High-order-modulation large MIMO detector based on physics-inspired methods

Qing-Guo Zeng,[1] Xiao-Peng Cui,[2] Xian-Zhe Tao,[1] Jia-Qi Hu,[3] Shi-Jie Pan,[4] Wei E.I. Sha,[5,*] and Man-Hong Yung[1,6,7,8,†]

[1] *Shenzhen Institute for Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China*
[2] *Department of Physics, Fudan University, Shanghai 200433, China*
[3] *Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University, Beijing 100084, China*
[4] *State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[5] *Key Laboratory of Micro-Nano Electronic Devices and Smart Systems of Zhejiang Province, College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, Zhejiang, China*
[6] *International Quantum Academy, Shenzhen 518048, China*
[7] *Guangdong Provincial Key Laboratory of Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China*
[8] *Shenzhen Key Laboratory of Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China*

(Received 24 February 2025; revised 10 July 2025; accepted 18 August 2025; published 10 September 2025)

Applying quantum annealing or current quantum- and physics-inspired algorithms for multiple-input multiple-output (MIMO) detection always abandon the direct Gray-coded bit-to-symbol mapping in order to obtain Ising form, leading to inconsistency errors. This often results in slow convergence rates and error floor, particularly with high-order-modulations. We propose HOPbit, an alternative MIMO detector designed to address this issue by transforming the MIMO detection problem into a higher-order Ising problem while maintaining Gray-coded bit-to-symbol mapping. The method then employs the simulated probabilistic bits (p-bits) algorithm to directly solve higher-order Ising problem without degradation. This innovative strategy enables HOPbit to achieve rapid convergence and attain near-optimal maximum-likelihood performance in most scenarios, even those involving high-order-modulations. The experiments show that HOPbit surpasses ParaMax by several orders of magnitude in terms of bit error rate (BER) in the context of 12-user massive and large MIMO systems even with less computing resources. In addition, HOPbit achieves lower BER rates compared to other traditional detectors.

DOI: 10.1103/zmgv-m2ql

I. INTRODUCTION

The past decade has witnessed an exponential growth of global mobile data traffic along with the tremendous rise in the number of mobile users and data-intensive content [1]. This urges the fast development of wireless technologies. Multiple input multiple output (MIMO) is a key technique to satisfy the massive demand for wireless data traffic in current and future generation of wireless systems in which multiple antennas at both transmitter and receiver are combined to improve the capacity of a radio link based on multipath propagation.

However, it is challenging for the receiver to reproduce the transmitted symbols based on channel state information (CSI) and the received signal with the presence of noise and interference, which is referred to as MIMO detection problem [2]. The exact algorithm, maximum-likelihood detector (MLD), has been proven to be NP-hard [3]. Linear detectors such as zero forcing (ZF) and minimum mean squared error (MMSE) perform nearly to MLD in massive MIMO systems at the cost of a large number of receiver antennas to support relatively fewer users [4,5]. That is, their detection accuracy becomes poor in large MIMO systems where the number of receiver antennas is the same as that of user antennas [6]. Despite decades of research, practical methods with near-optimal detection accuracy for larger MIMO systems remain elusive [2].

In recent years, some researchers have begun to investigate the application of quantum annealing to MIMO detection [7]. QuAMax, a quantum-annealing-based MIMO detector, uses a linear transformation function to reduce

*Contact author: weisha@zju.edu.cn
†Contact author: yung@sustech.edu.cn

the MIMO detection problem into an Ising problem and then exploits a quantum annealer to solve it [8]. Tabi *et al.* improve QuAMax by combining the single qubit correction postprocessing technique and applying it to larger MIMO systems and higher-order-modulation (64-QAM) cases [9]. As a result, correction-enhanced QuAMax detector achieves a lower bit error rate (BER). The quantum-annealing-based method is promising and has the potential to achieve polynomial or exponential speedups with extremely near-optimal performance in large MIMO systems [10,11]. Considering the sensitivity of quantum devices to noise, a hybrid quantum-classical detector has been proposed in Ref. [12] to combine the advantages of both hardware. However, in the noisy intermediate-scale quantum (NISQ) era, these are not optimistic for practical applications due to the limited qubit connectivity and physical noises [13]. Fortunately, some Ising machines and quantum- and physical-inspired algorithms, such as the coherent Ising machine and its simulated algorithms [14–17], simulated bifurcation machine [18,19], etc., have demonstrated their superiority over the current quantum computing devices in specific problems [20]. By adopting the same transformation, these methods and solvers can also solve the MIMO detection problem. The detector, ParaMax, employs the parallel tempering algorithm to solve Ising problems. One key advantage of this approach is its capacity to generate a soft output based on confidence information that is useful for iterative MIMO detection processes [21]. Thus, the author of Ref. [21] also proposes 2R-ParaMax, a MIMO detector that uses the parallel tempering algorithm in two rounds. It fixes some spins based on spinwise detection confidence and updates the rest spins in the second round. Sing *et al.* begin to investigate the disadvantage of straightforward ML-to-Ising transformation and find out an error floor problem [22]. As a result, they add a regularized term to mitigate this issue. In order to reduce MLD to Ising form, all of the aforementioned methods have to give up two-dimensional Gray-code mapping, resulting in the inconsistency error. Inconsistency error refers to the mismatch between the geometric distance in the signal constellation space and the Hamming distance in the bitstring representation when using binary coding in MIMO detection problems. In binary coding, adjacent modulation symbols may correspond to bitstrings with inconsistent or large Hamming distances, even if they are close in Euclidean space. This inconsistency leads to a distortion in the objective function (e.g., Ising or higher-order Ising energy), causing optimization algorithms to misinterpret the true proximity between symbols. Take a simple one-dimensional 4-PAM modulation (part of a 16-QAM scheme) as an example. The four symbols (bitstrings) are -3 (00), -1 (01), $+1$ (10), and $+3$ (11). The Hamming distance between 00 (for -3) and 01 (for -1) is 1 while that between 01 (for -1) and 10 (for $+1$), the Hamming distance is 2. In addition, Norimoto *et al.* point out that the methods based on the linear transformation function need extra encoding and decoding steps to transform the solution back to Gray-coded form [23]. As a result, they directly use Gray-coded bit-to-symbol mapping to reduce the MIMO detection problem into the higher-order Ising problem, and then solve it by utilizing a quantum algorithm, grover adaptive search [24,25]. However, Grover's algorithm remains more of a theoretical milestone than a practical tool since its meaningful acceleration requires the use of a fault-tolerant quantum computer with a large number of qubits, which are still far away [26].

Hence, we propose HOPbit, a detector that also transforms the MIMO detection problem to the higher-order Ising problem while maintaining the Gray-coded bit-to-symbol mapping to mitigate error floor. Current quantum annealing or quantum- and physics-inspired methods for solving the higher-order Ising problem, where interactions involve more than two spins, typically introduce auxiliary variables to transform it into a standard pairwise Ising form. However, this transformation leads to a proportional increases in the number of auxiliary bits and interaction terms, which grows with the number of antennas and significantly increases the problem complexity. [27–31]. In this work, we implement a simulator of high-order p-bits to solve it directly. Besides, we apply precomputation of the input signal of p-bit for further acceleration [32].

The rest of this paper is organized as follows. The MIMO detection problem is described in the next section. In Sec. III, we presented each component and detailed process of the HOPbit. Its performances are investigated and compared with the Ising solver and other conventional detectors in Sec. IV. Finally, we conclude and discuss the future work in the last section.

