A Probabilistic Reformulation Technique for Discrete RIS Optimization in Wireless Systems

Anish Pradhan, *Graduate Student Member, IEEE,* and Harpreet S. Dhillon, *Fellow, IEEE*

***Abstract*—The use of reconfigurable intelligent surfaces (RIS) can improve wireless communication by modifying the wireless link to create virtual line-of-sight links, bypass blockages, suppress interference, and enhance localization. However, enabling the RIS to modify the wireless channel requires careful optimization of the RIS phase-shifts. Although discrete RIS is more practical given hardware limitations, continuous RIS phase-shift optimization has attracted significantly more attention than discrete RIS optimization, which suffers from issues like quantization error and scalability. To overcome these issues, we develop a comprehensive probabilistic technique to transform discrete optimization problems into optimization problems of continuous domain probability parameters by interpreting the discrete optimization variable as a categorical random vector and computing expectations with respect to those parameters. We rigorously establish that for the unconstrained case, the optimal points of the reformulation and the original problem coincide. For the constrained case, we prove that the transformed problem is a relaxation of the original problem. We apply the proposed technique to two canonical discrete RIS applications: SINR maximization and overhead-aware rate and energy efficiency (EE) maximization. The reformulation enables both stochastic and analytical interpretations of the original problems, as we demonstrate in our RIS applications. The former interpretation yields a stochastic sampling technique, whereas the latter yields an analytical gradient descent (GD) approach that employs closed-form approximations for the expectation. We have explicitly derived the worst-case computational complexities of the proposed algorithms. The numerical results demonstrate that the proposed technique is applicable to a variety of discrete RIS optimization problems and outperforms other general approaches, such as closest point projection (CPP) and semidefinite relaxation (SDR) methods.**

***Index Terms*—Reconfigurable intelligent surface, discrete optimization, categorical random variables.**

I. Introduction

An RIS is a large array composed of low-cost reflecting elements, each of which can impart controllable phase-shifts to the incident signal, thereby modifying the propagation channel. However, because of hardware constraints, the phase-shift induced by each reflective element is normally limited to a set of discrete values. When configured appropriately, RISs can create multiple virtual LoS links [2], improve channel rank [3], transform a fast-fading channel to a slow-fading one [4], suppress co-channel interference [5], enhance localization performance [6]–[8], etc. However, optimizing the RIS phase-shifts is the first step to reaping these benefits. Despite the fact that RIS optimization has been extensively investigated in the literature, the majority of its attention has been directed toward the scenario of continuous phase-shifts as this scenario allows easier insights and upper-bounds on the performance of a wireless network. As a consequence, the discrete RIS optimization techniques often appear as an afterthought and the existing techniques that deal with discrete RIS optimizations suffer from various issues, such as scalability and arbitrarily bad performance due to quantization error.

Motivated by the scarcity of scalable and reliable discrete RIS optimization techniques [9], stochastic interpretation of semidefinite relaxation technique [10], and recent efforts to approach binary optimization problems [11] with a lens of probability, we develop a comprehensive technique to transform optimization problems of discrete variables into optimization problems of continuous domain probability parameters. We also rigorously prove that in terms of optimal points, the transformed problem is mathematically equivalent in the unconstrained case and a relaxation with respect to the original problem in the constrained case. Moreover, we gain further insights into our reformulation by investigating the simple two-way partitioning problem and report several moment and gradient results for quadratic forms in binary optimization problems. Ultimately, we apply this reformulation in two different canonical discrete RIS optimization problems demonstrating both the stochastic and analytical approaches. The numerical results confirm that the expectation-based algorithms outperform the conventional approaches. Note that, even though the proposed reformulation is inspired by discrete RISs, the scope of the reformulation is more general and could potentially find applications in other domains as well.

A. Related Work and Motivation

Although the study and design of discrete RIS phase-shifts have sparked some interest recently, a large portion of the literature relaxes the discrete constraint to a continuous one, solves the approximate problem, and then quantizes the solution to the closest discrete point. This two-fold approximation is shown to provide arbitrarily bad solutions in the worst-case scenario [9]. Yet, continuous RIS optimization remains a big part of the discrete RIS optimization literature. In light of this, some of the most used optimization strategies for continuous RIS phase-shits are combinations of a) SDR, b) minorize-maximization (MM) algorithm, c) penalty methods, d) manifold optimization, e) alternating direction method of multipliers (ADMM), and f) treating phase-shifts as optimization variables instead of the complex gains they provide.

In [2], [12], the authors jointly optimized the active beamforming vector and RIS-based passive beamforming vector in

A. Pradhan and H. S. Dhillon are with Wireless@VT, Department of ECE, Virginia Tech, Blacksburg, VA, USA (email: {pradhananish1, hdhillon}@vt.edu). The support of U.S. National Science Foundation (Grants ECCS-2030215 and CNS-2225511) is gratefully acknowledged. This paper was presented in part at the IEEE PIMRC 2023, Toronto, Canada [1].

multiple-antenna systems employing SDR. The authors of [13] utilized RISs to enhance the physical layer security of a multiuser multiple-input-single-output (MU-MISO) wireless system. In particular, they used a combination of a penalty-based approach, SDR, and successive convex approximation (SCA) to address the unit modulus constraint of RIS phase-shifts. Energy-efficient RIS designs for a MU-MISO wireless system are developed in [14]. In this paper, the authors developed two algorithms to maximize energy efficiency of the network. One of them uses the gradient descent method whereas the other uses the MM algorithm while considering unit modulus constraint and a realistic power consumption model. Similarly, in [15], the authors leveraged the MM algorithm and complex circle manifold (CCM) method to propose two algorithms that maximize the weighted sum-rate of a multicell MIMO network. In an RIS-assisted backscatter system, the RIS is optimized with a combination of SDR and ADMM technique in [16]. In [17], the weighted sum-rate is maximized in an RIS-aided cell-free network through ADMM. The authors of [18] tackled the resource allocation problem in an RIS-assisted wireless network. They developed an algorithm that uses the SCA and penalty method to jointly optimize phase-shifts and on-off status of RISs to maximize energy efficiency under a total power constraint. The authors of [19] enhanced the physical layer security by optimizing RIS phase-shifts with fractional programming and manifold optimization techniques. Using a similar technique, the authors of [20] developed an algorithm combining both SDR and manifold optimization techniques to optimize an RIS-aided edge caching system. Recently, the authors of this paper treated the vector of phase-shifts itself as an optimization variable instead of the vector of the complex gains they provide and optimized the RIS phase-shifts with the GD method to maximize the SINR [21] similar to the methodology used in [22].

