$

where $T_{jj} \in \mathbb{R}^{M \times M}$ is the j-th diagonal block of T. With [15, Theorem 2], problem (ERSDR3) is tight for problem (P) under condition (1.5) for $M \geqslant 2$.

Now, the relations between the series of "ERSDRs" and other formulations discussed above can be summarized in Figure 3. Problems (ERSDR1), (ERSDR2), and (ERSDR3) are SDRs of problems (RQP), (SQP1), and (SQP2), respectively. These "ERSDRs" are equivalent.
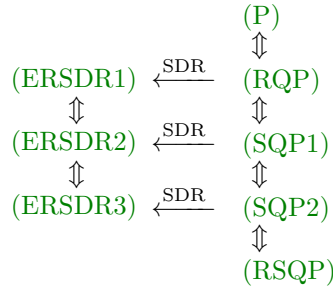
$$
\begin{array}{ccc}
 & & \text{(P)} \\
 & & \Updownarrow \\
\text{(ERSDR1)} & \xleftarrow{\text{SDR}} & \text{(RQP)} \\
\Updownarrow & & \Updownarrow \\
\text{(ERSDR2)} & \xleftarrow{\text{SDR}} & \text{(SQP1)} \\
\Updownarrow & & \Updownarrow \\
\text{(ERSDR3)} & \xleftarrow{\text{SDR}} & \text{(SQP2)} \\
 & & \Updownarrow \\
 & & \text{(RSQP)}
\end{array}
$$

Fig. 3. *Relations between the series of "ERSDRs" and other formulations.*

To conclude this section, we summarize the scale of the above problems in terms of the number of variables and the number of different types of constraints in Table 1. In Table 1, '=' means equality constraints, '$\succeq$' means positive semidefinite constraints, and '$\geqslant$' means lower bound constraints. It can be seen from Table 1 that problem (RSQP) only involves one vector variable $t \in \mathbb{R}^{nM}_+$ and n linear equality constraints, both of which are significantly smaller than those of other relaxation problems. In addition, all constraints in problem (RSQP) are linear. In sharp contrast, all SDR problems contain a positive semidefinite constraint. Our proposed relaxation problem (RSQP) enables us to develop fast algorithms for solving the large-scale MIMO detection problem. Indeed, our proposed PN-QP method for solving the MIMO detection problem is customized based on problem (RSQP), and as it will be shown in section 5, it is much more efficient compared to state-of-the-art ERSDR based approaches.

TABLE 1
Comparison of different relaxations.

Problem	Number of variables		Number of constraints		
	vector	matrix (size)	=	$\succeq$ (size)	$\geqslant$
(ERSDR1)	$2n + nM$	$1(2n \times 2n)$	$6n$	$1(4n \times 4n)$	nM
(ERSDR2)	nM	$1(nM \times nM)$	$n + nM(M+1)/2$	$1(nM \times nM)$	nM
(ERSDR3)	nM	$1(nM \times nM)$	$n + nM(M+1)/2$	$1(nM \times nM)$	nM
(RSQP)	nM	0	n	0	nM

4. Numerical Algorithm for Problem (RSQP). In this section, we present the numerical algorithm for solving problem (RSQP) and discuss its convergence result.

4.1. Quadratic Penalty Method. Recall that problem (RSQP) is a nonlinear programming problem with n simplex constraints. Hence, one can use a solver for constrained optimization problems like `fmincon` in MATLAB to solve it. However, due to the special property as stated in Remark 3.3, once the support set of the global minimizer of problem (RSQP) is correctly identified, we can apply the rounding procedure in (3.2) to obtain a global minimizer of problem (SQP2). Based on such observations, instead of directly solving problem (RSQP) by treating it as a general constrained optimization problem, we prefer to design an algorithm to (quickly) identify the support set of the global minimizer of problem (RSQP). Due to this, such an algorithm does not need to strictly satisfy the equality constraints during the algorithmic procedure, i.e., it is reasonable to allow the violations of the equality constraints to some extent. Therefore, we choose the quadratic penalty method to solve problem (RSQP). More precisely, at each iteration k, the quadratic penalty method solves the following subproblem:

$$\min_{t\in\mathbb{R}^{nM}} \quad f_{\omega_k}(t) \triangleq f(t) + \frac{\omega_k}{2}\sum_{j=1}^{n}(\boldsymbol{e}^\top\bar{t}_j - 1)^2$$

$$\text{s.t.} \quad t \geqslant 0,$$

where $\omega_k > 0$ is the penalty parameter. The above subproblem is in general unbounded when $t \to +\infty$. Therefore, we solve the following subproblem instead

$$(4.1)\qquad\qquad\qquad \min_{t\in\mathbf{B}} \quad f_{\omega_k}(t)$$

where $\mathbf{B} \triangleq \{t \in \mathbb{R}^{nM} \mid 0 \leqslant t_j \leqslant K,\ j = 1, \ldots, nM\}$, with $K \geqslant 1$ being a sufficiently large number to guarantee the boundedness of the feasible region. The above problem (4.1) is the penalized subproblem of the following problem

$$\min_{t\in\mathbb{R}^{nM}} \quad f(t)$$

$$\text{s.t.} \quad \boldsymbol{e}^\top\bar{t}_j = 1,\ j = 1, \ldots, n,$$

$$t \in \mathbf{B},$$

which is equivalent to problem (RSQP). Next, we provide more details on the stopping criteria of the quadratic penalty algorithm and the algorithm for solving the subproblem (4.1).

Let t^k be an approximate solution of subproblem (4.1). As for the stopping criteria, we check whether the support set of t^k is the same as that of the previous step and whether the size of the support set of $\bar{t}_j^k$ is equal to one for all $j = 1, \ldots, n$, i.e.,

$$\mathcal{K}(t^k) = \mathcal{K}(t^{k-1}) \text{ and } \|\bar{t}_j^k\|_0 = 1,\ j = 1, \ldots, n,$$

where $\mathcal{K}(t)$ is the support set of $t \in \mathbb{R}^{nM}$ defined as $\mathcal{K}(t) = \{\ell \mid t_\ell > 0,\ \ell = 1, \ldots, nM\}$. If the above conditions are satisfied, it implies that we reach a feasible point of problem (SQP2) with sparsity n, we terminate the iteration. From the numerical point of view, the condition $\|\bar{t}_j^k\|_0 = 1$ is implemented by

$$(4.2)\qquad\qquad\qquad \|\bar{t}_j^k\|_{0,\,\varepsilon} = 1,\ j = 1, \ldots, n,$$

where $\|\bar{t}_j^k\|_{0,\,\varepsilon}$ denotes the number of elements which are significantly larger than zero, that is, $(\bar{t}_j^k)_l > \varepsilon$, and $0 < \varepsilon < 1$ is a prescribed small number.

As for subproblem (4.1), one equivalent characterization of the stationary point is

$$t^k - \Pi_{\mathbf{B}}(t^k - \nabla f_{\omega_k}(t^k)) = 0,$$

where $\Pi_{\mathbf{B}}(t)$ denotes the projection of $t \in \mathbb{R}^{nM}$ onto the set $\mathbf{B}$. Here we solve subproblem (4.1) inexactly to get a solution t^k, that is, t^k satisfies

$$(4.3) \qquad \|t^k - \Pi_{\mathbf{B}}(t^k - \nabla f_{\omega_k}(t^k))\|_2 \leqslant \tau_k,$$

where $\tau_k \downarrow 0$.

Note that subproblem (4.1) is a non-convex quadratic programming problem with simple lower and upper bound constraints. As mentioned above, we prefer to identify the support set of the global minimizer of subproblem (4.1) rather than find the global minimizer itself (in order to reduce the computational cost). The strategy of identifying the active set is therefore crucial in solving subproblem (4.1). From this point of view, the active set methods are particularly suitable to solve subproblem (4.1). Therefore, we choose the typical active set method, the projected Newton method proposed in [2], which is demonstrated to be highly efficient in solving large-scale problems such as calibrating least squares covariance matrices [13].

Overall, we give the details of the PN-QP method in Algorithm 4.1.

Algorithm 4.1 PN-QP Method.

1: Initialization: $t^0 \in \mathbb{R}^{nM}$, $k := 1$, $\rho > 1$, $\tau_k \downarrow 0$, $\varepsilon > 0$, and $\omega_k > 0$;
2: **while** $k \leqslant$ maxiter **do**
3:     Solve subproblem (4.1) by the projected Newton method to get t^k such that t^k satisfies (4.3);
4:     **if** conditions $\mathcal{K}(t^k) = \mathcal{K}(t^{k-1})$ and (4.2) are satisfied **then**
5:         Break;
6:     **end if**
7:     Set $\omega_{k+1} := \rho\omega_k$, $k := k + 1$;
8: **end while**
9: **return** t^k.