II. MIMO DETECTION PROBLEM

We assume that in MIMO detection, N_t users each with a single antenna transmit the symbols $\mathbf{x} = [x_1, x_2, \ldots, x_{N_t}]^\top \in \mathbb{C}^{N_t}$ from constellation Ω and then the symbols is received by base station with N_r antennas. The received symbols can be presented $\mathbf{y} = [y_1, y_2, \ldots, y_{N_r}]^\top \in \mathbb{C}^{N_r} = \mathbf{y}^R + j\,\mathbf{y}^I$. The transmission characteristic between N_t user antennas and N_r receiver antennas is summarized in a channel matrix $\mathbf{H} \in \mathbb{C}^{N_r \times N_t} = \mathbf{H}^R + j\,\mathbf{H}^I$. Finally, this process can be formulated as

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

where $\mathbf{n} \in \mathbb{C}^{N_r}$ indicates the additive white Gaussian noise. The MIMO detection problem is to reproduce the transmission symbols when given the channel matrix and received symbols as far as possible.

With the presence of additive noise, the ML detector is to seek the optimal solution that minimizes the detection error as follows:

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \arg\min_{\mathbf{x}\in\Omega^{N_t}} ||\mathbf{y} - \mathbf{H}\mathbf{x}||^2. \tag{2}$$

It means that the detector searches for the solution in an exponentially large space with the size of the constellation and the number of user antennas even equipped with sphere decoder reduction in search space [33,34].

III. HOPBIT

It is impractical to apply ML detector for massive and large MIMO systems. In this section, we propose an alternative detector, HOPbit, that reduces ML detection to the higher-order Ising problem, thereby eliminating the inconsistency error. Firstly, a standard higher-order Ising problem form and the transformation of ML into the higher-order Ising problem are given. Then, we introduce a simulated p-bits solver. Finally, we describe the complete pipeline of HOPbit.

A. ML-to-higher-order-Ising transformation

A higher-order Ising problem is the task of minimizing a pseudo-Boolean function (Hamiltonian) that includes interaction terms involving more than two Ising spins [35,36]:

$$\mathcal{H}(\mathbf{q}) = \sum_{S\subseteq 1,\ldots,N,|S|\leq 1} J_S \prod_{i\in S} q_i, \tag{3}$$

where N is the number of binary variables, S is a nonempty subset of indices $1, 2, \ldots, N$, $J_S \in \mathbb{R}$ is the real coefficient associated with subset S.

The most critical thing for reducing the ML problem into higher-order Ising form is to find a bit-to-symbol function that represents the transmission symbols using the bit string. As mentioned before, this function directly influences the final detection accuracy. In this work, we use the bit-to-symbol mapper proposed in 5G NR standard [37]. For phase-shift-keying (PSK) modulation, this mapper is equivalent to ML-to-Ising reduction in Ref. [8] but is quite different in quadrature amplitude modulation (QAM).

For $|\Omega|$-QAM, the number of bits to represent each symbol is $m = \log_2 |\Omega|$, then the total number of bits for the $N_r \times N_t$ MIMO system is $N_r \times m$. The ML-to-Ising transformation is

$$\begin{aligned} x_i(\{q_{mi-a}\}_{a=1}^{m}) &= x_i^R + jx_i^I \\ &= \phi(\{q_{mi-a}\}_{a=1}^{m/2}) + j\,\phi(\{q_{mi-a}\}_{a=m/2+1}^{m}) \\ &= \underbrace{\sum_{a=1}^{m/2} 2^{\frac{m}{2}+1-a} q_{mi-a}}_{\text{real part}} + j \underbrace{\sum_{a=1}^{m/2} 2^{\frac{m}{2}+1-a} q_{mi-\frac{m}{2}-a}}_{\text{image part}}, \end{aligned} \tag{4}$$

while the Gray-coded bit-to-symbol function is

$$\begin{aligned} x_i(\{q_{mi-a}\}_{a=1}^{m}) &= x_i^R + jx_i^I \\ &= \psi(\{q_{mi-a}\}_{a=1}^{m/2}) + j\,\psi(\{q_{mi-a}\}_{a=m/2+1}^{m}) \\ &= \underbrace{\sum_{a=1}^{m/2} 2^{\frac{m}{2}-a} \prod_{b=1}^{a} q_{mi-b}}_{\text{real part}} + j \underbrace{\sum_{a=1}^{m/2} 2^{\frac{m}{2}-a} \prod_{b=1}^{a} q_{mi-\frac{m}{2}-b}}_{\text{image part}}, \\ & i = 1, \ldots, N_r. \end{aligned} \tag{5}$$

It is worth noting that distinct bit-to-symbol functions can yield identical constellation while producing divergent loss landscapes, consequently leading to varying levels of optimization complexity. The ML-to-Ising reduction method has shown promising performance in large MIMO systems in low-order-modulation by using quantum annealing or quantum-inspired optimization methods [7,12,21,22] and thus, we will only discuss and compare the different bit-to-symbol functions for QAM in this paper.

Substituting bit-to-symbol function [Eq. (5)] into the objective function of ML [Eq. (2)], we can get the real-valued objective function in higher-order Ising form [Eq. (A1) in Appendix A].

B. Solver

When using a quantum annealer or Ising machines to solve the higher-order Ising problem, the problems always are transformed into the Ising problem by introducing auxiliary spin variables. However, p-bits is able to solve it directly [38]. Next, we will briefly introduce the mechanism of p-bits and the simulated algorithm of it.

1. P-bits

P-bit is an unconventional computation scheme that utilizes probabilistic bits, which is intermediate between classical bit and quantum bits (qubit). Unlike classical bits, which are strictly 0 or 1, or qubits, which exist in quantum superposition, p-bits fluctuate stochastically between 0 and 1 according to tunable probability distributions. And it has been demonstrated that p-bits can efficiently solve combinatorial optimization problems by accelerating randomized algorithms through their inherent noise and parallelism [39].

Every individual bit in p-bits is built with a spintronic device named magnetic tunnel junction (MTJ). According to tunneling magnetoresistance effect, MTJs exhibit two different resistance states of high and low due to the relative direction of the two ferromagnetic layers, which can be used to represent binary variables naturally. MTJs are unstable; they are always fluctuating between the two states due to their intrinsic randomness. According to Arrhenius' law, the average time interval between spontaneous switches is about

$$\tau = \tau_0 \exp(\Delta E/k_B T), \quad \text{with } \tau_0 \approx 1 \text{ ns}, \tag{6}$$

where ΔE is the energy barrier, k_B is Boltzmann's constant, and T is the temperature. By tuning the energy barrier via voltage or current bias, the switching probability can be precisely controlled.

The intrinsic thermal noise in MTJs enables p-bits to produce inherently stochastic outputs, making them ideal for implementing hardware-friendly versions of algorithms like simulated annealing, Boltzmann machines, and other stochastic local search heuristics. Furthermore, p-bit networks are inherently suitable for asynchronous and parallel operation, which offers significant computational speedups when deployed in hardware.

2. Simulated p-bits algorithm

In order to understand and predict the performance of p-bits on MIMO detection problems, we implemented a simulator of high-order p-bits [40], namely simulated p-bit algorithm, based on the mechanism of p-bits.