In the realm of discrete RIS optimization literature, the optimization tools often used are exhaustive search, CPP from a continuous relaxation, and branch-and-bound (BB) methods. In [23], the authors investigated a practical discrete RIS-aided wideband orthogonal frequency division multiplexing (OFDM) system and optimized the discrete RIS element-wise exhaustive search in an alternating optimization framework. The authors of [24], [25] optimized the RIS phase-shifts in a MISO wireless network using BB methods that scale exponentially with the number of RIS elements. An RIS-aided MIMO system with low-resolution digital-to-analog converters (DACs) is jointly optimized with particle swarm optimization (PSO) algorithm in [26]. This algorithm is shown to work with both continuous and discrete RIS phase-shifts. For maximizing the achievable rate in a MIMO system, the authors in [27] relaxed the discrete RIS constraints to continuous ones and used a projected gradient method (PGM) to solve the problem. Similar continuous relaxation and CPP method was conducted in [28]. Specifically, the authors in [28] investigated the spectral efficiency and energy efficiency trade-off in a MU-MIMO setup. In that paper, the discrete RIS phase-shift constraint was relaxed to a continuous one and then solved by MM and accelerated gradient method. The mentioned optimization problems are either not scalable or suffer from arbitrarily bad performance due to the mentioned two-fold approximation. However, there are some recent efforts [9], [29], [30] that provide scalable optimal discrete RIS beamforming optimization for single-input-single-output systems as the resulting objective function has a low-rank matrix with $d \leq 2$ based on the fixed-rank result of [31]. These strategies suffer from being too specialized for single-antenna scenarios and do not work in multi-antenna scenarios that are more practical.

Going beyond the RIS literature, there have been some interesting approaches to binary quadratic optimization problems [10], [11], [32]–[34]. The authors of [10] show that SDR formulation is actually a stochastic version of the original non-convex quadratic program. Even when the non-convexity originates from binary variables, the stochastic interpretation still works. The authors of [32], [33] approach binary quadratic programs with a probabilistic data association (PDA) algorithm that treats the optimization variable as a binary random variable and iteratively updates the probabilities with Gaussian noise approximation. This is shown to achieve near-optimal results. Recently, the authors of [11] provide a stochastic gradient descent framework for binary optimization problems by similarly treating the optimization variable as a random variable and then taking expectation on it. Inspired by these probabilistic optimization techniques along with the lack of novel generalized discrete RIS optimization techniques, we develop a comprehensive probabilistic technique to transform discrete optimization problems that opens up new avenues to approach these problems in the continuous probability parameter domain. This technique is then used in discrete RIS cases to showcase its general nature and effectiveness.

B. Contributions

We approach the general discrete optimization problems with a different perspective of probability. This results in a comprehensive probabilistic reformulation technique with a wide applicability, including to the discrete RIS problems, which was our original inspiration behind this work. Our key contributions in this paper are listed next.

1) A comprehensive probabilistic technique for general discrete optimization problems: We develop a comprehensive probabilistic technique to reformulate general discrete optimization problems (that are not limited to binary programs) into continuous domain problems. In particular, we re-imagine the entries of the optimization variable as independently but not identically distributed (i.n.i.d) categorical random variables and replace the objective function and constraints, if any, with their expectations. We rigorously establish the equivalence between a general unconstrained problem with a unique optimal solution and the reformulated problem in terms of the optimal point. Additionally, when the original problem is constrained, we prove that the primal solution of the transformed problem is bounded between the dual and primal solution of the original problem. We also show that when strong duality holds, the transformed problem has the same optimal objective value as the original problem. Utilizing this technique, random sampling from a non-degenerate probability parameter solution can provide a better solution with the number of samples similar to Gaussian randomization in SDR.

2) Derivation of various analytical moments and their gradients associated with the quadratic form and binary random vectors: As discrete RIS problems often deal with binary phase-shifts, using our reformulation technique naturally gives rise to expectations associated with the quadratic form and the binary random vectors. For example, both the denominator and numerator of the SINR or secrecy rate often contain quadratic forms [22]. The quadratic form is a *canonical construct* that appears in the wireless literature frequently. For a gradient-based optimization approach, the gradients of these expectations will also be required. For this reason, we also derive the first and second moments of the said quadratic forms along with their gradients. These key intermediate results are later used in one of our algorithms demonstrating their importance.

3) GD algorithm for the SINR maximization: As the first canonical case study, we apply this technique to an SINR maximization problem and propose a stochastic GD and an analytical GD approach to solve the reformulated problem. We derive and use the first and second-order Taylor approximations of the expectation of the SINR in the analytical GD algorithm while an estimator of the gradient is used in the stochastic approach. The expectation-based algorithms are shown to perform better than the conventional practical approaches evaluated.

4) Stochastic sampling approach for ternary random vectors for EE and rate maximization: We also apply this technique to our second case study, an overhead-aware rate and EE maximization problem which leads to expectations associated with a ternary random vector. As deriving the analytical expectation was challenging for this specific case study, we develop a stochastic sampling approach for such a ternary random vector where the gradient is estimated with Monte Carlo (MC) samples, thereby demonstrating the versatility of the proposed approach. We demonstrate that this framework is well-suited for non-smooth objective forms and performs well in both interference-free and interference-rich scenarios. Moreover, the developed stochastic approach is demonstrated to work with different objective functions like rate and EE without the need for changing the algorithm.

5) Computational complexity discussion: We have also derived worst-case computational complexities with big-O notation for all the proposed algorithms.

Notations: The distribution of a standard complex normal random variable is denoted by $\mathcal{CN}(0,1)$. The matrix, scalar, and vector entities are denoted by $\mathbf{X}$, x, and $\mathbf{x}$, respectively. All the vectors are column vectors unless defined explicitly. For a vector $\mathbf{x}$, $\operatorname{diag}(\mathbf{x})$ denotes a diagonal matrix with the entries of $\mathbf{x}$ as its diagonal elements. For a matrix $\mathbf{X}$, $\mathbf{X}^H$, $\mathbf{X}^T$, $\operatorname{Re}(\mathbf{X})$, $\operatorname{Tr}(\mathbf{X})$, $\operatorname{diag}(\mathbf{X})$, and $\mathbf{X} \succeq 0$ denote its conjugate transpose, transpose, real part, trace, diagonal elements as a vector, and positive semidefiniteness, respectively. Additionally, $\mathbf{X}_{wd} = \mathbf{X} - \operatorname{diag}(\mathbf{X})$. The expectation operation is denoted by $\mathrm{E}[\cdot]$, $\operatorname{var}(\cdot)$ denotes a total variance operator which evaluates the trace of the variance-covariance matrix of the random vector argument, and the operator $\odot$ denotes element-wise multiplication between two matrices. The L0 and L2 norm are denoted by $\|\cdot\|_0$ and $\|\cdot\|_2$, respectively. The identity matrix and all-one column vector of dimension N are denoted by $\mathbf{I}_N$ and $\mathbf{1}_N$, respectively.