Remark 4.1. Here we would like to highlight that due to the special strategy in identifying the active set for lower and upper constraints, the projected Newton method [2] is guaranteed to identify the active set of the stationary point of the problem in the form of subproblem (4.1) [2, Proposition 2]. Moreover, since the second-order information is employed in the projected Newton method, under reasonable assumptions, it is able to converge to a local optimal solution of subproblem (4.1) [2, Propositions 3, 4].

We have the following classic convergence result of the quadratic penalty method [25, Chapter 17]. Due to the limitation on the length of the paper, we omit the proof here.

THEOREM 4.2. *Suppose that in Algorithm 4.1, the sequence $\{t^k\}$ satisfies (4.3), $\tau_k \downarrow 0$, and $\omega_k \uparrow +\infty$. Then any accumulation point of the sequence generated by Algorithm 4.1 is a stationary point of problem* (RSQP).

For Algorithm 4.1, it is possible that the sparsity of the resulting stationary point of $\{t^k\}$ may be greater than n. Below we design a special rounding algorithm, which is guaranteed to return a feasible point of problem (SQP2) with sparsity n.

4.2. Rounding Algorithm. To present the rounding algorithm, we need the following equivalent characterization of stationary points [6, Lemma 1 (i)]. For $\nu \in \mathbb{R}^M$, let $\Gamma(\nu) = \{l \mid \nu_l > 0\}$ and $\mathcal{I}(\nu) = \{l \mid \nu_l = 0\}$.

PROPOSITION 4.3. *A vector $t \in \mathbb{R}^{nM}$ is a stationary point of problem* (RSQP) *if and only if the following conditions hold at t for $j = 1, \ldots, n$,*

$$(4.4) \qquad (\nabla_{\bar{t}_j} f(t))_s = (\nabla_{\bar{t}_j} f(t))_k, \ \forall \ s, k \in \Gamma(\bar{t}_j),$$

$$(4.5) \qquad (\nabla_{\bar{t}_j} f(t))_l \geqslant (\nabla_{\bar{t}_j} f(t))_s, \ \forall \ l \in \mathcal{I}(\bar{t}_j), \ \forall \ s \in \Gamma(\bar{t}_j).$$

The following rounding algorithm for computing a stationary point of problem (RSQP) is slightly different from (3.2). In particular, instead of picking any index of the nonzero entries, we shall pick an index whose corresponding gradient is the smallest, which will help in obtaining a smaller function value.

Algorithm 4.2 Rounding Algorithm

1: Initialization: a stationary point t^0 of problem (RSQP), $j = 1$;
2: **while** $j \leqslant n$ **do**
3:     Let $s_j \in \arg\min_{l \in \{1, \ldots, M\}} (\nabla_{\bar{t}_j} f(t^{j-1}))_l$;
4:     Define $\nu \in \mathbb{R}^M$ by $\nu_{s_j} = 1$, $\nu_l = 0$, $l \neq s_j$, $l \in \{1, \ldots, M\}$;
5:     Define $t^j \in \mathbb{R}^{nM}$ by

$$(4.6) \qquad \bar{t}_j^j = \nu, \ \bar{t}_k^j = \bar{t}_k^{j-1}, \ k \neq j, \ k \in \{1, \ldots, n\} \backslash \{j\};$$

6:     Let $j := j + 1$;
7: **end while**
8: **return** t^n.

We have the following properties about Algorithm 4.2.

PROPOSITION 4.4. *Let t^0 be a stationary point of problem* (RSQP). *Running Algorithm 4.2 with input t^0, we have $\|t^n\|_0 = n$ and*

$$(4.7) \qquad f(t^n) \leqslant f(t^0).$$

Proof. The sparsity of t^n can be obtained directly by the process of Algorithm 4.2. For (4.7), by the definition of t^j in (4.6), Proposition 2.3, as well as Proposition 4.3, one can obtain that $f(t^j)$ does not exceed $f(t^{j-1})$, $j = 1, \ldots, n-1$, giving (4.7). $\square$

Remark 4.5. Proposition 4.4 reveals that for a stationary point t^0 returned by Algorithm 4.1, the rounding procedure Algorithm 4.2 will return a feasible point, whose sparsity is n and whose function value does not exceed $f(t^0)$.

4.3. Exact Detection of PN-QP. To further discuss under which condition the PN-QP method has an exact detection guarantee, i.e., the PN-QP method is guaranteed to return the optimal solution t^* corresponding to the vector of transmitted signals x^*, let us denote

$$(4.8) \qquad \mathring{Q} = H^\dagger H - \text{Diag}(\text{diag}(H^\dagger H)) \text{ and } \overline{Q} = H^\dagger H - \frac{1}{2}\text{Diag}(\text{diag}(H^\dagger H)).$$

We need the following two lemmas whose proofs are elementary and therefore were provided in a separate technical report[1].

LEMMA 4.6. *Let x^* be the vector of transmitted signals satisfying* (1.1) *and $t^* \in \mathbb{R}^{nM}$ be the corresponding vector by* (2.7). *Let x be any feasible point of problem* (P) *with $x \neq x^*$ and (similarly) $t \in \mathbb{R}^{nM}$ be the corresponding vector, that is,*

$$\begin{bmatrix} \mathrm{Re}(x) \\ \mathrm{Im}(x) \end{bmatrix} = Pt = \begin{bmatrix} At \\ Bt \end{bmatrix},$$

where A, B, and P are defined in (2.5). *Furthermore, assume that the phases for x_j and x_j^* are θ_j and θ_j^*, respectively, $j = 1, \ldots, n$. We have the following results:*

$$(4.9) \qquad \|x^* - x\|_2^2 = 4 \sum_{j=1}^n \sin^2 \left(\frac{\theta_j^* - \theta_j}{2} \right),$$

$$(4.10) \qquad \|x^* - x\|_1 = 2 \sum_{j=1}^n \sin \left(\left| \frac{\theta_j^* - \theta_j}{2} \right| \right),$$

$$(4.11) \qquad (t^* - t)^\top \widetilde{G}(t^* - t) = \langle x^* - x, \mathring{Q}(x^* - x) \rangle,$$

$$(4.12) \qquad (t^* - t)^\top (Gt^* + w) = \langle -H^\dagger v, x^* - x \rangle,$$

$$(4.13) \qquad 2(t^*)^\top \widetilde{D}(t^* - t) = \langle x^* - x, \mathrm{Diag}(\mathrm{diag}(Q))(x^* - x) \rangle.$$

LEMMA 4.7. $|\langle \xi, \xi^* \rangle| \leqslant \sqrt{2} \|\xi\|_\infty \|\xi^*\|_1, \ \forall \ \xi, \ \xi^* \in \mathbb{C}^n.$

With the above two lemmas, we have the following result.

LEMMA 4.8. *Let t, t^*, x, and x^* be defined as in* Lemma 4.6. *If the following condition holds*

$$(4.14) \qquad \sqrt{2} \lambda_{\min}(\overline{Q}) \sin \left(\frac{\pi}{M} \right) > \|H^\dagger v\|_\infty,$$

then there is $\nabla f(t)^\top (t^* - t) < 0$.

Proof. With $\nabla f(t) = 2\widetilde{G}t + 2w$, there is

$\nabla f(t)^\top (t^* - t)$

$= 2(\widetilde{G}t + w)^\top (t^* - t)$

$= 2(\widetilde{G}(t - t^*))^\top (t^* - t) + 2(\widetilde{G}t^* + w)^\top (t^* - t)$

$= -2(t^* - t)^\top \widetilde{G}(t^* - t) + 2(Gt^* + w)^\top (t^* - t) - 2(\widetilde{D}t^*)^\top (t^* - t)$ (by (2.9))

$= -2\langle x^* - x, \mathring{Q}(x^* - x) \rangle + 2\langle -H^\dagger v, x^* - x \rangle - \langle x^* - x, \mathrm{Diag}(\mathrm{diag}(Q))(x^* - x) \rangle$ (by *Lemma* 4.6)

$= -2\langle x^* - x, \overline{Q}(x^* - x) \rangle + 2\langle -H^\dagger v, x^* - x \rangle$ (by (4.8))

$\leqslant -2\lambda_{\min}(\overline{Q})\|x^* - x\|_2^2 + 2\sqrt{2}\|H^\dagger v\|_\infty \|x^* - x\|_1$ (by *Lemma* 4.7)

$= -8\lambda_{\min}(\overline{Q}) \sum_{j=1}^n \sin^2 \left(\frac{\theta_j^* - \theta_j}{2} \right) + 4\sqrt{2}\|H^\dagger v\|_\infty \sum_{j=1}^n \sin \left(\left| \frac{\theta_j^* - \theta_j}{2} \right| \right)$ (by *Lemma* 4.6)