By analyzing of stochastic processes of p-bits, we can calculate the switch probabilities of each bit. The only parameter that matters is ΔE so that the switch probabilities can be finely controlled by its value. In p-circuits, ΔE is tuned by the current signal automatically. In other words, a p-bit can be described as a binary stochastic neuron in mathematics,

$$q_i = \text{sgn}[\tanh(\beta I_i) - r], \tag{7}$$

where $r \sim \mathcal{U}(-1, 1)$ represents a random number. I_i is the input signal, a continuous value, and m_i is the output signal with binary value. β is the inverse temperature, which increases during the annealing. The circuit that calculates the input signal encodes a special problem. For a higher-order Ising problem, I_i should be

$$I_i = -\frac{\partial \mathcal{H}(\mathbf{q})}{\partial q_i}. \tag{8}$$

where $\mathcal{H}(\mathbf{q})$ is a Hamiltonian of higher-order Ising problem. Although evaluating this gradient appears costly in general, the specific structure of the MIMO MLD problem allows for an efficient implementation. We first convert the complex-valued channel model into an equivalent real-valued representation,

$$\tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{H}^R & -\mathbf{H}^I \\ \mathbf{H}^I & \mathbf{H}^R \end{bmatrix}, \quad \tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y}^R \\ \mathbf{y}^I \end{bmatrix}, \quad \tilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x}^R \\ \mathbf{x}^I \end{bmatrix}, \tag{9}$$

and let $\mathbf{A} = \mathbf{H}\mathbf{H}^\top$. Then, when the jth spin variable q_i in ith symbol x_i flips (the new spin variable and new symbol denote by $\tilde{q}_j$ and $\tilde{x}_i$, respectively), the changes of Hamiltonian is

$$\begin{aligned} I_{i_j} &= ||\tilde{\mathbf{H}}\psi(\mathbf{q}) - \tilde{\mathbf{y}}||^2 - ||\tilde{\mathbf{H}}\psi(\hat{\mathbf{q}}) - \tilde{\mathbf{y}}||^2 \\ &= \Delta x_i^j (2(\mathbf{A}_{i\cdot}\mathbf{q} - \tilde{\mathbf{y}}) + \Delta x_i^j A_{ii}), \end{aligned} \tag{10}$$

where $\mathbf{A}_{i\cdot}$ denotes the ith row of $\mathbf{A}$, and the scalar Δx_i^j is the symbol difference resulting from the bit flip. It can be computed as

$$\begin{aligned} \Delta x_i^j &= \psi(\{q_{mi-1}\}_{a=1}^{m/2}) - \psi(\{\tilde{q}_{mi-1}\}_{a=1}^{m/2}) \\ &= \sum_{a=j}^{m/2} 2^{\frac{m}{2}-a+1} \prod_{b=1}^{a} q_{mi-b} \\ &= -2q_{mi-j} \cdot ((\cdots((q_{mi-m} + 2) \cdot q_{mi-m+1} + 2^2) \cdots \\ &\quad q_{mi-j-1} + 2^{m-j})q_{mi-j+1} + 2^{m-j+2}) \cdots q_{mi-\frac{m}{2}} \end{aligned} \tag{11}$$

Although this expression appears complex due to the recursive binary-to-symbol mapping (e.g., Gray coding), it is important to note that only a small, constant subset of terms is affected by each individual spin flip. Specifically, the computation of Δx_i^j involves at most $m/2$ additions and multiplications. Furthermore, if we precompute and store the mapping between all $2^{m/2}$ possible bit strings and their corresponding real-valued PAM symbols, the evaluation of Δx_i^j reduces to a single subtraction between two stored values. This small lookup table requires minimal memory and can be efficiently cached. As a result, the update of the input signal I_{i_j} for each spin flip—both in QOPbit

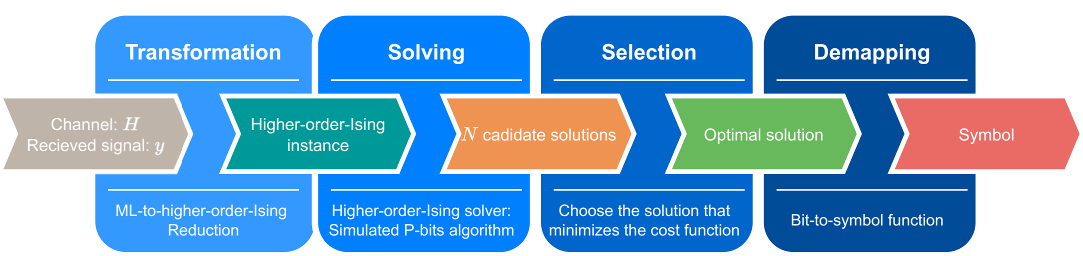

FIG. 1. The flowchart of HOPbit.

ALGORITHM 1. Simulated P-bit.

```
Input: Initial replicas with random states
Output: Final states of the P-bits
β ← β_0;
for i = 1 to max_iters do
    for each p-bit do
        Compute input signal I_i;
        Compute the output state
          q_i = sgn[tanh(βI_i) − r];
    β ← β + Δ_β;
```

and HOPbit—can be performed with a time complexity of $\mathcal{O}(N_t)$, despite the high-order nature of the underlying Ising formulation.

When we run p-bits to solve a combinatorial optimization problem, actually it is going through an annealing process, which helps the algorithm to escape from local minima and converge to the solution with lower energy. Although both simulated p-bit and simulated annealing are thermally inspired, they fundamentally differ in how spin updates are performed [41]. In simulated annealing, a trial flip is proposed and accepted or rejected based on the Metropolis criterion, where the acceptance probability depends on the energy difference. In contrast, the simulated p-bit algorithm directly samples the new spin state from a stochastic activation function, which avoids the explicit accept or reject step and enables continuous probabilistic updating. This difference results in distinct behavior in the fact that simulated p-bit algorithm tends to maintain exploration even at low temperatures, while simulated annealing becomes increasingly deterministic. The simulated p-bits algorithm is described in Algorithm 1.

The value of input signal I_i is computed by traversing the neighbors of spin i and this operation accounts for the majority of the computation time for each spin update. In practice, the average spin-flip acceptance rate over the entire annealing schedule is typically around 15%, although the instantaneous probability is higher at the beginning due to higher effective temperature and decreases as the system cools down. Additionally, there are numerous shared terms in the input signals across different spins, which could be reused to reduce redundant calculations. Therefore, it is more efficient to update the corresponding terms only when the spin is flipped.

C. The pipeline of HOPbit

The pipeline of HOPbit contains three parts, including transformation, solving, selection, and demapping (Fig. 1). Specifically, it is listed as follows:

(a) Reduce MIMO detection problem into the higher-order Ising form.