II. Probabilistic Reformulation for Discrete Optimization

A. The Case of Unconstrained Discrete Optimization Problem

We begin with a general unconstrained discrete optimization problem where we make no assumptions about the objective function's convexity. The optimization variable is a vector of length n and each of the entries can take a discrete value among the set $\mathcal{C} = \{c_1, c_2, \ldots, c_b\}$.

$$\min_{\mathbf{x} \in \mathcal{C}^n} \quad f(\mathbf{x}). \tag{1}$$

Our main goal is to reformulate the problem in a form that does not deal with the discrete domain and shares the optimal solution with the original problem. To that end, we propose to re-imagine entries of $\mathbf{x}$ as i.n.i.d categorical random variables with the following joint probability density function (PDF):

$$\mathbb{P}(\mathbf{x}|\mathbf{P}) = \prod_{i=1}^{n} \sum_{j=1}^{b} \delta(x_i - c_j) p_{i,j}, p_{i,j} \in [0,1], \sum_{j=1}^{b} p_{i,j} = 1, \tag{2}$$

where the (i,j)-th entry of the matrix $\mathbf{P}$ is denoted by $p_{i,j}$, the i-th entry of $\mathbf{x}$ is denoted by x_i, and $\delta(\cdot)$ is the Dirac delta function. We then reformulate the original problem into a stochastic optimization problem:

$$\min_{p_{i,j} \in \mathcal{F}} \quad \xi(\mathbf{P}) = \mathrm{E}_{\mathbf{x} \sim \mathbb{P}(\mathbf{x}|\mathbf{P})} \left[f(\mathbf{x}) \right], \tag{3}$$

where $\mathcal{F}$ is the set of possible $p_{i,j}$'s defined by (2). The connection between (1) and (3) and their solution sets are summarised in the following lemma.

Lemma 1. *The solution sets of the problems* (1) *and* (3) *are denoted by* $\Omega_{\mathbf{x}}$ *and* $\Omega_{\mathbf{P}}$ *and,*

$$\Omega_{\mathbf{x}} \subseteq \Omega_{\mathbf{P}}.$$

Moreover if the unique optimal solution of (1) *is* $\mathbf{x}_{\text{opt}}$*, then* $\mathbf{P}_{\text{opt}} = \operatorname{Degen}(\mathbf{x}_{\text{opt}})$ *is the unique optimal solution of* (3)*, where the* $\mathbf{P} = \operatorname{Degen}(\mathbf{x})$ *operation implies that the* (i,j)*-th entry of* $\mathbf{P}$ *is defined as* $p_{i,j} = 1$ *only when* $x_i = c_j$ *while all the other entries are zero.*

Proof: We observe that $\Omega_{\mathbf{x}}$ has b^n elements and each of them corresponds to one of the possible b^n combinations that $\mathbf{x}$ can take. In (3), the same objective values can be attained by the corresponding $\mathbf{P} = \operatorname{Degen}(\mathbf{x})$ which is the parameter matrix of n degenerate categorical distributions. Let us illustrate this with an example. Suppose we have a vector $\mathbf{x} = [-1, 1, -1]^T$. Each element of $\mathbf{x}$ can adopt either $c_1 = 1$ or $c_2 = -1$. By referring to the definition of the $\operatorname{Degen}(\cdot)$ function given in Lemma 1, the resulting parameter matrix for this vector is:

$$\operatorname{Degen}(\mathbf{x}) = \mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The interpretation of this matrix is as follows: the first element of $\mathbf{x}$ is -1 with a probability of 1, the second element is 1 with a probability of 1, and so forth. The parameter matrix thereby represents the degenerate distributions of the elements of $\mathbf{x}$ if it was reimagined as a random vector. Importantly, this parameter matrix can be translated back into the original $\mathbf{x}$ vector. Consequently, for every possible permutation of $\mathbf{x}$ (as given in equation (1)), there exists a corresponding parameter matrix in equation (3) that provides the same objective value. Upon reflecting on this, it becomes evident that the set $\Omega_{\mathbf{x}}$ must be included in $\Omega_{\mathbf{P}}$. In mathematical terms, this relationship can be represented as $\Omega_{\mathbf{x}} \subseteq \Omega_{\mathbf{P}}$.

For any feasible $\mathbf{P}$, it can be shown that,

$$\min_{\mathbf{x}} f(\mathbf{x}) \le \xi(\mathbf{P}) = \sum_{k=1}^{b^n} f(\mathbf{x}\{k\})\mathbb{P}(\mathbf{x}=\mathbf{x}\{k\}|\mathbf{P}) \le \max_{\mathbf{x}} f(\mathbf{x}), \tag{4}$$

where $\mathbf{x}\{k\}$ denotes the k-th combination out of possible b^n combinations of $\mathbf{x}$. This stems from the observations that the expectation is nothing but a convex combination of all the possible values of $f(\mathbf{x})$. This is possible because the probability terms, which are always nonnegative, sum up to one, enabling the expression of the expectation as this sum. Such a convex combination of scalar values essentially represents a probability-weighted average. Each scalar is weighed by its corresponding probability or chance of occurrence. These probabilities fundamentally dictate the placement of the weighted average on the line between the minimum and maximum scalar values. Due to the constraint of the probabilities adding up to one, this average cannot exist outside this range. For instance, when the probabilities tend to favor larger scalar values, the resulting combination leans closer toward the maximum and vice versa. It is crucial to clarify that equating the expectation to a convex combination does not imply that the expectation is a convex function. We are not discussing the convexity of the expectation itself. Yet, due to the inherent properties of convex combinations, inequality (4) consistently holds, as indicated in [35]. Additionally, Carathéodory's theorem offers further proof of this fact [11], [35].

Now assume that $\mathbf{x}_{\text{opt}}$ is the unique optimal solution of (1). It follows that, $\mathbf{P}_{\text{opt}} = \text{Degen}(\mathbf{x}_{\text{opt}})$ is an optimal solution of (3). Consider that $\exists \mathbf{P}_0 \neq \mathbf{P}_{\text{opt}}$, such that, $\xi(\mathbf{P}_0) = \xi(\mathbf{P}_{\text{opt}}) = f(\mathbf{x}_{\text{opt}})$. The parameter matrix $\mathbf{P}$ cannot denote n degenerate categorical distributions as the corresponding $\mathbf{x}_0 = \text{Degen}^{-1}(\mathbf{P}_0)$ would violate the uniqueness assumption on $\mathbf{x}_{\text{opt}}$. We then consider the non-degenerate distribution case. As the optimal value p^* is shown to be the same for both of these problems, we can assume that $p^* = f(\mathbf{x}\{k_0\})$ without any loss of generality. Then,

$$\xi(\mathbf{P}_0) = \sum_{k=1}^{b^n} f(\mathbf{x}\{k\})\mathbb{P}(\mathbf{x}=\mathbf{x}\{k\}|\mathbf{P}_0) = f(\mathbf{x}\{k_0\}) = p^* \tag{5}$$

$$\implies \sum_{k=1, k\neq k_0}^{b^n} (f(\mathbf{x}\{k\}) - f(\mathbf{x}\{k_0\}))\,\mathbb{P}(\mathbf{x}=\mathbf{x}\{k\}|\mathbf{P}_0) = 0. \tag{6}$$

As for some k, the value $f(\mathbf{x}\{k\})$ needs to be equal to $f(\mathbf{x}\{k_0\})$ for (6) to be true, this would also violate the uniqueness assumption on k_0. ■