$= 4\sqrt{2} \sum_{j=1}^n \left(-\sqrt{2}\lambda_{\min}(\overline{Q}) \sin^2 \left(\frac{\theta_j^* - \theta_j}{2} \right) + \|H^\dagger v\|_\infty \sin \left(\left| \frac{\theta_j^* - \theta_j}{2} \right| \right) \right)$

$\triangleq 4\sqrt{2} \sum_{j=1}^n \Psi \left(\sin \left(\left| \frac{\theta_j^* - \theta_j}{2} \right| \right) \right).$

[1] http://lsec.cc.ac.cn/~yafliu/technical_report_MIMO.pdf

If (4.14) holds, there is $\lambda_{\min}(\overline{Q}) > 0$ and

$$\frac{\|H^\dagger v\|_\infty}{\sqrt{2}\lambda_{\min}(\overline{Q})} < \sin\left(\frac{\pi}{M}\right).$$

Consequently, $\Psi(\cdot)$ is decreasing over the interval $\left[\sin\left(\frac{\pi}{M}\right), +\infty\right)$, implying that

$$\Psi\left(\sin\left(\frac{\pi}{M}\right)\right) < 0.$$

Therefore, with the fact that $x^* \neq \tilde{x}$ and $x^*, \tilde{x} \in \mathcal{X}$, we have

$$\nabla f(t)^\top (t^* - t) \leqslant 4\sqrt{2}\sum_{j=1}^n \Psi\left(\sin\left(\left|\frac{\theta_j^* - \theta_j}{2}\right|\right)\right)$$

$$= 4\sqrt{2}\sum_{\theta_j^* \neq \theta_j} \Psi\left(\sin\left(\left|\frac{\theta_j^* - \theta_j}{2}\right|\right)\right)$$

$$\leqslant 4\sqrt{2}\sum_{\theta_j^* \neq \theta_j} \Psi\left(\sin\left(\frac{\pi}{M}\right)\right) < 0.$$

The proof is completed.   □

With Theorem 3.7, Theorem 4.2, and Lemma 4.8, we have the following result.

THEOREM 4.9. *Under conditions (1.5) and (4.14), the sequence generated by Algorithm 4.1 will converge to the unique global minimizer t^* of problem (RSQP), which corresponds to the vector of transmitted signals x^* in (1.1).*

Proof. Under condition (1.5), Theorem 3.7 implies that t^* is the unique global minimizer of problem (RSQP). Together with Lemma 4.8, any feasible point of problem (RSQP) with sparsity n other than t^* is not a stationary point, since $t^* - t$ is a descent direction of the function f at t satisfying $(\nabla f(t))^\top (t^* - t) < 0$. Consequently, among all the points with sparsity n, t^* is the unique stationary point of problem (RSQP). By Theorem 4.2, the accumulation point of the sequence generated by Algorithm 4.1 will converge to t^*, which corresponds to the vector of transmitted signals x^*. The proof is completed.   □

Remark 4.10. As mentioned in [27], if H has i.i.d. standard complex Gaussian entries, then $H^\dagger H$ is very close to a diagonal matrix with a very high probability. Assume that $H^\dagger H$ is a diagonal matrix, there is $Q = \text{Diag}(q_{11}, \ldots, q_{nn}) = 2\overline{Q}$. In this case, condition (4.14) reduces to $\lambda_{\min}(H^\dagger H)\sin\left(\frac{\pi}{M}\right) > \sqrt{2}\|H^\dagger v\|_\infty$, which is in general stronger than condition (1.5).

5. Numerical Results. In this section, we conduct extensive numerical tests to verify the efficiency of the proposed PN-QP algorithm. The algorithm is implemented in MATLAB (R2017a) and all the experiments are preformed on a Lenovo ThinkPad laptop with Intel dual core i5-6200 CPU (2.30 GHZ and 2.40 GHz) and 8 GB of memory running in Windows 10. We generate the instances of problem (P) following the way in [14, 17], which is detailed as follows:

Step 1: Generate each entry of the channel matrix $H \in \mathbb{C}^{m \times n}$ according to the complex standard Gaussian distribution (with zero mean and unit variance);
Step 2: Generate each entry of the noise vector $v \in \mathbb{C}^m$ according to the complex Gaussian distribution with zero mean and variance σ^2;

Step 3: Choose k_j uniformly and randomly from $\{0, 1, \ldots, M - 1\}$, and set $x_j^* = \exp\left(\frac{2\pi k_j i}{M}\right)$ for each $j \in \{1, \ldots, n\}$, where x^* is the vector of transmitted signals;

Step 4: Compute the vector of received signals $r \in \mathbb{C}^m$ as in (1.1).

Generally, in practical digital communications, M is taken as an exponential power of 2. Therefore, in our following tests, we always choose $M = 2^l$, where l is a positive integer. In our setting, we define the *signal-to-noise ratio* (SNR) as

$$\text{SNR} = 10 \log_{10}\left(\frac{\mathbb{E}[\|Hx^*\|_2^2]}{\mathbb{E}[\|v\|_2^2]}\right) = 10 \log_{10}\left(\frac{m\sigma_x^2}{\sigma_v^2}\right),$$

where $\sigma_x^2 = \mathbb{E}[\|x^*\|_2^2]$, $\sigma_v^2 = \mathbb{E}[\|v\|_2^2]$, and $\mathbb{E}[\cdot]$ is the expectation operator. Then according to our ways of generating instances (i.e., $\sigma_x^2 = n$, and $\sigma_v^2 = m\sigma^2$), we have $\text{SNR} = 10 \log_{10}\left(\frac{n}{\sigma^2}\right)$ in our tests. Generally, the MIMO detection problem is more difficult when the SNR is low and when the numbers of inputs and outputs are equal (i.e., $m = n$).

5.1. Performance of PN-QP. We set $\rho = 3$, $\omega_0 = 10$, $\varepsilon = 0.01$ and $\tau_k = 0.01$ in Algorithm 4.1 and apply Algorithm 4.2 as the rounding procedure after running Algorithm 4.1.

First, we demonstrate the efficiency of PN-QP by an example with $(m, n, M) = (4, 4, 8)$, and SNR = 30 dB. The initial point of PN-QP is chosen as $t^0 = t_e$, where

$$(5.1) \qquad\qquad t_e = \left(\frac{1}{0.2 + M}\right)e \in \mathbb{R}^{nM}.$$

We choose such an initial point t_e since it is a feasible point of subproblem (4.1), and it approximately satisfies the equality constraints in (RSQP), i.e., $(\bar{t}_e)_j^\top e \approx 1$, $j = 1, \ldots, n$. We selectively plot the iterates $\{t^k\}$ in Figure 4 with $k = 0$, 10, and 25. In Figure 4, the '$*$' denotes the vector t^* corresponding to the vector of transmitted signals x^* in (1.1) and the '$\circ$' denotes the iterate t^k generated by PN-QP. It can be seen from Figure 4 that as the iteration goes on, t^k becomes more and more sparse, and eventually, the support set of t^k at $k = 25$ coincides with that of the true minimizer of problem (RSQP).

5.2. Comparison with Other Algorithms. In this subsection, we will compare the numerical performance of PN-QP for solving problem (RSQP) with different models and the corresponding algorithms, which are detailed below.

- Problem (P) solved by GPM [14]: GPM is essentially a gradient projection method whose projection step is taken directly over the discrete set $\mathcal{X}$. Due to its low computational complexity, it is able to solve the large-scale problem. Moreover, in our implementation, we modify its output by choosing the best point among all generated iterates (to improve its performance), instead of simply using the last iterate x^k as the output.
- Problem (P) solved by SD[2] [7]: SD is a typical tree search based method which searches for constellation points limited to a sphere with a predetermined radius centered on the vector of received signals r to find the *global*

---

[2]The code is downloaded from https://ww2.mathworks.cn/matlabcentral/fileexchange/22890-sphere-decoderfor-mimo-systems and modified by adopting the techniques proposed in [4] to further improve its efficiency.

(a) $k = 0$ (the initial point).

(b) $k = 10$.