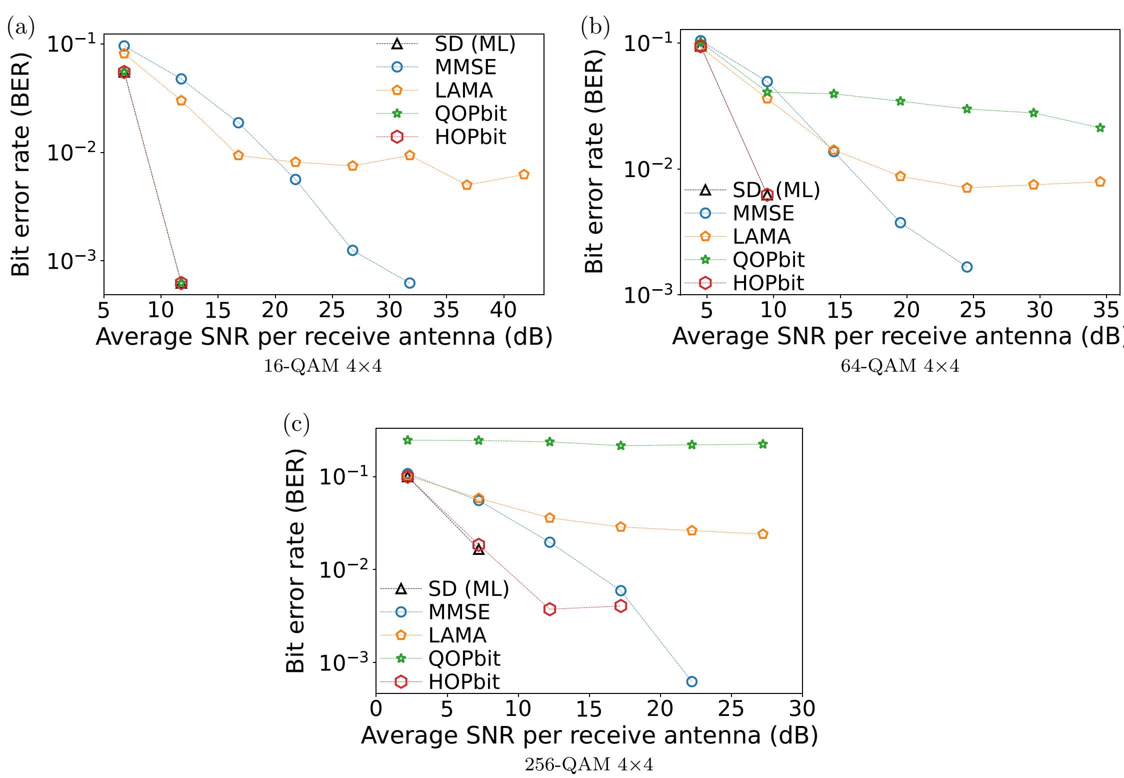

FIG. 2. Comparisons of BER of different detectors on different large MIMO systems with high-order-modulations on the Gaussian channel. The missing data points mean no error for detectors within the experimental instances.

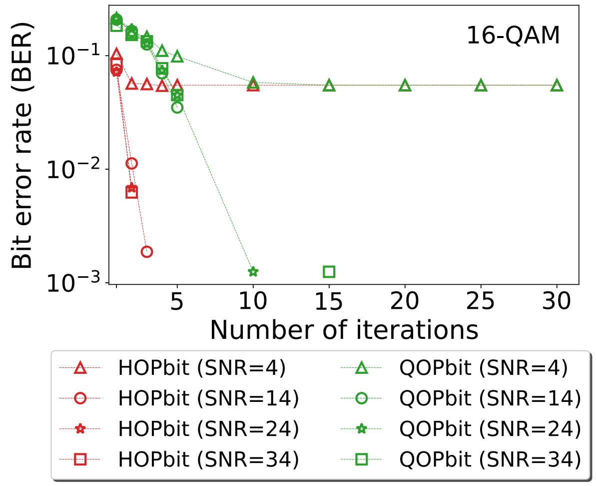

FIG. 3. BER as a function of the number of iterations for HOPbit and QOPbit detectors in four-user 16-QAM MIMO systems with varying high-order-modulations. The missing data points mean no error for detectors within the experimental instances.

(b) Perform N runs of simulated p-bits algorithm to get a set of candidate solutions.

(c) Select the best solution from the candidate solutions based on Eq. (3).

(d) Directly transform bits into symbols based on Eq. (5).

IV. EXPERIMENTS

To investigate the effectiveness of HOPbit, some evaluations and comparisons with other state-of-the-art detectors are conducted on large MIMO systems. In this section, we first present the experimental dataset, and then introduce evaluation metrics, and finally present the results.

A. Experimental setup

We simulate wireless MIMO channels between users and receiver antennas through independent and identically distributed Gaussian channels as well as additive white Gaussian noise (AWGN) for SNR ranging from 15 to 55 dB. Each SNR setting contains 100 different instances in order to assess the average performance of the detectors. The MIMO sizes are 4×4 for high-order-modulations including 16-QAM, 64-QAM, and 256-QAM. In order to verify the ability of HOPbit to eliminate the inconsistency error, we compare HOPbit with the detector based on ML-to-Ising transformation and use the same solver as HOPbit, which is denoted as QOPbit here. In addition, some other state-of-the-art detectors in large and massive MIMO systems, including MMSE, large MIMO approximate message passing (LAMA) [42,43] and the optimal sphere decoder- (SD) based maximum-likelihood MIMO detector [44], are included in the comparison. LAMA iteratively estimates the probability distribution of signals based on approximate message passing. It has low complexity and is readily implementable in practice, but it works well only in the specific scenario [45]. SD essentially searches for the most likely transmitted symbol by considering a hypersphere in the signal space and iteratively refining the search space based on pruned tree search [46]. However, it still suffers from exponentially increasing in computational demand.

To further show the advantage of HOPbit, we implement the same experiment setting as ParaMax and compare the

TABLE I. Computational complexity analysis for MMSE, QOPbit, and HOPbit with and without precomputation. I is the number of iterations and $\tilde{I}(\ll 2IN_t m)$ denotes the number of updated bits during the iterations.

	Preprocessing (update H)		Optimization	
Methods	Multiplication	Addition	Multiplication	Addition
MMSE	$24N_r^2N_t + \frac{8}{3}N_t^3$ $+3N_t^2 - \frac{13}{6}N_t$	$24N_r^2N_t + \frac{8}{3}N_t^3$ $+2N_t^2 - 4N_r^2$ $-8N_tN_r + 2N_r$ $-\frac{5}{6}N_t$	$4N_rN_t + 2N_t$	$2N_t(2N_r + 2^{\frac{m}{2}} - 2)$
QOPbit	$8N_rN_t^2$	$8N_rN_t^2 - 4N_t^2$	$2ImN_t^2 + 8N_rN_t +$ $(\frac{Im^2}{2} + 6Im + m - 2)N_t$	$2ImN_t^2 + 8N_rN_t +$ $(\frac{Im^2}{2} + 5Im + m - 4)N_t$
HOPbit	$8N_rN_t^2$	$8N_rN_t^2 - 4N_t^2$	$2ImN_t^2 + 8N_rN_t +$ $(7Im + m - 2)N_t$	$2ImN_t^2 + 8N_rN_t +$ $(5Im + m - 4)N_t$
QOPbit precomputation	$8N_rN_t^2$	$8N_rN_t^2 - 4N_t^2$	$4N_rN_t + 2mN_t^2 +$ $(2Im + 6m)N_t + 2\tilde{I}$	$4N_rN_t + 2mN_t^2 +$ $(2Im + 2\tilde{I} + m - 2)N_t$
HOPbit precomputation	$8N_rN_t^2$	$8N_rN_t^2 - 4N_t^2$	$4N_rN_t + 2mN_t^2 +$ $(2Im + 6m)N_t + \frac{m}{2}\tilde{I}$	$4N_rN_t + 2mN_t^2 +$ $(2Im + 2\tilde{I} + m - 2)N_t$ $+ \left(\frac{m}{2} - 2\right)\tilde{I}$

HOPbit with ZF, ParaMax, and 2R-ParaMax in 12-user 16-QAM. That is, we set the number of iterations to 50.