B. The Case of Constrained Discrete Optimization Problem

Next, we explore whether such a coincident optimal solution through such a reformulation is valid for constrained problems as well. To that end, we write a general optimization problem with constraints without assuming convexity below:

$$\begin{aligned} \min_{\mathbf{x} \in \mathcal{C}^n} \quad & f_0(\mathbf{x}), \\ \text{s.t.} \quad & f_i(\mathbf{x}) \le 0 \quad \forall i = 1,2,\dots,m, \\ & h_j(\mathbf{x}) = 0 \quad \forall j = 1,2,\dots,r, \end{aligned} \tag{7}$$

where the optimal value and solution set of this problem are denoted by p_c^*, and $\Psi_{\mathbf{x}}$, respectively. The transformed formulation is expressed as:

$$\begin{aligned} \min_{p_{i,j} \in \mathcal{F}} \quad & \mathrm{E}_{\mathbf{x}\sim\mathbb{P}(\mathbf{x}|\mathbf{P})}[f_0(\mathbf{x})], \\ \text{s.t.} \quad & \mathrm{E}_{\mathbf{x}\sim\mathbb{P}(\mathbf{x}|\mathbf{P})}[f_i(\mathbf{x})] \le 0 \quad \forall i = 1,2,\dots,m, \\ & \mathrm{E}_{\mathbf{x}\sim\mathbb{P}(\mathbf{x}|\mathbf{P})}[h_j(\mathbf{x})] = 0 \quad \forall j = 1,2,\dots,r, \end{aligned} \tag{8}$$

with optimal value p_e^*, and solution set $\Psi_{\mathbf{P}}$.

Lemma 2. *The original problem* (7) *and the transformed problem* (8) *share the same dual problem with the dual solution* d^*. *Moreover,*

$$d^* \le p_e^* \le p_c^*. \tag{9}$$

Proof: Similar to the proof of Lemma 1, it can be readily seen that for every feasible $\mathbf{x}$ in (7), there is a corresponding feasible parameter matrix $\mathbf{P} = \text{Degen}(\mathbf{x})$ in (8). It directly follows from this observation with similar reasoning in the previous proof that the solution set $\Psi_{\mathbf{x}}$ is a subset of $\Psi_{\mathbf{P}}$. Consequently, the transformed optimization problem can be seen as a relaxation of the original constrained problem (7) and provides a better optimal value. Thus it can be established that $p_e^* \le p_c^*$.

Next, we investigate the dual function of the original problem by expressing it as the infimum of the Lagrangian [36], [37]:

$$\begin{aligned} g_c(\boldsymbol{\lambda}, \mathbf{v}) &= \inf_{\mathbf{x}\in\mathcal{C}^n} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) \\ &= \inf_{\mathbf{x}\in\{c_1,c_2,\dots,c_b\}^n} f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^{r} v_j h_j(\mathbf{x}), \end{aligned} \tag{10}$$

where λ_i is the i-th entry of $\boldsymbol{\lambda}$ and v_j is the j-th entry of $\mathbf{v}$. The vectors discussed here are the dual variables related to our problem. We can think of the dual function as a *softened* form of equation (7), which consists of more stringent or *hard* constraints [36]. Crucially, for all non-negative vectors $\boldsymbol{\lambda}$, the dual function serves as a consistent lower bound for the optimal value of the primal problem, denoted p_c^*. For a more in-depth treatment of this well-known result, readers can refer to [36]. Next, using Lemma 1, we can reformulate the above dual function into the following expression:

$$\begin{aligned} g_c(\boldsymbol{\lambda}, \mathbf{v}) = \inf_{p_{i,j}\in\mathcal{F}} & \mathrm{E}_{\mathbf{x}\sim\mathbb{P}(\mathbf{x}|\mathbf{P})}[f_0(\mathbf{x})] + \sum_{i=1}^{m} \lambda_i \mathrm{E}_{\mathbf{x}\sim\mathbb{P}(\mathbf{x}|\mathbf{P})}[f_i(\mathbf{x})] + \\ & \sum_{j=1}^{r} v_j \mathrm{E}_{\mathbf{x}\sim\mathbb{P}(\mathbf{x}|\mathbf{P})}[h_j(\mathbf{x})], \end{aligned} \tag{11}$$

where the optimal dual solution after maximizing the concave dual function is denoted by d^*. We note that the dual function of (8) is equivalent to (11). Given the lower bound characteristic of the dual function, it is logical to conclude $d^* \leq p_e^*$. Moreover, strong duality ensures equality. Compiling these inequalities, we deduce $d^* \leq p_e^* \leq p_c^*$, which consequently proves the Lemma. It further implies that the relaxation (8) is non-trivial and it is bounded by d^*, given that the dual solution is bounded. ■

C. Discussion on the Two-way Partitioning Example

In this subsection, we will focus on the simple two-way partitioning problem to demonstrate the technique. We focus on this foundational example to facilitate a more comprehensive understanding and to draw parallels with other probabilistic methods more effortlessly. While the selected problem covers a wide array of applications, such as binary phase beamforming in an RIS-aided network [9], we delve into two more complex applications in Section III. We begin with the description of the two-way partitioning problem below:

$$\max_{\mathbf{x} \in \{-1,1\}^n} \quad \mathbf{x}^T \mathbf{W} \mathbf{x}, \tag{12}$$

where $\mathbf{W} \in \mathbb{R}^{n \times n}$ is a symmetric matrix. Next, we derive our reformulation of (12) based on Lemma 1 below starting with the following result:

$$\begin{aligned} &\mathrm{E}_{\mathbf{x} \sim \mathbb{P}(\mathbf{x}|\mathbf{p}_x)} \left[\sum_{j=1}^{n} \sum_{i=1}^{n} x_i x_j W_{ij} \right] \\ &= \sum_{j=1}^{n} \sum_{\substack{i=1 \\ i \neq j}}^{n} (2p_{x,i} - 1)(2p_{x,j} - 1) W_{ij} + \sum_{j=1}^{n} W_{ii}, \end{aligned} \tag{13}$$

where $\mathbf{p}$ is the vector of parameters with $p_{x,i} = \mathbb{P}[x_i = 1]$ denoting the i-th entry, x_i denotes the i-th entry of $\mathbf{x}$, and W_{ij} denotes the i,j-th entry of $\mathbf{W}$. This result directly follows from the facts that $\mathrm{E}[x_i] = 2p_{x,i} - 1$, $\mathrm{E}[x_i^2] = 1$, and the entries are i.n.i.d.