(c) $k = 25$.
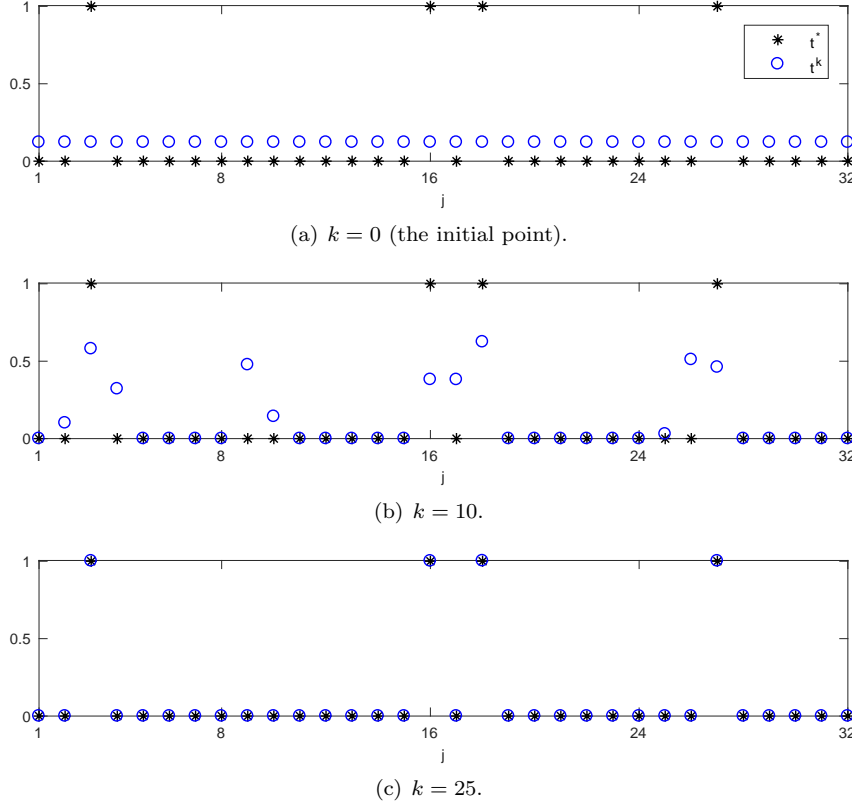
FIG. 4. *Entries in t^k generated by* PN-QP *with $k = 0$, 10, and 25.*

solution. However, the complexity of SD is generally exponential, and hence it is impractical to solve the large-scale problem.

- Problem (ERSDR1) solved by MOSEK[3] [24]: There are several state-of-the-art solvers for SDRs including MOSEK and SDPNAL+ [28, 33, 36, 37]. Here we choose MOSEK since it is a classical interior-point algorithm based solver and is faster than SDPNAL+ for solving problem (ERSDR1). As shown in Table 1, the size of problems (ERSDR2) and (ERSDR3) is significantly larger than that of problem (ERSDR1), and the three problems are (mathematically) equivalent due to [15, Theorem 1] and the discussions in subsection 3.3. Therefore, in our following tests, we do not compare the performance of solving problems (ERSDR2) and (ERSDR3). For (t^k, y^k, Y^k) of problem (ERSDR1) returned by MOSEK, we perform the following rounding procedure to obtain a feasible point: project (y_j^k, y_{n+j}^k) to $\mathcal{Y}$ in (2.1), and get $(y_j^\star, y_{n+j}^\star)$ by

$$(y_j^\star, y_{n+j}^\star) \in \arg \min_{(u_1, u_2) \in \mathcal{Y}} (u_1 - y_j^k)^2 + (u_2 - y_{n+j}^k)^2, \ j = 1, \dots, n.$$

Then return $x_j = y_j^\star + \mathrm{i} y_{n+j}^\star$, $j = 1, \dots, n$.

In our experiments, we set $M \in \{2, 8, 16\}$, $n \in \{16, 32, 64, 128, 256, 512\}$, and

---

[3]https://www.mosek.com

$m = n$ or $m = 2n$. We use the following two metrics to evaluate the performance of different algorithms: the symbol error rate (denoted by SER) [14, 31] as well as the running time in seconds (denoted by Time). More specifically, the SER is used to evaluate the detection error rate of different algorithms, which is calculated by

$$\frac{\text{The number of incorrectly recovered entries compared to } x^*}{\text{The length of transmitted signals } n}.$$

Time is used to evaluate the speed of different algorithms, which is particularly important for solving large-scale problems. We limit the maximum running time of each algorithm to be 3600 seconds. That is, we will terminate the algorithm if its running time is over 3600 seconds and we use "—" to denote such case. The reported results below are obtained by averaging over 100 randomly generated instances.

Initial Points. Since both PN-QP and GPM require the initial points, we first compare the effects of different initial points on the detection results of the two algorithms. We test three initial points: the zero vector $\mathbf{0}$, the vector $t_{\boldsymbol{e}}$ defined as in (5.1), and the approximate solution t_{ml} obtained by the minimum mean square error (MMSE) detector [14]. We report the results in Table 2, where for each setting (m, n, M, SNR) and each method, the winner is marked in bold among three initial points.

It can be observed from Table 2 that in terms of the time for PN-QP, for small scale of (m, n), t_{ml} takes the smallest time whereas for large scale of (m, n), $t_{\boldsymbol{e}}$ and t_{ml} are comparable. Coming to the SER for PN-QP, $t_{\boldsymbol{e}}$ is more favorable among the three choices. In comparison, the SER by GPM varies quite a lot among the three choices of the initial point, and t_{ml} is definitely the best initial point for GPM which will lead to a much smaller SER. In terms of the running time, it seems that t_{ml} for PN-QP is preferable for small-scale problems whereas $t_{\boldsymbol{e}}$ leads to the smallest running time for PN-QP among the three choices when the size of problem n is large (i.e., $n = 512$). For GPM, t_{ml} is also the winner from the perspective of running time. Based on the above observations, in our following test, we choose $t_{\boldsymbol{e}}$ as the initial point for PN-QP and the MMSE estimator t_{ml} as the initial point for GPM. Here we would like to highlight that the numerical results with $n \geqslant 128$ which we show in this paper have not appeared in literature.

Results on Problems with $M \leqslant 8$. Next, we compare the performance of the four algorithms in the cases that $M = 2$ and $M = 8$. We also report the *no interference lower bound* (LB) results. This approach solves the MIMO detection problem with respect to each component x_j assuming all the others being fixed to be the true transmitted signals. Again, the SER is obtained by dividing the total number of incorrectly estimated elements over the length of transmitted signals. The solution returned by LB can be viewed as the best possible result that the MIMO detection problem can be solved theoretically. Therefore, the above no interference LB can be used as the theoretical (and generally unachievable especially in the low SNR scenarios) lower bound of the SER of all the other approaches.

It can be observed from Table 3 and Table 4 that as the SNR decreases, the SER achieved by each algorithm increases, implying that the problem becomes more difficult. As shown in Table 3, when $M = 2$, all methods perform well in terms of the SER.

From Table 3 and Table 4, one can see that the running time for SD becomes longer and even prohibitively high as the SNR decreases/n increases, despite that SD provides the best SER among the four algorithms. For the other three algorithms, in general PN-QP provides the best SER, and MOSEK performs better than GPM

TABLE 2
Comparison of initial points for PN-QP and GPM.