To evaluate the performance of HOPbit in more realistic channel models, we compare HOPbit with QOPbit as well as some traditional detectors (e.g., ZF, LMMSE) in flat fading MIMO channels. We simulate $N_t = 16$ MIMO systems with QAM-16, QAM-64, and QAM-256 for various values of N_r ($N_r = 16, 64$) at ranging from 5 to 30 dB SNR by Sionna [47], a Python library. For each setting, we randomly generate 100 000 instances in order to achieve BER up to near 10^{-6}.

For both QOPbit and HOPbit, the annealing schedule was set to be linear. The annealing parameters, specifically the initial and final inverse temperature, denoted by β_0 and β_T, were selected via a grid search within range $(0, 10)$. The step size is set as $\Delta_\beta = (\beta_T - \beta_0)/\text{max\_iters}$. This tuning procedure was applied consistently across all scenarios to ensure that each solver operated under its best-performing configuration in a fair and comparable manner. It is also worth noting that, for a given MIMO configuration (i.e., number of users, antennas, and modulation scheme), we used the same annealing schedule for all tested SNR setting. This choice avoids overfitting the schedule to specific noise levels and ensures that the observed performance improvements of HOPbit and QOPbit are robust across a range of channel conditions.

B. Results

The detection accuracy is measured by BER, the ratio of the bits that have been detected by mistake relative to the total number of bits transmitted in a transmission. The number of iterations and anneals of HOPbit and QOPbit are the same, 30 and 100, respectively, in order to compare detectors fairly. QOPbit exhibits inferior performance compared to HOPbit (Fig. 2). The performance of QOPbit deteriorates significantly with increasing modulation order, as the inconsistency error accumulates. In contrast, HOPbit

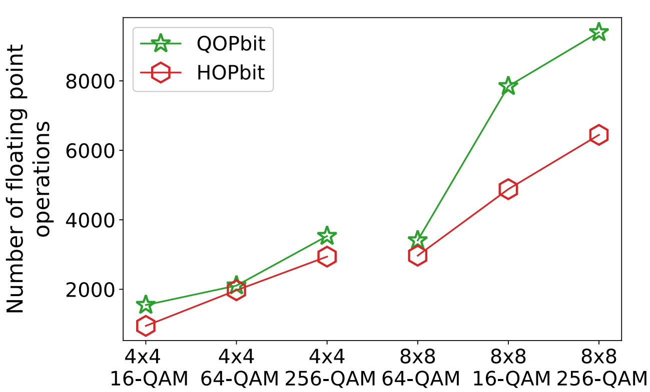

FIG. 4. Comparison of computational complexity between QOPbit and HOPbit with precomputation in varying MIMO systems.

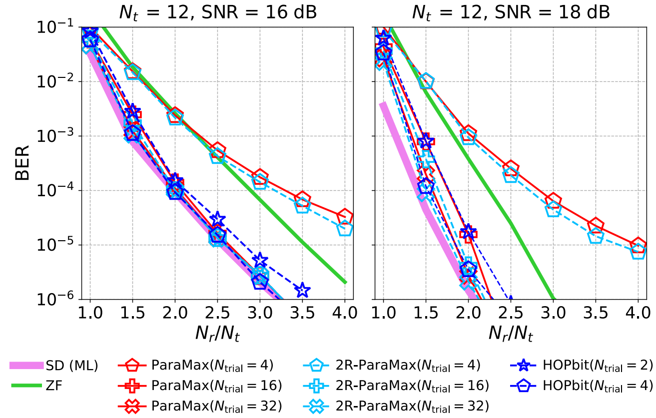

FIG. 5. Comparisons of BER for HOPbit and ZF in 12-user MIMO regimes and/or SNRs with 16-QAM. It is worth noting that the results of SD, ParaMax, and 2R-ParaMax are obtained from Ref. [21].

leverages ML-to-higher-order-Ising transformation to mitigate the inconsistency error, enabling rapid convergence to an approximate optimal solution. Consequently, HOPbit also outperforms MMSE and LAMA, and exhibits a similar detection accuracy as SD. As a result, it further extends QIA-based detectors” feasible MIMO regimes. The relatively diminished performance of HOPbit at high SNR arises from its fixed parameter setting for all instances with the same MIMO size (Fig. 2). It is reasonable to tailor the annealing schedule according to the noise level to enhance performance.

Although HOPbit and QOPbit reach the same BER in four-user 16-QAM MIMO systems, we evaluate BER as a function of iterations for them and find that HOPbit converges extremely fast, with only three iterations approximately across varying SNR settings while QOPbit converges after 15–20 iterations (Fig. 3). It further illustrates that ML-to-Ising transformation can accelerate the convergence of detectors.

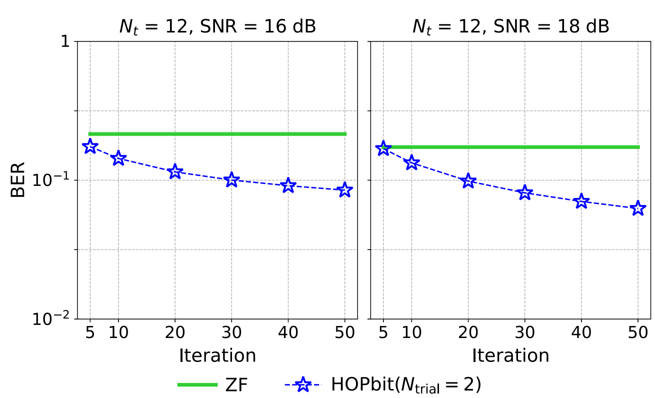

FIG. 6. BER as a function of N_{trial} varying SNRs in 12×12 large MIMO.

Then, we analyze the computational complexity of MMSE, QOPbit and HOPbit with and without precomputation (Table I). With precomputation, QOPbit and HOPbit can only compute the terms related to the flipped bits while keeping other terms unchanged, which can further reduce the computational complexity. For MMSE, the computational complexity for optimizing is quite low after preprocessing the channel matrix. For QOPbit and HOPbit, the computational overhead of preprocessing is low, but the optimization is time consuming. To sum up, the time complexity of MMSE and QOPbit and HOPbit is $O(N_r^2 N_t)$ and $O(N_r N_t^2)$, respectively, which means that QOPbit and HOPbit are much more efficient than MMSE. The computation complexity of QOPbit and HOPbit is quite close in a single iteration. Taking the faster convergence of HOPbit into account, the whole processing time of HOPbit is much lower than QOPbit, especially for a larger system with higher modulations (Fig. 4). On the other hand, QOPbit and HOPbit naturally benefit from parallel computing through running multiple independent anneals in parallel