Remark 1. *Note that, if all the entries of* $\mathbf{y}$ *are either* $+1$ *or* -1, $\mathbf{y} = \mathrm{Degen}^{-1}(\mathbf{p}_{\mathrm{x}})$. *In other words, in that case,* $\mathbf{y}$ *is a feasible* $\mathbf{x}$ *and vice versa.*

With this result, the transformed problem is as follows:

$$\max_{\mathbf{y} \in [-1,1]^n} \quad \mathbf{y}^T \mathbf{W}_{wd} \mathbf{y}, \tag{14}$$

where $\mathbf{y} = 2\mathbf{p}_x - \mathbf{1}$, and $\mathbf{W}_{wd}$ is the matrix $\mathbf{W}$ with its diagonal elements set to zero. Note that, we effectively converted a binary quadratic program (BQP) to a non-convex box-constrained quadratic program (BoxQP) emphasizing the ability of our reformulation to change the structure of a problem while being equivalent in terms of optimal point. However, this is a known result in the optimization community [38], [39] and our reformulation provides a probabilistic proof. Other than this structural change, we can obtain more insights about our reformulation by focusing on the SDR of the original problem which can be derived by considering the following stochastic program by taking expectation on the objective value and the domain of (12) [10]:

$$\begin{aligned} \max_{\mathbf{X} \succeq 0} \quad & \mathrm{E}_{\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \mathbf{X})} \left[\boldsymbol{\zeta}^T \mathbf{W} \boldsymbol{\zeta} \right], \\ \text{s.t.} \quad & \mathrm{E}_{\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \mathbf{X})} \left[\boldsymbol{\zeta} \odot \boldsymbol{\zeta} \right] = \mathbf{1}_n, \end{aligned} \tag{15}$$

where $\mathbf{X}$ is an arbitrary symmetric positive semidefinite matrix, and $\boldsymbol{\zeta}$ is a random vector drawn from a normal distribution with zero mean and covariance $\mathbf{X}$. Through the simple observation $\mathrm{E}_{\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \mathbf{X})} \left[\boldsymbol{\zeta} \boldsymbol{\zeta}^T \right] = \mathbf{X}$, this is equivalent to the classic SDR problem described below:

$$\begin{aligned} \max_{\mathbf{X}} \quad & \mathrm{Tr}\left(\mathbf{W}\mathbf{X}\right) \\ \text{s.t.} \quad & \mathbf{X} \succeq 0, \\ & \mathbf{X}_{ii} = 1, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{16}$$

In the above formulation, the addition of the rank-one constraint $\mathrm{rank}(\mathbf{X}) = 1$ would make the problem equivalent to (12). However, this relaxed formulation is solvable in polynomial time unlike (12). SDR can also be seen as a relaxation of the original problem when we allow x_i to be a multidimensional vector with a unit norm. These vectors can be found from The Cholesky decomposition of the solution of SDR. If the angle between two such vectors is really small, that implies that those two entries of $\mathbf{x}$ are more likely to be in the same group [40]. In other words, the SDR provides us with pairwise correlation information. In contrast, a relaxed version of our reformulation (14) will provide us the probabilities with which each entry of the original vector $\mathbf{x}$ will be $+1$ or -1.

The findings of this subsection reveal that our proposed reformulation has the potential to alter the structure of optimization problems. Even if the change is trivial in the case of simple objective forms, it is expected that more complex objective forms will yield non-trivial changes, which can have a significant impact on the efficiency and effectiveness of the optimization process. We will demonstrate these non-trivial structural changes in canonical case studies related to discrete RIS optimization in Section III. Furthermore, the results indicate that although our reformulation differs from SDR in terms of the information it provides, they share similarities in the formulation from a stochastic standpoint. These results encourage further exploration of our approach in more complex objective forms.

D. Some Useful Results for Quadratic Expressions for Binary Random Vectors

As most discrete RIS applications deal with binary phase-shift RIS $\{-1, +1\}$ due to its simplicity in operation and implementation, it is only appropriate to derive the analytical moments and their gradients associated with the binary random vectors defined in (14). They can be used in different optimization contexts with such expectation-based formulations. The higher moment results are motivated by the previous subsection and will be heavily used in the next section. We begin with the covariance matrix next.

Remark 2. *For a random vector* $\mathbf{x} \in \{-1, +1\}^n$ *with i.n.i.d* *entries and expectation* $\mathrm{E}[\mathbf{x}] = \mathbf{y}$*, the covariance matrix is*

$$\mathrm{E}[\mathbf{x}\mathbf{x}^T] = (\mathbf{y}\mathbf{y}^T) \odot \mathbf{E}_m + \mathbf{I}_N, \tag{17}$$

where $\mathbf{E}_m$ *is the all-one matrix with a hollow diagonal and* $\mathbf{p}$ *is defined similarly to* (29)*.*

Now, we state the first moment and its gradient in Lemma 3 without proof due to its trivial nature and partial proof in (13).

Lemma 3. *For a random vector* $\mathbf{x} \in \{-1, +1\}^n$ *with i.n.i.d entries and expectation* $\mathrm{E}[\mathbf{x}] = \mathbf{y}$*, the expectation and the gradient of a sum between a quadratic form and a linear form are*

$$\mu_{qf}(\mathbf{G}, \mathbf{z}, \mathbf{y}) = \mathrm{E}[\mathbf{x}^T\mathbf{G}\mathbf{x} + \mathbf{z}^T\mathbf{x}] = \mathbf{y}^T\mathbf{G}_{wd}\mathbf{y} + \mathrm{Tr}(\mathbf{G}) + \mathbf{z}^T\mathbf{y}, \tag{18}$$

$$\vartheta_{qf}(\mathbf{G}, \mathbf{z}, \mathbf{y}) = \nabla_{\mathbf{y}}\mathrm{E}[\mathbf{x}^T\mathbf{G}\mathbf{x}] = (\mathbf{G}_{wd} + \mathbf{G}_{wd}^T)\mathbf{y} + \mathbf{z}. \tag{19}$$

where $\mathbf{G}$ *is a real symmetric matrix.*

Next, we derive an expectation that is very important for covariance calculations between a quadratic form and a linear form in the next theorem.

Theorem 1. *For a random vector* $\mathbf{x} \in \{-1, +1\}^n$ *with i.n.i.d entries and expectation* $\mathrm{E}[\mathbf{x}] = \mathbf{y}$*, the expectation of a product between a quadratic form and a linear form is*

$$\mu_{ql}(\mathbf{G}, \mathbf{z}, \mathbf{y}) = \mathrm{E}[\mathbf{x}^T\mathbf{G}\mathbf{x}\mathbf{z}^T\mathbf{x}] = 2\mathbf{y}^T\mathbf{G}_{wd}\mathbf{z} + \mathbf{z}^T\mathbf{y}\mathrm{Tr}(\mathbf{G}) + \\ \mathbf{1}^T\{(\mathbf{G}_{wd}\mathbf{Y}_{wd}) \odot \mathbf{Y}_{wd}\}(\mathbf{y} \odot \mathbf{z}), \tag{20}$$

where $\mathbf{G}$ *is a real symmetric matrix and* $\mathbf{Y} = \mathbf{y}\mathbf{1}^T$*.*

Proof: See Appendix A. ■

We just state the gradient of the above expectation without proof in Corollary 1.