(m, n, M)	SNR (dB)	Time(s)						SER(%)					
		PN-QP (RSQP)			GPM (P)			PN-QP (RSQP)			GPM (P)		
		0	t_e	t_{ml}	**0**	t_e	t_{ml}	**0**	t_e	t_{ml}	**0**	t_e	t_{ml}
(64, 32, 8)	24	0.034	0.034	**0.012**	0.001	0.001	**0.000**	0.00	0.00	0.00	12.44	12.31	**0.00**
	20	0.029	0.033	**0.011**	0.001	0.001	**0.000**	0.00	0.00	0.00	12.13	13.13	**0.00**
	16	0.030	0.036	**0.011**	0.001	0.001	**0.000**	0.00	0.00	0.00	7.06	9.16	**0.00**
	12	0.035	0.038	**0.012**	0.001	0.001	**0.000**	0.25	0.25	0.25	5.72	5.47	**0.91**
(32, 32, 8)	24	0.046	0.050	**0.014**	0.001	0.001	**0.000**	0.19	**0.00**	**0.00**	56.22	58.28	**1.25**
	20	0.045	0.048	**0.017**	0.001	0.001	**0.000**	1.06	**0.25**	**0.25**	57.50	57.19	**2.78**
	16	0.058	0.052	**0.023**	0.001	0.001	0.001	3.59	**2.00**	4.25	57.63	55.75	**11.09**
	12	0.063	0.063	**0.031**	0.001	0.001	0.001	17.59	**16.13**	20.28	52.13	52.16	**25.44**
(64, 32, 16)	24	0.079	0.091	**0.027**	0.001	0.001	**0.000**	0.00	0.00	0.00	40.75	40.41	**0.00**
	20	0.082	0.095	**0.028**	0.001	0.001	**0.000**	0.00	0.00	0.00	30.06	30.72	**0.06**
	16	0.096	0.115	**0.036**	0.001	0.001	**0.000**	2.63	**2.31**	2.78	32.06	29.94	**3.22**
	12	0.114	0.123	**0.057**	0.001	0.001	0.001	18.13	**17.94**	18.88	28.16	29.91	**18.84**
(32, 32, 16)	24	0.133	0.142	**0.095**	0.002	0.001	**0.000**	1.44	**0.53**	5.28	67.28	66.00	**15.91**
	20	0.171	0.176	**0.133**	0.002	0.001	**0.000**	6.75	**5.19**	13.09	68.53	69.34	**22.03**
	16	0.190	0.196	**0.172**	0.002	**0.001**	**0.001**	26.91	**25.41**	32.41	72.69	71.69	**38.63**
	12	0.198	0.203	**0.158**	0.001	0.001	0.001	48.34	**46.53**	48.84	68.91	69.53	**52.56**
(1024, 512, 8)	24	0.746	0.653	**0.245**	1.623	1.802	**0.136**	0.00	0.00	0.00	14.38	15.52	**0.00**
	20	0.767	0.676	**0.240**	0.891	1.175	**0.134**	0.00	0.00	0.00	5.82	9.28	**0.00**
	16	0.837	0.748	**0.259**	0.543	0.523	**0.145**	0.00	0.00	0.00	1.86	2.28	**0.00**
	12	1.153	1.053	**0.583**	0.351	0.351	**0.201**	0.26	**0.25**	0.27	0.42	0.40	**0.38**
(512, 512, 8)	24	1.235	1.108	**0.461**	1.519	1.471	**0.116**	0.00	0.00	0.00	57.48	57.74	**0.00**
	20	1.317	1.190	**0.723**	1.525	1.492	**0.141**	0.00	0.00	0.00	57.78	58.20	**0.00**
	16	1.810	**1.621**	1.676	1.555	1.430	**0.923**	0.10	0.10	0.10	58.14	58.38	**8.55**
	12	**3.101**	3.155	3.520	1.514	1.470	**0.733**	17.41	**17.00**	20.69	58.51	59.03	**24.07**
(1024, 512, 16)	24	1.469	1.431	**0.512**	3.646	3.900	**0.136**	0.00	0.00	0.00	55.38	55.37	**0.00**
	20	1.714	1.664	**0.651**	3.815	3.948	**0.168**	0.01	0.01	0.01	53.86	54.08	**0.01**
	16	2.616	2.492	**1.852**	4.094	3.983	**0.295**	**1.92**	1.95	2.11	47.75	48.16	**2.04**
	12	4.082	**3.896**	4.412	1.394	1.430	**0.407**	**18.64**	18.64	19.69	25.36	25.53	**19.82**
(512, 512, 16)	24	3.403	**3.116**	3.193	2.311	2.401	**2.243**	0.00	0.00	0.00	77.61	77.68	**23.54**
	20	7.071	**6.434**	8.134	2.426	2.450	**2.235**	**3.13**	**3.13**	8.98	77.97	77.97	**33.35**
	16	6.525	**6.355**	7.881	2.337	2.586	**2.270**	29.20	**29.18**	32.58	77.72	77.69	**45.58**
	12	6.900	**6.686**	8.725	2.326	**2.199**	2.360	48.27	**48.14**	50.17	78.29	78.13	**55.01**

in terms of the SER. For example, for $(m, n, M) = (128, 128, 8)$ and SNR $= 14$ dB, PN-QP takes 0.277 seconds to return a solution with SER $= 4.57\%$, whereas MOSEK takes about one second to return a solution with a larger SER 7.14%, and GPM returns a solution with the SER being 17.13% instantly (i.e., 0.012 seconds). For such example, SD fails to return a solution within one hour. For the running time, GPM for solving problem (P) is the fastest one since it only involves gradient calculation and projection onto a discrete set. PN-QP for solving problem (RSQP) is fairly fast and its running time increases slowly as n increases. This is due to the vector formulation of problem (RSQP). MOSEK for solving problem (ERSDR1) is not as fast as GPM and PN-QP. As n increases, the running time increases much faster than that for PN-QP. Comparing $(m, n, M) = (256, 256, 8)$ with $(m, n, M) = (512, 512, 8)$, it can be observed that as n increases from 256 to 512, the running time for MOSEK increases from about 10 seconds to about 85 seconds. This can be explained by the numbers of the variables (one $2n \times 2n$ matrix variable and one $2n + nM$ vector variable) and constraints (one $4n \times 4n$ positive semidefinite constraint, $6n$ equality constraints, and nM inequality constraints) in Table 1. Therefore, in our subsequent test, we will not include SD and MOSEK.

To better understand the detection performance of the four algorithms, we plot Figure 5, showing the SER with respect to the SNR for each algorithm. It can be seen from Figure 5 that for $(m, n, M) = (16, 16, 8)$, SD performs the best since its SER curve coincides with the LB when the SNR is large. MOSEK also performs very well since the curve of MOSEK becomes parallel to the LB, i.e., a constant SER gap. However, this is not the case for PN-QP and GPM. For $(m, n, M) = (32, 16, 8)$, PN-QP, MOSEK, and SD are competitive, whereas SD is the best one.

(a) $(m, n, M) = (16, 16, 8)$.   (b) $(m, n, M) = (32, 16, 8)$.
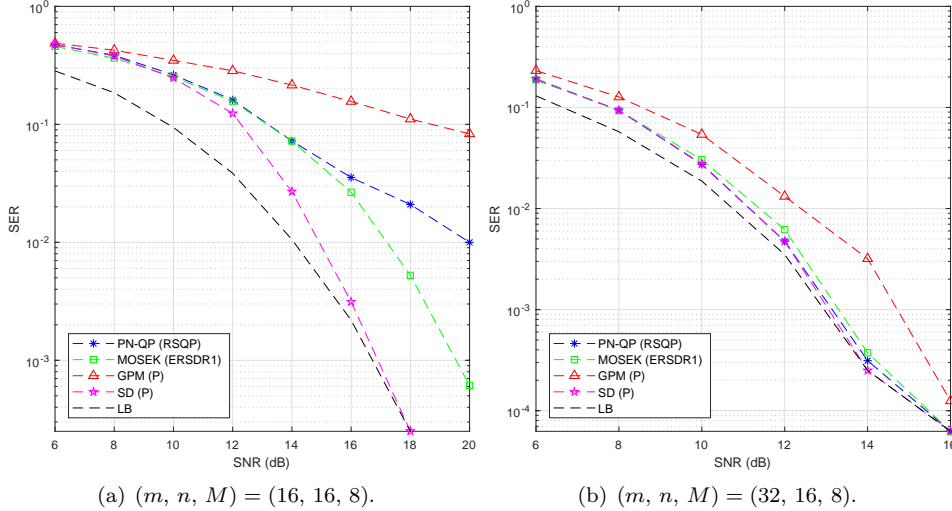
FIG. 5. *The* SER *performance of the four algorithms under different* SNR*s.*

Results on Problems with $M \geqslant 8$. We further compare the performance of PN-QP and GPM on $M = 16$. According to Table 5, as the SNR decreases, PN-QP provides a lower SER than GPM, implying that PN-QP achieves better detection performance. For instance, for $(m, n, M) = (512, 512, 16)$ with SNR $= 20$ dB, PN-QP returns a solution with SER $= 3.46\%$, whereas the SER of GPM is 33.67%. We also present more comparisons of the two algorithms in Figure 6. It can be observed

TABLE 3
Time and SER comparison of PN-QP, MOSEK, GPM, *and* SD *for solving MIMO detection problems with* $M = 2$.

(m, n, M)	SNR (dB)	Time(s)				SER(%)				
		PN-QP (RSQP)	MOSEK (ERSDR1)	GPM (P)	SD (P)	PN-QP (RSQP)	MOSEK (ERSDR1)	GPM (P)	SD (P)	LB
(32, 32, 2)	22	0.004	0.338	0.001	0.003	0.00	0.00	0.00	0.00	0.00
	20	0.004	0.329	0.001	0.003	0.00	0.00	0.00	0.00	0.00
	18	0.004	0.335	0.001	0.005	0.00	0.00	0.00	0.00	0.00
	16	0.005	0.338	0.001	0.012	0.00	0.00	0.00	0.00	0.00
	14	0.004	0.348	0.001	0.047	0.00	0.00	0.03	0.00	0.00
	12	0.004	0.369	0.001	0.110	0.00	0.00	0.16	0.00	0.00
(64, 64, 2)	22	0.008	0.456	0.001	0.007	0.00	0.00	0.00	0.00	0.00
	20	0.008	0.460	0.001	0.013	0.00	0.00	0.00	0.00	0.00
	18	0.008	0.469	0.001	0.035	0.00	0.00	0.00	0.00	0.00
	16	0.008	0.503	0.001	0.158	0.00	0.00	0.00	0.00	0.00
	14	0.008	0.534	0.001	1.392	0.00	0.00	0.00	0.00	0.00
	12	0.008	0.607	0.001	10.251	0.00	0.00	0.09	0.00	0.00
(128, 128, 2)	22	0.025	1.256	0.003	0.815	0.00	0.00	0.00	0.00	0.00
	20	0.025	1.319	0.003	12.260	0.00	0.00	0.00	0.00	0.00
	18	0.026	1.500	0.003	284.234	0.00	0.00	0.00	0.00	0.00
	16	0.030	1.723	0.004	2348.488	0.00	0.00	0.00	0.00	0.00
	14	0.030	1.783	0.004	—	0.00	0.00	0.00	—	0.00
	12	0.027	2.158	0.003	—	0.00	0.00	0.02	—	0.00
(256, 256, 2)	22	0.111	7.249	0.013	—	0.00	0.00	0.00	—	0.00
	20	0.112	7.408	0.013	—	0.00	0.00	0.00	—	0.00
	18	0.113	7.683	0.014	—	0.00	0.00	0.00	—	0.00
	16	0.115	8.502	0.014	—	0.00	0.00	0.00	—	0.00
	14	0.114	12.064	0.014	—	0.00	0.00	0.00	—	0.00
	12	0.117	15.292	0.015	—	0.00	0.00	0.01	—	0.00
(512, 512, 2)	22	0.198	58.730	0.092	—	0.00	0.00	0.00	—	0.00
	20	0.200	61.319	0.092	—	0.00	0.00	0.00	—	0.00
	18	0.201	64.035	0.092	—	0.00	0.00	0.00	—	0.00
	16	0.207	71.433	0.097	—	0.00	0.00	0.00	—	0.00
	14	0.216	113.457	0.106	—	0.00	0.00	0.00	—	0.00
	12	0.230	135.229	0.110	—	0.00	0.00	0.00	—	0.00