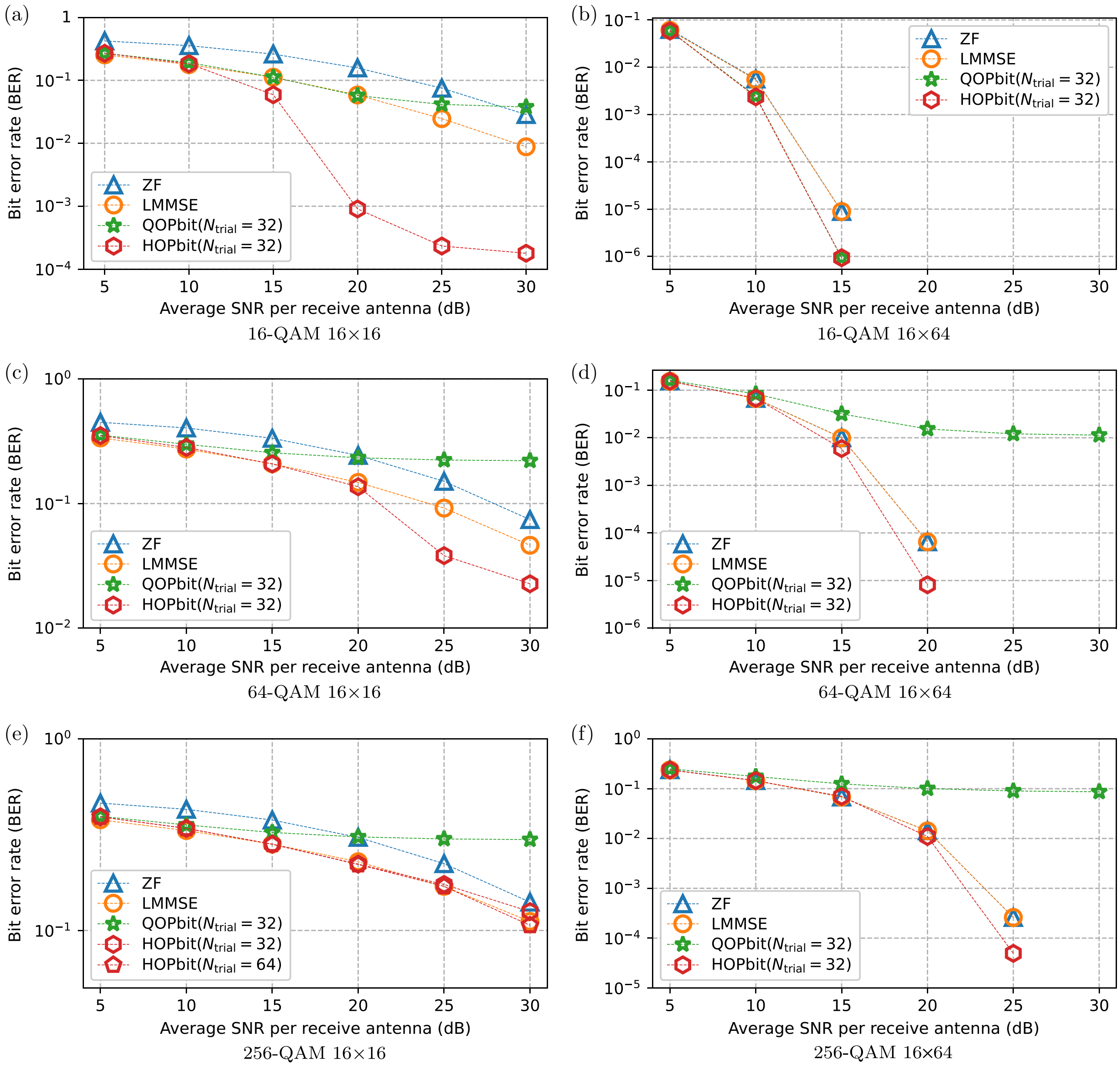

FIG. 7. Comparisons of BER of different detectors on different large MIMO systems with high-order-modulations on flat fading channel. The missing data points mean no error for detectors within the experimental instances. (a) 16-QAM 16 × 16. (b) 16-QAM 16 × 64. (c) 64-QAM 16 × 16. (d) 64-QAM 16 × 64. (e) 256-QAM 16 × 16. (f) 256-QAM 16 × 64.

that improves both performance and execution time. In contrast, MMSE is a deterministic method and does not gain from such sampling diversity. Equipped with enough processing elements (the number of processing elements is no more than the number of anneals in this work), the computational time of HOPbit with precomputation is enough for LTE requirements.

Besides, we evaluate the BER of ZF and HOPbit as well as the results from Ref. [21] in various 12-user 16-QAM MIMO systems. The BER curve of ZF in Fig. 5 is almost the same as shown in the previous paper [21], which further illustrates the consistency of the experimental setup. Both ParaMax and 2R-ParaMax underperform by several orders of magnitude BER in terms of massive MIMO compared to ZF with four trials (each processing element contains two trials, so it adds up to four trials in two processing elements). Only when the number of trials increases to 16, the BER performance of Paramax becomes better than ZF and close to SD. For HOPbit, it can be observed that it outperforms ZF in all the 12-user MIMO regimes with 16-QAM even only with two trials. Equipped with four trials, HOPbit obtains the lower BER, which extremely approximates the performance of SD. Besides, thanks to the fast convergence of HOPbit, we only need five–ten iterations with two trials to outperform ZF (Fig. 6).

Finally, we evaluate the performance of various detectors on flat fading MIMO channels (shown in Fig. 7). Due to the large size of MIMO systems, ML detectors in Sionna are unable to solve it. Therefore, the comparison is conducted among HOPbit, QOPbit, ZF, and LMMSE. The conclusion is the same as that in Gaussian channels. QOPbit only performs well as HOPbit in small MIMO systems with low-order-modulations. As the order of modulations or the size of MIMO systems increases, QOPbit suffers from error floor and performance deteriorates, even worse than ZF and LMMSE. Although equipped with more trials, it cannot overcome error floor. However, HOPbit solves error floor and always achieves the best performance among all the detectors we involved. In Fig. 7(e), we increase the number of trials to improve the performance of HOPbit but it will not increase the computing time since it can be processed in parallel.

V. CONCLUSION

Previous research endeavors have revealed the challenges that quantum- and physics-inspired methods encounter when dealing with high-order-modulations, even when subjected to increased iterations and trials. In response, this study introduces the HOPbit detector, an alternative approach designed to address these issues. HOPbit first leverages the ML-to-higher-order-Ising transformation and then employs the simulated p-bits algorithm to effectively solve the large and massive MIMO detection problem.

Remarkably, the HOPbit detector demonstrates swift convergence and achieves near-maximum-likelihood performance, even in the presence of high-order-modulation schemes within the context of large MIMO systems. Experimental results confirm that HOPbit not only outperforms Ising solvers, including QOPbit, ParaMax, and 2R-ParaMax configured with identical settings, but also surpasses traditional detectors, such as ZF, MMSE, and LAMA.

However, there are still several challenges or improvements that can be discussed here.

(a) **Parameters setting:** The performance of HOPbit is sensitive to the parameters, including annealing schedule-related and the number of iterations, especially for the higher-order-modulations and large MIMO systems. Therefore, it is necessary to develop an automatic parameter tuning method for annealing-based detectors. In addition, the parameters should be set according to the noise level, like MMSE and LAMA.

(b) **Connection between different anneals:** The anneals in this work are independent, meaning that each run proceeds without sharing information with others. While this allows for fully parallel execution, solving harder problems often requires a larger number of independent samples to increase the probability of finding better solutions. Although this parallelism can be advantageous, it also leads to higher hardware resource consumption when implemented on specialized devices such as FPGAs or quantum-inspired hardware. To address this, future improvements could incorporate ideas from parallel tempering [48], reverse annealing [49], or memory-based heuristics such as tabu search [50], which may enhance sampling diversity and reduce the dependence on naive repeated sampling from scratch.