Corollary 1. *The gradient of the derived expectation in Theorem 1 can be calculated as:*

$$\vartheta_{ql}(\mathbf{G}, \mathbf{z}, \mathbf{y}) = 2\mathbf{G}_{wd}\mathbf{z} + \mathbf{z}\mathrm{Tr}(\mathbf{G}) + ((\mathbf{G}_T^T \odot \mathbf{E}_m)\mathbf{y}) \odot \mathbf{z} + \\ \mathrm{diag}(\mathbf{G}_T\mathrm{diag}(\mathbf{y} \odot \mathbf{z})\mathbf{E}_m) + (\mathbf{G}_T \odot \mathbf{E}_m)(\mathbf{y} \odot \mathbf{z}), \tag{21}$$

where $\mathbf{G}_T = \mathbf{G}_{wd}\mathbf{T}_0$*, and* $\mathbf{T}_0 = \mathrm{diag}(\mathbf{y})\mathbf{E}_m$*.*

Next, we focus on the second moment of a quadratic form in Theorem 2.

Theorem 2. *For a random vector* $\mathbf{x} \in \{-1, +1\}^n$ *with i.n.i.d entries and expectation* $\mathrm{E}[\mathbf{x}] = \mathbf{y}$*, the second moment of a quadratic form is*

$$\mu_{qs}(\mathbf{G}, \mathbf{y}) = \mathrm{E}[(\mathbf{x}^T\mathbf{G}\mathbf{x})^2] = \mathbf{y}^T\left(\mathbf{G}_s - \mathbf{F}(\mathbf{y})\right)\mathbf{y} + \mathrm{Tr}(\mathbf{G})^2 + \\ 2\mathrm{Tr}(\mathbf{Z}) + (\mathbf{y}^T\mathbf{G}\mathbf{y})^2 - \mathbf{d}^T\mathbf{G}_g\mathbf{d}, \tag{22}$$

where $\mathbf{G}$ *is a real symmetric matrix,* $\mathbf{d} = \mathbf{y} \odot \mathbf{y}$*,* $\mathbf{G}_s = 2\mathrm{Tr}(\mathbf{G})\mathbf{G}_{wd} + 4\mathbf{Z}_{wd}$*,* $\mathbf{Z} = \mathbf{G}_{wd}\mathbf{G}_{wd}^T$*,* $\mathbf{F}(\mathbf{y}) = (\mathbf{y} \odot \mathbf{y})^T\mathrm{diag}(\mathbf{G})(\mathbf{G} + \mathbf{G}_{wd}) + 4\mathbf{U}_{wd}$*,* $\mathbf{U} = [\mathbf{I}_N \otimes (\mathbf{y} \odot \mathbf{y})^T]\mathbf{B}$*, and* $\mathbf{G}_g = 2\mathbf{G}_{wd} \odot \mathbf{G}_{wd}$*. The matrix* $\mathbf{B}$ *is defined through blocks as*

$$\mathbf{B} = \begin{bmatrix} \mathbf{b}_{1,1}, \ldots, \mathbf{b}_{1,N} \\ \cdots, \cdots, \cdots, \\ \mathbf{b}_{N,1}, \ldots, \mathbf{b}_{N,N}. \end{bmatrix}, \tag{23}$$

where the i*-th element of* $\mathbf{b}_{k,j}$ *is* $\mathbf{b}_{k,j}^i = G_{wd_{ij}}G_{wd_{ki}}$*.*

Proof: See Appendix B. ■

Now, we derive the gradient of the second moment in the Corollary 2.

Corollary 2. *The gradient of the derived expectation in Theorem 2 can be calculated as:*

$$\vartheta_{qs}(\mathbf{G}, \mathbf{y}) = (\mathbf{G}_s + \mathbf{G}_s^T)\mathbf{y} + 2\mathbf{y}^T\mathbf{G}\mathbf{y}(\mathbf{G} + \mathbf{G}^T)\mathbf{y} - \\ 2\mathbf{y}^T(\mathbf{G} + \mathbf{G}_{wd})\mathbf{y}(\mathrm{diag}(\mathbf{G}) \odot \mathbf{y}) - \mathbf{d}^T\mathrm{diag}(\mathbf{G})(\mathbf{G} + \mathbf{G}_{wd})\mathbf{y} \\ - \mathrm{diag}(\mathbf{G})^T\mathbf{d}(\mathbf{G} + \mathbf{G}_{wd})^T\mathbf{y} - 2((\mathbf{G}_g + \mathbf{G}_g^T)\mathbf{d}) \odot \mathbf{y} - \\ 8\mathbf{y} \odot \mathbf{b}_s - 4(\mathbf{U}_{wd} + \mathbf{U}_{wd}^T)\mathbf{y}, \tag{24}$$

where $\mathbf{d} = \mathbf{y} \odot \mathbf{y}$*, and* i*-th entry of* $\mathbf{b}_s$ *is* $\mathbf{y}^T\mathbf{B}_t[i]\mathbf{y} - \mathrm{Tr}(\mathbf{B}_t[i])$*. The matrix* $\mathbf{B}_t[i]$ *can be derived by multiplicating the* i*-th column of* $\mathbf{G}_{wd}$ *with the* i*-th row of* $\mathbf{G}_{wd}$*.*

Proof: See Appendix C. ■

III. Applications of The Proposed Reformulation

In a MIMO communication scenario, optimizing RIS phase shifts can be a challenging task, particularly when dealing with discrete RISs. Discrete RIS problems are generally more difficult to solve, making it necessary to split the original problem into smaller sub-problems that can be handled separately. Therefore, we focus on the canonical forms of discrete RIS sub-problems that frequently appear in the literature. For a unified treatment, we have chosen a signal model capable of representing a range of RIS-aided scenarios and sub-problems, including a device-to-device communication link, a cellular network where each antenna serves a different user through antenna selection, and a wireless communication link with interferers while the receive beamformer vector remains fixed [21]. This signal model can be expressed in the following point-to-point representation:

$$y_r = (h_{d_0} + \mathbf{h}_0^H\mathrm{diag}(\boldsymbol{\theta})\mathbf{f}_0)x_{s,0} + \\ \sum_{i=1}^{N_I}(h_{d_i} + \mathbf{h}_i^H\mathrm{diag}(\boldsymbol{\theta})\mathbf{f}_i)x_{s,i} + w, \tag{25}$$

where y_r is the received signal from the Tx of interest (denoted by $i = 0$), h_{d_i} denotes the direct channel between the i-th Tx and Rx, $\mathbf{h}_i$ is the Tx-RIS channel, $\mathbf{f}_i$ denotes the RIS-Rx channel, $x_{s,i}$ is the data for the i-th Tx, $\mathrm{E}[x_{s,i}^2] = \beta_i$, $\boldsymbol{\theta}$ is the N-element discrete RIS phase configuration vector, N_I is the number of interferers, and w is the additive noise. Note that the users and interferers always transmit at their maximum power p. For a general MIMO communication scenario, these channels can be seen as the actual channels pre-multiplied and post-multiplied by precoding and receiver beamformer vectors, respectively. With such a versatile signal model, two use cases for RIS-assisted wireless communication systems are explored, namely, SINR maximization and overhead-aware RIS optimization. Note that, in both cases, RISs are assumed to be controlled by the receiver through an RIS controller [41].

A. SINR Maximization with RIS Optimization

1) System model: We consider a generic system model dictated by the signal model (25). We consider that the RIS phase vector $\boldsymbol{\theta} = [\theta_1 \quad \theta_2 \ldots \theta_n \ldots \theta_N]^T$ and $\theta_n \in \{-1, +1\}$.