from Figure 6 that as the SNR increases, the SER curve of PN-QP tends to coincide with the LB. In contrast, there is a large gap between the SER curve of GPM with the LB, especially in the case where $m = n$.

Based on the above comparisons, our proposed algorithm for solving the MIMO detection problem is more efficient than existing algorithms for solving large-scale problems. Specifically, compared to MOSEK and SD, PN-QP is more efficient; compared to GPM, PN-QP achieves better detection performance. To conclude, PN-QP is demonstrated to be a competitive candidate for solving the large-scale MIMO detection problem.

6. Conclusions. In this paper, we proposed an efficient algorithm called PN-QP for solving the *large-scale* MIMO detection problem, motivated by the massive MIMO technology. The proposed algorithm is essentially a quadratic penalty method applied to solve an SQP relaxation, i.e., problem (RSQP), of the original problem. Two key features of the proposed algorithm, which make it particularly suitable to solve the large-scale problems, are: (i) it is based on the relaxation problem (RSQP), whose numbers of variables and constraints are significantly less than those of the SDRs; and (ii) our proposed algorithm is custom-designed to identify the support set of the optimal solution by judiciously exploiting the special structure of the problem,

TABLE 4
Time and SER comparison of PN-QP, MOSEK, GPM, *and* SD *for solving MIMO detection problems with* $M = 8$.

(m, n, M)	SNR (dB)	Time(s)				SER(%)				
		PN-QP (RSQP)	MOSEK (ERSDR1)	GPM (P)	SD (P)	PN-QP (RSQP)	MOSEK (ERSDR1)	GPM (P)	SD (P)	LB
(32, 32, 8)	22	0.044	0.362	0.001	0.208	0.00	0.00	1.31	0.00	0.00
	20	0.050	0.307	0.001	1.666	0.00	0.13	2.13	0.00	0.00
	18	0.051	0.340	0.001	19.923	0.00	0.25	4.06	0.00	0.00
	16	0.051	0.335	0.001	305.642	1.63	1.94	9.59	0.09	0.03
	14	0.060	0.403	0.002	2495.598	3.13	6.88	21.88	3.13	0.00
	12	0.067	0.351	0.001	—	13.94	14.81	23.47	—	3.69
(64, 64, 8)	22	0.126	0.513	0.002	769.366	0.00	0.00	0.22	0.00	0.00
	20	0.134	0.537	0.002	—	0.00	0.00	0.06	—	0.00
	18	0.132	0.507	0.003	—	0.00	0.39	3.09	—	0.00
	16	0.137	0.519	0.002	—	0.78	1.84	7.94	—	0.02
	14	0.150	0.510	0.003	—	5.92	7.28	15.53	—	0.66
	12	0.188	0.506	0.002	—	16.34	16.44	24.97	—	3.38
(128, 128, 8)	22	0.156	1.763	0.004	—	0.00	0.00	0.11	—	0.00
	20	0.164	1.730	0.005	—	0.00	0.00	0.95	—	0.00
	18	0.172	1.701	0.006	—	0.01	0.24	2.64	—	0.01
	16	0.189	1.675	0.010	—	0.38	1.88	6.73	—	0.06
	14	0.277	1.651	0.012	—	4.57	7.14	17.13	—	0.63
	12	0.307	1.555	0.008	—	15.89	15.47	23.63	—	3.27
(256, 256, 8)	22	0.342	10.558	0.018	—	0.00	0.00	0.00	—	0.00
	20	0.353	10.290	0.029	—	0.00	0.02	0.98	—	0.00
	18	0.379	10.153	0.041	—	0.00	0.32	2.02	—	0.00
	16	0.450	10.144	0.076	—	0.08	2.08	7.54	—	0.05
	14	0.587	9.799	0.097	—	4.11	7.45	15.85	—	0.75
	12	0.660	9.748	0.067	—	16.60	16.25	23.22	—	3.18
(512, 512, 8)	22	1.193	88.809	0.134	—	0.00	0.00	0.00	—	0.00
	20	1.303	91.365	0.177	—	0.00	0.01	0.09	—	0.00
	18	1.436	90.253	0.297	—	0.00	0.25	0.88	—	0.00
	16	1.825	88.638	0.929	—	0.08	2.23	6.68	—	0.06
	14	3.177	86.758	1.455	—	2.46	7.29	16.82	—	0.66
	12	3.482	85.408	0.765	—	16.41	16.15	22.97	—	3.17

instead of finding the solution itself, which thus substantially reduces the computational complexity of the proposed algorithm. The above two reasons lead to the better numerical performance of the proposed algorithm. In particular, our extensive simulation results show that our proposed algorithm compares favorably with the state-of-the-art algorithms (including SD and SDR based approaches) for solving the MIMO detection problem. When applied to solve large-scale problems, our proposed algorithm achieves significantly better detection performance than GPM.

Appendix A. Proof of Theorem 2.2. We need the following results to prove Theorem 2.2.

PROPOSITION A.1. *Let* Q, $\widehat{Q}$, *and* G *be defined in* (1.3), (2.2), *and* (2.6), *respectively. We have*
  (i) $q_{jj} = \sum_{k=1}^{m} |h_{kj}|^2$ *for* $j = 1, \ldots, n$;
  (ii) $\hat{q}_{kj} = \hat{q}_{jk} = \hat{q}_{(n+j)(n+k)} = \hat{q}_{(n+k)(n+j)}$, $\hat{q}_{(n+j)k} = \hat{q}_{k(n+j)} = -\hat{q}_{(n+k)j} = -\hat{q}_{j(n+k)}$, $\hat{q}_{jj} = \hat{q}_{(n+j)(n+j)} = q_{jj}$, *and* $\hat{q}_{(n+j)j} = \hat{q}_{j(n+j)} = 0$ *for* j, $k = 1, \ldots, n$;
  (iii) $S_{jk} = \hat{q}_{jk}\alpha\alpha^\top + \hat{q}_{(n+j)k}\beta\alpha^\top + \hat{q}_{j(n+k)}\alpha\beta^\top + \hat{q}_{(n+j)(n+k)}\beta\beta^\top$ *and* $S_{jk} = S_{kj}^\top$ *for* j, $k = 1, \ldots, n$; *and*
  (iv) $\text{diag}(S_{jj}) = q_{jj}\boldsymbol{e}$ *for* $j = 1, \ldots, n$.

Time and SER comparison of PN-QP *and* GPM *for solving MIMO detection problems with* $M = 16$.

(m, n, M)	SNR (dB)	Time(s)		SER(%)		
		PN-QP (RSQP)	GPM (P)	PN-QP (RSQP)	GPM (P)	LB
(256, 128, 16)	30	0.203	0.005	0.00	0.00	0.00
	25	0.210	0.004	0.00	0.00	0.00
	20	0.235	0.004	0.02	0.02	0.01
	15	0.395	0.007	4.49	5.23	2.66
(128, 128, 16)	30	0.386	0.007	0.00	2.81	0.00
	25	0.406	0.018	0.00	17.41	0.00
	20	0.549	0.023	8.42	34.49	0.74
	15	0.572	0.023	33.09	46.80	12.58
(512, 256, 16)	30	0.432	0.023	0.00	0.00	0.00
	25	0.466	0.024	0.00	0.00	0.00
	20	0.518	0.026	0.02	0.02	0.01
	15	0.774	0.046	4.27	4.61	2.80
(256, 256, 16)	30	0.862	0.047	0.00	1.60	0.00
	25	0.986	0.179	0.00	18.45	0.00
	20	1.513	0.186	4.80	34.61	0.65
	15	1.603	0.190	34.89	48.95	11.55
(1024, 512, 16)	30	1.355	0.156	0.00	0.00	0.00
	25	1.416	0.155	0.00	0.00	0.00
	20	1.722	0.191	0.01	0.01	0.01
	15	3.273	0.397	4.48	4.76	2.85
(512, 512, 16)	30	2.640	0.380	0.00	1.28	0.00
	25	2.980	2.298	0.00	19.43	0.00
	20	6.337	2.337	3.46	33.67	0.59
	15	6.718	2.390	33.88	48.36	12.09

Proof. (i) By the definition of Q in (1.3), there is

$$q_{jj} = \sum_{k=1}^{m} h_{kj}^{\dagger} h_{kj} = \sum_{k=1}^{m} |h_{kj}|^2,$$

which gives (i).