(c) **Adaptability to emerging hardware:** Although HOPbit does not require any specialized hardware, it is adaptable with the emerging p-bits solvers [40]. In addition, it is also compatible with other specialized hardware, such as the CMOS-based higher-order Ising machine [51] and the high-order oscillator Ising machine [52].

DATA AVAILABILITY

The data that support the findings of this article are not publicly available upon publication because it is not technically feasible and/or the cost of preparing, depositing, and hosting the data would be prohibitive within the terms of this research project. The data are available from the authors upon reasonable request.

APPENDIX: THE OBJECTIVE FUNCTION OF MIMO DETECTION IN HIGHER-ORDER ISING FORM

$$\begin{aligned}
&\min_{\mathbf{q}} 2\sum_{i=1}^{N_t}\sum_{j=1}^{N_r}\left[(H_{ij}^R y_j^R + H_{ij}^I y_j^I)x_i^R + (H_{ij}^R y_j^I + H_{ij}^I y_j^R)x_i^I\right] \\
&+ 2\sum_{i=1}^{N_r}\sum_{j=1}^{N_t}\sum_{k=j+1}^{N_t}\left[(H_{ij}^R H_{ik}^R + H_{ij}^I H_{ik}^I)(x_j^R x_k^R + x_j^I x_k^I) + (H_{ij}^R H_{ik}^I - H_{ij}^I H_{ik}^R)(x_j^I x_k^R - x_j^R x_k^I)\right] \\
&+ \sum_{i=1}^{N_t}\sum_{j=1}^{N_r}(H_{ij}^{R^2} + H_{ij}^{I^2})(x_i^{R^2} + x_i^{I^2}) \\
&= \min_{\mathbf{q}} -2\sum_{i=1}^{N_t}\sum_{j=1}^{N_r}\left[(H_{ij}^R y_j^R + H_{ij}^I y_j^I)\left(\sum_{a=1}^{m/2} 2^{\frac{m}{2}-a}\prod_{b=1}^{a} q_{mi-b}\right) + (H_{ij}^R y_j^I + H_{ij}^I y_j^R)\left(\sum_{a=1}^{m/2} 2^{\frac{m}{2}-a}\prod_{b=1}^{a} q_{mi-\frac{m}{2}-b}\right)\right] \\
&+ 2\sum_{i=1}^{N_r}\sum_{j=1}^{N_t}\sum_{k=j+1}^{N_t}\Bigg[(H_{ij}^R H_{ik}^R + H_{ij}^I H_{ik}^I)\left(\sum_{a=1}^{m/2}\sum_{b=1}^{m/2} 2^{m-a-b}\,(\prod_{c=1}^{a}\prod_{d=1}^{b} q_{mj-c}q_{mk-d} + \prod_{c=1}^{a}\prod_{d=1}^{b} q_{mj-\frac{m}{2}-c}q_{mk-\frac{m}{2}-d})\right) \\
&+ (H_{ij}^R H_{ik}^I - H_{ij}^I H_{ik}^R)\left(\sum_{a=1}^{m/2}\sum_{b=1}^{m/2} 2^{m-a-b}(\prod_{c=1}^{a}\prod_{d=1}^{b} q_{mj-\frac{m}{2}-c}q_{mk-d} - \prod_{c=1}^{a}\prod_{d=1}^{b} q_{mj-c}q_{mk-\frac{m}{2}-d})\right)\Bigg] \\
&+ \sum_{i=1}^{N_t}\sum_{j=1}^{N_r}(H_{ij}^{R^2} + H_{ij}^{I^2})\left[\sum_{k=1}^{m/2-1}\sum_{l=k+1}^{m/2} 2^{m-k-l}\left(\prod_{a=k}^{l-1} q_{mi-a} + \prod_{a=k}^{l-1} q_{mi-\frac{m}{2}-a}\right)\right].
\end{aligned} \tag{A1}$$

[1] M. A. Albreem, M. Juntti, and S. Shahabuddin, Massive MIMO detection techniques: A survey, IEEE Commun. Surv. Tutor. **21**, 3109 (2019).

[2] B. Trotobas, A. Nafkha, and Y. Louët, in *Advanced Radio Frequency Antennas for Modern Communication and Medical Systems* (IntechOpen, Rijeka, Croatia, 2020).

[3] S. Yang and L. Hanzo, Fifty years of MIMO detection: The road to large-scale MIMOs, IEEE Commun. Surv. Tutor. **17**, 1941 (2015).

[4] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, Massive MIMO for next generation wireless systems, IEEE Commun. Mag. **52**, 186 (2014).

[5] X. Gao, Z. Lu, Y. Han, and J. Ning, in *2014 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting* (IEEE, Beijing, China, 2014), p. 1.

[6] K. Nikitopoulos, J. Zhou, B. Congdon, and K. Jamieson, Geosphere: Consistently turning MIMO capacity into throughput, ACM SIGCOMM Comput. Commun. Rev. **44**, 631 (2014).

[7] M. Kim, S. Kasi, P. A. Lott, D. Venturelli, J. Kaewell, and K. Jamieson, Heuristic quantum optimization for 6G wireless communications, IEEE Netw. **35**, 8 (2021).

[8] M. Kim, D. Venturelli, and K. Jamieson, in *Proceedings of the ACM Special Interest Group on Data Communication* (Association for Computing Machinery, Beijing, China, 2019), p. 241.

[9] Z. I. Tabi, Á. Marosits, Z. Kallus, P. Vaderna, I. Gódor, and Z. Zimborás, Evaluation of quantum annealer performance via the massive MIMO problem, IEEE Access **9**, 131658 (2021).

[10] E. Crosson and A. W. Harrow, in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, Piscataway, USA, 2016), p. 714.

[11] T. Albash and D. A. Lidar, Demonstration of a scaling advantage for a quantum annealer over simulated annealing, Phys. Rev. X **8**, 031016 (2018).

[12] M. Kim, D. Venturelli, and K. Jamieson, in *Proceedings of the 19th ACM Workshop on Hot Topics in Networks* (Association for Computing Machinery, Virtual Event, USA, 2020), p. 110.

[13] N. N. Hegade, K. Paul, Y. Ding, M. Sanz, F. Albarrán-Arriagada, E. Solano, and X. Chen, Shortcuts to adiabaticity in digitized adiabatic quantum computing, Phys. Rev. Appl. **15**, 024038 (2021).

[14] A. Marandi, Z. Wang, K. Takata, R. L. Byer, and Y. Yamamoto, Network of time-multiplexed optical parametric oscillators as a coherent Ising machine, Nat. Photonics **8**, 937 (2014).

[15] E. S. Tiunov, A. E. Ulanov, and A. Lvovsky, Annealing by simulating the coherent Ising machine, Opt. Express **27**, 10288 (2019).

[16] S. Reifenstein, S. Kako, F. Khoyratee, T. Leleu, and Y. Yamamoto, Coherent Ising machines with optical error correction circuits, Adv. Quantum Technol. **4**, 2100077 (2021).