For ease of notation, we also define $\mathbf{h}_{c_i} = \left(\mathbf{h}_i^H \mathrm{diag}(\mathbf{f}_i)\right)^H$ With this discrete RIS, the SINR can be expressed as,

$$\gamma = \frac{\beta_0 |h_{d_0} + \mathbf{h}_{c_0}^H \boldsymbol{\theta}|^2}{\sum_{i=1}^{N_I} \beta_i |h_{d_i} + \mathbf{h}_{c_i}^H \boldsymbol{\theta}|^2 + \sigma_w^2} = \frac{f_s(\boldsymbol{\theta})}{f_I(\boldsymbol{\theta})} = \frac{\boldsymbol{\theta}^T \mathbf{R}_0 \boldsymbol{\theta} + \mathbf{c}_0^T \boldsymbol{\theta}}{\boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta} + \mathbf{s}^T \boldsymbol{\theta}}, \tag{26}$$

where $\mathbf{R}_i = \beta_i \mathrm{Re}\left(\mathbf{h}_{c_i}\mathbf{h}_{c_i}^H + \frac{|h_{d_i}|^2}{N}\mathbf{I}_N\right)$, $\mathbf{K} = \sum_{i=1}^{N_I} \mathbf{R}_i + \frac{\sigma_w^2}{N}\mathbf{I}_N$, σ_w^2 is the variance of the additive Gaussian noise, $\mathbf{c}_i = 2\beta_i \mathrm{Re}(\mathrm{conj}(h_{d_i}\mathbf{h}_{c_i}))$, and $\mathbf{s} = \sum_{i=1}^{N_I} \mathbf{c}_i$.

```
Algorithm 1: E-GD
  Input:  $\mathbf{R}_i, \mathbf{c}_i, \varrho, \varepsilon, \epsilon_{th}, \beta_{\text{init}}, G\ \forall i$
  Output: $\bar{\boldsymbol{\theta}}_{i+1}$
  Initialize $t = 1$, $\delta_{GD} = 1$, and $\mathbf{y}_s^{(t)} = \mathbf{y}_{\text{init}}$.
  while $\delta_{GD} \leq \epsilon_{th}$
   do
      Initialize $\beta^{(1)} = \beta_{\text{init}}, d_f = -1$.
      Calculate $\nabla_{\mathbf{y}_s} \mathcal{J}_l(\mathbf{y}_s^{(t)})$ from (32) or (33).
      while $d_f \leq 0$
       do
          $\mathbf{y}_{new} = \mathbf{y}_s^{(t)} - \beta^{(t)} \nabla_{\mathbf{y}_s}\left(\mathbf{y}_s^{(t)}\right)$.
          Find $\mathbf{y}_{proj}$ by clipping the vector $\mathbf{y}_{new}$ in
          $[-\mathbf{1}_N, +\mathbf{1}_N]$.
          $d_f =$
          $-\mathcal{J}_l(\mathbf{y}_s^{(t)}) - \varepsilon\beta^{(t)} \|\nabla_{\mathbf{y}_s}\mathcal{J}_l(\mathbf{y}_s^{(t)})\|_2^2 + \mathcal{J}_l(\mathbf{y}_{proj})$.
          $\beta^{(t)} = \varrho\beta^{(t)}$.
      $\mathbf{y}_s^{(t+\frac{1}{2})} = \mathbf{y}_s^{(t)} - \beta^{(t)}\nabla_{\mathbf{y}_s}\mathcal{J}_l(\mathbf{y}_s^{(t)})$.
      $\mathbf{y}_s^{(t+1)} \in \min_{\mathbf{v}_y \in [-1,+1]^N} \|\mathbf{v}_y - \mathbf{y}_s^{(t+\frac{1}{2})}\|$.
      $t = t + 1$.
      $\delta_{GD} = \|\mathbf{y}_s^{(t+1)} - \mathbf{y}_s^{(t)}\|_2^2$.
  $\mathbf{p}_s = \frac{\mathbf{y}_s^{(t)}+1}{2}$.
  Based on this probability parameter vector $\mathbf{p}$, sample
  $G$ RIS phase-shift vectors.
  Choose the best RIS phase-shift vector $\boldsymbol{\theta}_{best}$ among
  them based on the resulting SINR.
  $\bar{\boldsymbol{\theta}}_{i+1} = \boldsymbol{\theta}_{best}$.
```

2) RIS optimization: In this subsection, our objective is to maximize the SINR given in (26) while the RIS elements are discrete in nature. The optimization problem is described below:

$$\min_{\boldsymbol{\theta} \in \{-1,+1\}^N} \quad -\frac{f_s(\boldsymbol{\theta})}{f_I(\boldsymbol{\theta})}. \tag{27}$$

As the domain of this problem is discrete and the problem is a fractional quadratic program, a common way to solve this problem is to relax the discrete domain and then project the solution to the closest discrete point. The relaxed version is solved through GD in [21]. Note that, we also consider SDR in the simulation results. We approach this problem with our reformulation (3) and transform this problem into a continuous domain problem. The reformulated problem is as follows:

$$\min_{\mathbf{y}_s \in [-1,+1]^N} \quad -\mathrm{E}_{\boldsymbol{\theta} \sim \mathbb{P}_B(\boldsymbol{\theta}|\mathbf{p}_s)}\left[\frac{f_s(\boldsymbol{\theta})}{f_I(\boldsymbol{\theta})}\right], \tag{28}$$

where $\mathbf{y}_s = 2\mathbf{p}_s - 1$ and $\boldsymbol{\theta}$ is assumed to be distributed with the joint PDF

$$\mathbb{P}_B(\boldsymbol{\theta}|\mathbf{p}_s) = \prod_{n=1}^{N} \left(\delta(\theta_n - 1)p_{s,n} + \delta(\theta_n + 1)(1 - p_{s,n})\right), \tag{29}$$

where $p_{s,n} \in [0,1]$ is the n-th entry of $\mathbf{p}_s$ and $\theta_n \in \{-1,+1\}$. We propose two approaches to solve (28): a) stochastic sampling approach (SSA), and b) analytical gradient descent approach. The former approach generally does not require an explicit expression of the gradient whereas the latter does. In SSA, a typical gradient estimator, which is based on the log-derivative trick and Monte Carlo sampling as described in [42], is often used in the GD algorithm. Following that, we have formulated an approach, SSA-B, which is a special case of SSA for binary variables using the same log-derivative trick and Monte Carlo sampling. We have omitted the details of SSA-B here, as this variant of SSA for binary variables has already appeared in a different context - the Bayesian optimal design of experiments, in [11]. For a detailed understanding of SSA-B, readers can refer to [11, Algorithm 3.1]. Next, we detail the analytical optimization approach for this case study. We explicitly develop the SSA for a non-binary random vector in the next case study, which is a direct result of the general probabilistic reformulation technique that we rigorously devised in Section II.

