QUBO-Formulated MIMO Detection Under Precision-Constrained Hardware

Seyedkhashayar Hashemi[(1)], Elisabetta Valiante[(2)], Ignacio Rozada[(2)], and Moslem Noori[(2)]

[(1)] University of Alberta, Edmonton, AB, Canada [(2)] 1QB Information Technologies (1QBit), Vancouver, BC, Canada

In communication systems, the Multiple-Input Multiple-Output (MIMO) detection problem refers to the scenario where multiple single-antenna users transmit signals to a multi-antenna base station (BS), and the goal is to detect the transmitted signal of each user separately. The maximum likelihood (ML) detector is known as the optimal detection strategy; however, it involves an exhaustive search over the entire search space, and the complexity of this method grows exponentially with the number of transmit antennas and the modulation order. Linear, low-complexity detectors would be near-optimal in an underloaded system, where the number of BS antennas is much larger than the number of users. However, the performance of these detectors significantly degrades when the system is fully loaded and the number of users is almost equal to the number of BS antennas. There are Nonlinear methods that offer better performance compared to the linear ones, but they would be computationally expensive in fully loaded massive MIMO systems.[1]

Formulating MIMO detection as a Quadratic Unconstrained Binary Optimization (QUBO) problem[2] enables the use of highly parallel, physics-inspired hardware-accelerated solvers and non-von Neumann architectures, in particular efficient in-memory computing solvers[3]. However, embedding the continuous-valued QUBO coefficients into the hardware introduces quantization noise due to the hardware's finite precision for embedding the coefficients. This constraint can severely degrade the detection accuracy, and the extent of this degradation requires careful study.

We present a rigorous analysis of the effect of using a finite-precision hardware-accelerated QUBO solver on the MIMO detection performance. First, we analytically derive the probability distribution functions of the QUBO matrix entries and introduce novel homogeneous and heterogeneous quantization schemes accustomed to either each QUBO matrix realization or just using the QUBO matrix entries' statistical information. In addition, theoretical analysis is conducted to derive a sufficient condition on the required precision to maintain the optimal solution for the MIMO detection problem. Extensive numerical experiments, across various MIMO system sizes and modulation orders (up to 256-QAM), validate our analysis. The results demonstrate that a heterogeneous quantization—applying different precision levels to diagonal and off-diagonal terms of the QUBO matrix—matches the full-precision baseline bit-error rate (BER) using significantly fewer bits than homogeneous approaches. Finally, we provide hardware-aware guidelines for selecting the optimal quantization strategy to balance a trade-off between the throughput and computational resources for the hardware-accelerated QUBO solvers.

[1] S. Yang and L. Hanzo, IEEE Commun. Surv. Tutor., 17, 1941-1988, 2015.

[2] M. Kim et al., Proceeding of the ACM Special Interest Group on Data Communication, SIGCOMM '19, 241-255, 2019.

[3] M. Hizzani et al., 2024 IEEE ISCAS, 5, 1-5, 2024.