[17] T. Inagaki, Y. Haribara, K. Igarashi, T. Sonobe, S. Tamate, T. Honjo, A. Marandi, P. L. McMahon, T. Umeki, K. Enbutsu, *et al.*, A coherent Ising machine

for 2000-node optimization problems, Science **354**, 603 (2016).
[18] H. Goto, K. Endo, M. Suzuki, Y. Sakai, T. Kanao, Y. Hamakawa, R. Hidaka, M. Yamasaki, and K. Tatsumura, High-performance combinatorial optimization based on classical mechanics, Sci. Adv. **7**, eabe7953 (2021).
[19] H. Goto, K. Tatsumura, and A. R. Dixon, Combinatorial optimization by simulating adiabatic bifurcations in nonlinear Hamiltonian systems, Sci. Adv. **5**, eaav2372 (2019).
[20] P. L. McMahon, A. Marandi, Y. Haribara, R. Hamerly, C. Langrock, S. Tamate, T. Inagaki, H. Takesue, S. Utsunomiya, K. Aihara, *et al.*, A fully programmable 100-spin coherent Ising machine with all-to-all connections, Science **354**, 614 (2016).
[21] M. Kim, S. Mandrà, D. Venturelli, and K. Jamieson, in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, MobiCom '21 (Association for Computing Machinery, New York, NY, USA, 2021), p. 42.
[22] A. K. Singh, K. Jamieson, P. L. McMahon, and D. Venturelli, Ising machines' dynamics and regularization for near-optimal MIMO detection, IEEE Trans. Wirel. Commun. **21**, 11080 (2022).
[23] M. Norimoto, R. Mori, and N. Ishikawa, Quantum algorithm for higher-order unconstrained binary optimization and MIMO maximum likelihood detection, IEEE Trans. Commun. **71**, 1926 (2023).
[24] A. Gilliam, S. Woerner, and C. Gonciulea, Grover adaptive search for constrained polynomial binary optimization, Quantum **5**, 428 (2021).
[25] D. Bulger, W. P. Baritompa, and G. R. Wood, Implementing pure adaptive search with Grover's quantum algorithm, J. Optim. Theory Appl. **116**, 517 (2003).
[26] R. Babbush, J. R. McClean, M. Newman, C. Gidney, S. Boixo, and H. Neven, Focus beyond quadratic speedups for error-corrected quantum advantage, PRX Quantum **2**, 010103 (2021).
[27] E. Rodríguez-Heck, Linear and quadratic reformulations of nonlinear optimization problems in binary variables, 4OR-A Q. J. Oper. Res. **17**, 221 (2019).
[28] A. Mandal, A. Roy, S. Upadhyay, and H. Ushijima-Mwesigwa, in *Proceedings of the 17th ACM International Conference on Computing Frontiers* (Association for Computing Machinery, Catania, Sicily, Italy, 2020), p. 126.
[29] A. Verma and M. Lewis, Penalty and partitioning techniques to improve performance of QUBO solvers, Discrete Optim. **44**, 100594 (2022).
[30] M. Ayodele, in *European Conference on Evolutionary Computation in Combinatorial Optimization (Part of EvoStar)* (Springer, Madrid, Spain, 2022), p. 159.
[31] M. D. García, M. Ayodele, and A. Moraglio, in *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (Association for Computing Machinery, Boston, Massachusetts, 2022), p. 184.
[32] S. V. Isakov, I. N. Zintchenko, T. F. Rønnow, and M. Troyer, Optimised simulated annealing for Ising spin glasses, Comput. Phys. Commun. **192**, 265 (2015).
[33] M. O. Damen, H. El Gamal, and G. Caire, On maximum-likelihood detection and the search for the closest lattice point, IEEE Trans. Inf. Theory **49**, 2389 (2003).
[34] S. Verdú, Computational complexity of optimum multiuser detection, Algorithmica **4**, 303 (1989).
[35] P. L. Hammer, I. Rosenberg, and S. Rudeanu, On the determination of the minima of pseudo-Boolean functions, Stud. Cerc. Mat. **14**, 359 (1963).
[36] E. Boros and P. L. Hammer, Pseudo-Boolean optimization, Discret. Appl. Math. **123**, 155 (2002).
[37] 3GPP, 5G; NR; Physical Channels and Modulation, Tech. Rep. 38.211 (3rd Generation Partnership Project (3GPP), 2018), https://www.etsi.org/standards-search.
[38] W. A. Borders, A. Z. Pervaiz, S. Fukami, K. Y. Camsari, H. Ohno, and S. Datta, Integer factorization using stochastic magnetic tunnel junctions, Nature **573**, 390 (2019).
[39] J. Kaiser and S. Datta, Probabilistic computing with p-bits, Appl. Phys. Lett. **119**, 150503 (2021).
[40] S. Nikhar, S. Kannan, N. A. Aadit, S. Chowdhury, and K. Y. Camsari, All-to-all reconfigurability with sparse and higher-order Ising machines, Nat. Commun. **15**, 8977 (2024).
[41] N. Onizawa and T. Hanyu, Enhanced convergence in p-bit based simulated annealing with partial deactivation for large-scale combinatorial optimization problems, Sci. Rep. **14**, 1339 (2024).
[42] C. Jeon, R. Ghods, A. Maleki, and C. Studer, in *2015 IEEE International Symposium on Information Theory (ISIT)* (IEEE, Hong Kong, China, 2015), p. 1227.
[43] H. Bitra and P. Ponnusamy, Large scale MIMO analysis using enhanced lama, Wirel. Pers. Commun. **126**, 2469 (2022).
[44] A. Burg, M. Borgmann, M. Wenk, M. Zellweger, W. Fichtner, and H. Bolcskei, VLSI implementation of MIMO detection using the sphere decoding algorithm, IEEE J. Solid-State Circuits **40**, 1566 (2005).
[45] H. He, A. Kosasih, X. Yu, J. Zhang, S. Song, W. Hardjawana, and K. B. Letaief, in *2023 IEEE Wireless Communications and Networking Conference (WCNC)* (IEEE, Glasgow, Scotland, 2023), p. 1.
[46] C.-Y. Hung and T.-H. Sang, in *2006 IEEE International Symposium on Signal Processing and Information Technology* (IEEE, Vancouver, Canada, 2006), p. 502.
[47] J. Hoydis, S. Cammerer, F. Ait Aoudia, M. Nimier-David, L. Maggi, G. Marcus, A. Vem, and A. Keller, Sionna (2022), https://nvlabs.github.io/sionna/.
[48] R. H. Swendsen and J.-S. Wang, Replica Monte Carlo simulation of spin-glasses, Phys. Rev. Lett. **57**, 2607 (1986).
[49] D. Venturelli and A. Kondratyev, Reverse quantum annealing approach to portfolio optimization problems, Quantum Mach. Intell. **1**, 17 (2019).
[50] J. E. Beasley, Heuristic algorithms for the unconstrained binary quadratic programming problem, Tech. Rep. (Working Paper, The Management School, Imperial College, London, England, 1998).
[51] Y. Su, T. T.-H. Kim, and B. Kim, A reconfigurable CMOS Ising machine with three-body spin interactions for solving boolean satisfiability with direct mapping, IEEE Solid-State Circuits Lett. **6**, 221 (2023).
[52] C. Bybee, D. Kleyko, D. E. Nikonov, A. Khosrowshahi, B. A. Olshausen, and F. T. Sommer, Efficient optimization with higher-order Ising machines, Nat. Commun. **14**, 6033 (2023).