(ii) By the definitions of $\widehat{Q}$ in (2.2) and Q in (1.3), the first three results in (ii) hold naturally. Note that due to (i), there is $\mathrm{diag}(\mathrm{Im}(Q)) = \mathbf{0}$, implying the fourth result in (ii).

(iii) Let P have the partition as $P = \begin{bmatrix} P_1 & \cdots & P_n \end{bmatrix}$, where $P_j \in \mathbb{R}^{2n \times M}$ takes the following form:

(A.1)
$$P_j^{\top} = [\mathbf{0} \quad \cdots \quad \mathbf{0} \quad \underset{\underset{\text{the } j\text{-th block}}{\uparrow}}{\alpha} \quad \mathbf{0} \quad \cdots \quad \mathbf{0} \quad \underset{\underset{\text{the } (n+j)\text{-th block}}{\uparrow}}{\beta} \quad \mathbf{0} \quad \cdots \quad \mathbf{0}]^{\top}.$$

By the definition and partition of G in (2.6) and (2.8), we have

$$\begin{aligned} G = P^{\top} \widehat{Q} P &= \begin{bmatrix} P_1 & \cdots & P_n \end{bmatrix}^{\top} \widehat{Q} \begin{bmatrix} P_1 & \cdots & P_n \end{bmatrix} \\ &= \begin{bmatrix} P_1^{\top} \widehat{Q} P_1 & \cdots & P_1^{\top} \widehat{Q} P_n \\ \vdots & \ddots & \vdots \\ P_n^{\top} \widehat{Q} P_1 & \cdots & P_n^{\top} \widehat{Q} P_n \end{bmatrix} = \begin{bmatrix} S_{11} & \cdots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{n1} & \cdots & S_{nn} \end{bmatrix}, \end{aligned}$$
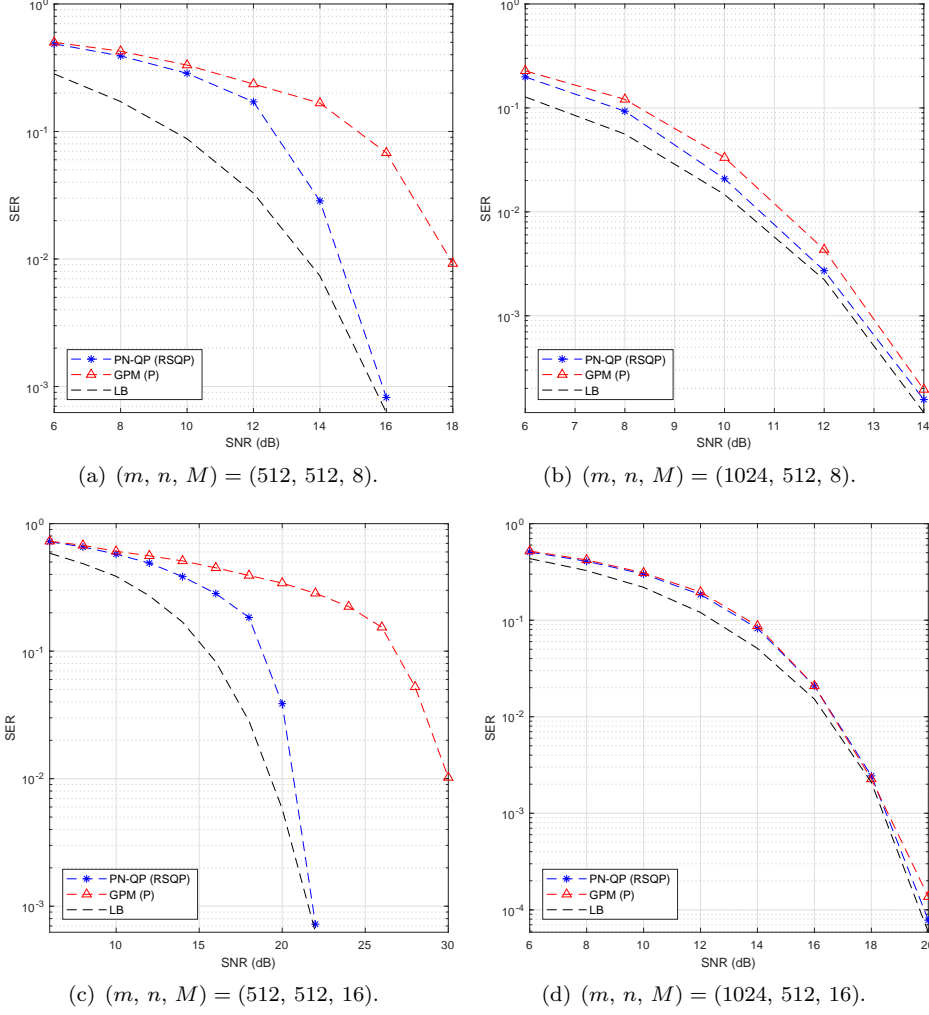
(a) $(m, n, M) = (512, 512, 8)$.

(b) $(m, n, M) = (1024, 512, 8)$.

(c) $(m, n, M) = (512, 512, 16)$.

(d) $(m, n, M) = (1024, 512, 16)$.

FIG. 6. *The* SER *performance of* PN-QP *and* GPM *under different* SNR*s.*

which, together with (2.2) and (A.1), further implies

$$
\begin{aligned}
S_{jk} &= P_j^\top \widehat{Q} P_k \\
&= \begin{bmatrix} \hat{q}_{j1}\alpha + \hat{q}_{(n+j)1}\beta & \cdots & \hat{q}_{j(2n)}\alpha + \hat{q}_{(n+j)(2n)}\beta \end{bmatrix} P_k \\
&= \hat{q}_{jk}\alpha\alpha^\top + \hat{q}_{(n+j)k}\beta\alpha^\top + \hat{q}_{j(n+k)}\alpha\beta^\top + \hat{q}_{(n+j)(n+k)}\beta\beta^\top.
\end{aligned}
$$

The proof of the first result in (iii) is finished. With the first result in (iii), as well as the first two results in (ii), there is

$$
\begin{aligned}
S_{jk} &= \hat{q}_{jk}\alpha\alpha^\top + \hat{q}_{(n+j)k}\beta\alpha^\top + \hat{q}_{j(n+k)}\alpha\beta^\top + \hat{q}_{(n+j)(n+k)}\beta\beta^\top \\
&= \hat{q}_{kj}\alpha\alpha^\top + \hat{q}_{k(n+j)}\beta\alpha^\top + \hat{q}_{(n+k)j}\alpha\beta^\top + \hat{q}_{(n+k)(n+j)}\beta\beta^\top \\
&= (\hat{q}_{kj}\alpha\alpha^\top + \hat{q}_{k(n+j)}\alpha\beta^\top + \hat{q}_{(n+k)j}\beta\alpha^\top + \hat{q}_{(n+k)(n+j)}\beta\beta^\top)^\top \\
&= S_{kj}^\top.
\end{aligned}
$$

We get the second result in (iii).

(iv) Due to (iii), there is

$$
\begin{aligned}
S_{jj} &= \hat{q}_{jj}\alpha\alpha^\top + \hat{q}_{(n+j)j}\beta\alpha^\top + \hat{q}_{j(n+j)}\alpha\beta^\top + \hat{q}_{(n+j)(n+j)}\beta\beta^\top \\
&= q_{jj}(\alpha\alpha^\top + \beta\beta^\top).
\end{aligned}
$$

Recall the definitions of α, β in (2.4) and θ_k in (2.1). The k-th diagonal entry of S_{jj} takes the following form:

$$
q_{jj}(\cos^2(\theta_k) + \sin^2(\theta_k)) = q_{jj}, \ \ k = 1,\ldots,M.
$$

Therefore, $\mathrm{diag}(S_{jj}) = q_{jj}\boldsymbol{e}$. The proof is finished.   □

PROPOSITION A.2. *Under the constraints in problem* (SQP2), *there is*

(A.2)
$$
\sum_{j=1}^{n} \bar{t}_j^\top S_{jj}\bar{t}_j = \|H\|_2^2.
$$

Proof. Under the constraints in problem (SQP2), for each $\bar{t}_j$, there exists $l_j \in \{1,\ldots,M\}$, such that $\bar{t}_j = e_{l_j}$, where e_{l_j} denotes the l_j-th column in the identity matrix $I \in \mathbb{R}^{M \times M}$. As a result, with (i) and (iv) in Proposition A.1, there is

$$
\sum_{j=1}^{n} \bar{t}_j^\top S_{jj}\bar{t}_j = \sum_{j=1}^{n} e_{l_j}^\top S_{jj}e_{l_j} = \sum_{j=1}^{n} q_{jj} = \sum_{j=1}^{n}\sum_{k=1}^{m} |h_{kj}|^2 = \|H\|_2^2,
$$

which gives (A.2). The proof is finished.   □

Now we are ready to prove Theorem 2.2.

Proof. Using Proposition A.2, we have

$$
h(t) - f(t) = \sum_{j=1}^{n} \bar{t}_j^\top S_{jj}\bar{t}_j = \|H\|_2^2.
$$

This, together with the fact that the constraints of problems (SQP1) and (SQP2) are the same, implies that problems (SQP1) and (SQP2) are equivalent.   □

REFERENCES

[1] M. A. ALBREEM, M. JUNTTI, AND S. SHAHABUDDIN, *Massive MIMO detection techniques: A survey*, IEEE Communications Surveys and Tutorials, 21 (2019), pp. 3109–3132.

[2] D. P. BERTSEKAS, *Projected Newton methods for optimization problems with simple constraints*, SIAM Journal on Control and Optimization, 20 (1982), pp. 221–246.

[3] S. A. BUSARI, K. M. S. HUQ, S. MUMTAZ, L. DAI, AND J. RODRIGUEZ, *Millimeter-wave massive MIMO communication for future wireless systems: A survey*, IEEE Communications Surveys and Tutorials, 20 (2018), pp. 836–869.

[4] A. M. CHAN AND I. LEE, *A new reduced-complexity sphere decoder for multiple antenna systems*, in Proceedings of IEEE International Conference on Communications, New York, 2002, pp. 460–464.

[5] Cisco, *Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022*, 2019, https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html (accessed 2019-12-10).

[6] C. Cui, Q.-N. Li, L. Qi, and H. Yan, *A quadratic penalty method for hypergraph matching*, Journal of Global Optimization, 70 (2018), pp. 237–259.

[7] O. Damen, A. Chkeif, and J.-C. Belfiore, *Lattice code decoder for space-time codes*, IEEE Communications Letters, 4 (2000), pp. 161–163.

[8] J. Jaldén, *Detection for multiple input multiple output channels*, PhD thesis, KTH Royal Institute of Technology, 2006.

[9] J. Jaldén and B. Ottersten, *On the complexity of sphere decoding in digital communications*, IEEE Transactions on Signal Processing, 53 (2005), pp. 1474–1484.

[10] J. Jaldén, D. Seethaler, and G. Matz, *Worst- and average-case complexity of LLL lattice reduction in MIMO wireless systems*, in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, 2008, pp. 2685–2688.

[11] R. Jiang, Y.-F. Liu, C. Bao, and B. Jiang, *Tightness and equivalence of semidefinite relaxations for MIMO detection*, 2021, https://arxiv.org/abs/2102.04586.

[12] A. K. Lenstra, H. W. Lenstra, and L. Lovász, *Factoring polynomials with rational coefficients*, Mathematische Annalen, 261 (1982), pp. 515–534.

[13] Q.-N. Li and D.-H. Li, *A projected semismooth Newton method for problems of calibrating least squares covariance matrix*, Operations Research Letters, 39 (2011), pp. 103–108.

[14] H. Liu, M.-C. Yue, A. M.-C. So, and W.-K. Ma, *A discrete first-order method for large-scale MIMO detection with provable guarantees*, in Proceedings of IEEE Workshop on Signal Processing Advances in Wireless Communications, Sapporo, 2017, pp. 669–673.

[15] Y.-F. Liu, Z. Xu, and C. Lu, *On the equivalence of semidifinite relaxations for MIMO detection with general constellations*, in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Brighton, 2019, pp. 4549–4553.

[16] M. P. Lotter and P. V. Rooyen, *Space division multiple access for cellular CDMA*, in Proceedings of IEEE International Symposium on Spread Spectrum Techniques and Applications, Sun City, 1998, pp. 959–964.

[17] C. Lu, Y.-F. Liu, W.-Q. Zhang, and S. Zhang, *Tightness of a new and enhanced semidefinite relaxation for MIMO detection*, SIAM Journal on Optimization, 29 (2019), pp. 719–742.

[18] C. Lu, Y.-F. Liu, and J. Zhou, *An efficient global algorithm for nonconvex complex quadratic problems with applications in wireless communications*, in Proceedings of IEEE/CIC International Conference on Communications in China, Qingdao, 2017, pp. 1–5.

[19] C. Lu, Y.-F. Liu, and J. Zhou, *An enhanced SDR based global algorithm for nonconvex complex quadratic programs with signal processing applications*, IEEE Open Journal of Signal Processing, 1 (2020), pp. 120–134.

[20] Z.-Q. Luo, X. Luo, and M. Kisialiou, *An efficient quasi-maximum likelihood decoder for PSK signals*, in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, 2003, pp. 561–564.

[21] W.-K. Ma, P.-C. Ching, and Z. Ding, *Semidefinite relaxation based multiuser detection for M-ary PSK multiuser systems*, IEEE Transactions on Signal Processing, 52 (2004), pp. 2862–2872.

[22] A. Mobasher, M. Taherzadeh, R. Sotirov, and A. K. Khandani, *A near-maximum-likelihood decoding algorithm for MIMO systems based on semi-definite programming*, IEEE Transactions on Information Theory, 53 (2007), pp. 3869–3886.

[23] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, *Hybrid beamforming for massive MIMO: A survey*, IEEE Communications Magazine, 55 (2017), pp. 134–141.

[24] MOSEK ApS, *MOSEK optimization toolbox for MATLAB Release 9.2.29*, 2020, https://docs.mosek.com/9.2/toolbox.pdf.

[25] J. Nocedal and S. Wright, *Numerical Optimization*, Springer-Verlag, New York, 2006.

[26] M. Pohst, *On the computation of lattice vectors of minimal length, successive minima and reduced bases with applications*, ACM Sigsam Bulletin, 15 (1981), pp. 37–44.

[27] A. M.-C. So, *Probabilistic analysis of the semidefinite relaxation detector in digital communications*, in Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, Austin, 2010, pp. 698–711.

[28] D. Sun, K.-C. Toh, Y. Yuan, and X.-Y. Zhao, *SDPNAL+: A MATLAB software for semidefinite programming with bound constraints (version 1.0)*, Optimization Methods and Software, 35 (2020), pp. 87–115.

[29] P. H. Tan and L. K. Rasmussen, *The application of semidefinite programming for detection in CDMA*, IEEE Journal on Selected Areas in Communications, 19 (2001), pp. 1442–1449.

[30] S. VERDÚ, *Computational complexity of optimum multiuser detection*, Algorithmica, 4 (1989), pp. 303–312.

[31] H.-T. WAI, W.-K. MA, AND A. M.-C. SO, *Cheap semidefinite relaxation MIMO detection using row-by-row block coordinate descent*, in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, 2011, pp. 3256–3259.

[32] Z. XIE, C. K. RUSHFORTH, R. T. SHORT, AND T. K. MOON, *Joint signal detection and parameter estimation in multiuser communications*, IEEE Transactions on Communications, 41 (1993), pp. 1208–1216.

[33] L. YANG, D. SUN, AND K.-C. TOH, *SDPNAL+: A majorized semismooth Newton-CG augmented Lagrangian method for semidefinite programming with nonnegative constraints*, Mathematical Programming Computation, 7 (2015), pp. 331–366.

[34] S. YANG AND L. HANZO, *Fifty years of MIMO detection: The road to large-scale MIMOs*, IEEE Communications Surveys and Tutorials, 17 (2015), pp. 1941–1988.

[35] H. YAO, *Efficient signal, code, and receiver designs for MIMO communication systems*, PhD thesis, Massachusetts Institute of Technology, 2003.

[36] X.-Y. ZHAO, *A semismooth Newton-CG augmented Lagrangian method for large scale linear and convex quadratic SDPs*, PhD thesis, National University of Singapore, 2009.

[37] X.-Y. ZHAO, D. SUN, AND K.-C. TOH, *A Newton-CG augmented Lagrangian method for semidefinite programming*, SIAM Journal on Optimization, 20 (2010), pp. 1737–1765.

[38] Q. ZHOU AND X. MA, *Element-based lattice reduction algorithms for large MIMO detection*, IEEE Journal on Selected Areas in Communications, 31 (2013), pp. 274–286.