In the analytical gradient descent approach, calculating the direct expectation of a ratio of correlated random variables is difficult. So, we consider the Taylor series approximations of such an expectation [43]. Both the first-order approximation $\mathcal{J}_1(\mathbf{y}_s)$ and second-order approximation $\mathcal{J}_2(\mathbf{y}_s)$ are stated below:

$$\begin{aligned}\mathcal{J}_1(\mathbf{y}_s) &= \frac{\mathrm{E}[f_s(\boldsymbol{\theta})]}{\mathrm{E}[f_I(\boldsymbol{\theta})]} = \frac{\mu_{qf}(\mathbf{R}_0, \mathbf{c}_0, \mathbf{y}_s)}{\mu_{qf}(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)},\\ \mathcal{J}_2(\mathbf{y}_s) &= \mathcal{J}_1(\mathbf{y}_s) - \frac{\mathrm{E}[f_s(\boldsymbol{\theta})f_I(\boldsymbol{\theta})]}{\mathrm{E}^2[f_I(\boldsymbol{\theta})]} + \frac{\mathrm{E}[f_I^2(\boldsymbol{\theta})]\mathrm{E}[f_s(\boldsymbol{\theta})]}{\mathrm{E}^3[f_I(\boldsymbol{\theta})]}.\end{aligned} \tag{30}$$

The second-order approximation requires two additional expectations that are derived along with their gradients in (31). Using the definitions in (31), we can express the gradients of the Taylor series approximations as follows:

$$\nabla_{\mathbf{y}_s}\mathcal{J}_1(\mathbf{y}_s) = \frac{\vartheta_{qf}(\mathbf{R}_0, \mathbf{c}_0, \mathbf{y}_s) - \mathcal{J}_1(\mathbf{y}_s)\vartheta_{qf}(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}{\mu_{qf}(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}, \tag{32}$$

$$\begin{aligned}&\nabla_{\mathbf{y}_s}\mathcal{J}_2(\mathbf{y}_s) = \nabla_{\mathbf{y}_s}\mathcal{J}_1(\mathbf{y}_s) - \frac{\vartheta_{cv}}{\mu_{qf}^2(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)} +\\ &\mu_{qf}(\mathbf{R}_0, \mathbf{c}_0, \mathbf{y}_s)\left(\frac{\vartheta_v}{\mu_{qf}^3(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)} - \frac{3v(\mathbf{y}_s)\vartheta_{qf}(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}{\mu_{qf}^4(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}\right) +\\ &\frac{2c_v(\mathbf{y}_s)\vartheta_{qf}(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}{\mu_{qf}^3(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)} + \frac{v(\mathbf{y}_s)\vartheta_{qf}(\mathbf{R}_0, \mathbf{c}_0, \mathbf{y}_s)}{\mu_{qf}^3(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}.\end{aligned} \tag{33}$$

Note that, they are stated without proof as they can be derived trivially with the basic chain rule. Armed with these gradients, we can develop simple update rules of a projected

$$c_v(\mathbf{y}_s)=\mathrm{E}[f_s(\boldsymbol{\theta})f_I(\boldsymbol{\theta})]=\frac{\mu_{qs}(\mathbf{R}_0+\mathbf{K},\mathbf{y}_s)-\mu_{qs}(\mathbf{R}_0+\mathbf{K},\mathbf{y}_s)}{4}+\mu_{ql}(\mathbf{R}_0,\mathbf{s},\mathbf{y}_s)+\mu_{ql}(\mathbf{K},\mathbf{c}_0,\mathbf{y}_s)+\mathbf{c}_0^T\left((\mathbf{y}_s\mathbf{y}_s^T)\odot\mathbf{E}_m+\mathbf{I}_N\right)\mathbf{s},$$
$$\vartheta_{cv}=\nabla_{\mathbf{y}_s}c_v(\mathbf{y}_s)=\frac{\vartheta_{qs}(\mathbf{R}_0+\mathbf{K},\mathbf{y}_s)-\vartheta_{qs}(\mathbf{R}_0+\mathbf{K},\mathbf{y}_s)}{4}+\vartheta_{ql}(\mathbf{R}_0,\mathbf{s},\mathbf{y}_s)+\vartheta_{ql}(\mathbf{K},\mathbf{c}_0,\mathbf{y}_s)+\mathbf{s}\odot(\mathbf{E}_m(\mathbf{c}_0\odot\mathbf{y}_s))+$$
$$\mathbf{c}_0\odot(\mathbf{E}_m(\mathbf{s}\odot\mathbf{y}_s)),\quad v(\mathbf{y}_s)=\mathrm{E}[f_I^2(\boldsymbol{\theta})]=\mu_{qs}(\mathbf{K},\mathbf{y}_s)+\mathbf{s}^T\left((\mathbf{y}_s\mathbf{y}_s^T)\odot\mathbf{E}_m+\mathbf{I}_N\right)\mathbf{s}+2\mu_{ql}(\mathbf{K},\mathbf{s},\mathbf{y}_s),$$
$$\vartheta_v=\nabla_{\mathbf{y}_s}v(\mathbf{y}_s)=\vartheta_{qs}(\mathbf{K},\mathbf{y}_s)+2\mathbf{s}\odot(\mathbf{E}_m(\mathbf{s}\odot\mathbf{y}_s))+2\vartheta_{ql}(\mathbf{K},\mathbf{s},\mathbf{y}_s). \tag{31}$$

GD algorithm next:

$$\mathbf{y}_s^{(t+\frac{1}{2})}=\mathbf{y}_s^{(t)}-\beta^{(t)}\nabla_{\mathbf{y}_s}\mathcal{J}_l(\mathbf{y}_s^{(t)}), \tag{34}$$

$$\mathbf{y}_s^{(t+1)}\in\min_{\mathbf{v}_y\in[-1,+1]^N}\|\mathbf{v}_y-\mathbf{y}_s^{(t+\frac{1}{2})}\|_2, \tag{35}$$

where $\mathbf{y}_s^{(t)}=2\mathbf{p}_s^{(t)}-\mathbf{1}$ is the transformed probability vector at the t-th iteration, $\beta^{(t)}$ is the step-size and $\nabla_{\mathbf{y}}\mathcal{J}_l(\mathbf{y}_s^{(t)})$ is the gradient of the l-th order Taylor approximation of the true expectation where $l\in\{1,2\}$. The steps (34) and the (35) are considered gradient step and projection step, respectively. For our box constraints, the projection turns out to be clipping the vector $\mathbf{y}_s^{(t+\frac{1}{2})}$ to -1 and $+1$. We also use Armijo-Goldstein (AG) line search [44] to find a good step-size while avoiding saddle points due to its diminishing nature [45]. Complete details of the GD approach are shown in Algorithm 1. Note that, according to the Remark 1, a feasible discrete $\boldsymbol{\theta}$ is also a feasible $\mathbf{x}$ and corresponds to the degenerate PDF itself that generates $\boldsymbol{\theta}$. So, we find the vector that aligns the phases of the reflected signals with the phase of the direct signal:

$$\varphi_n^{\text{init}}=e^{-j\left(\arg(\mathbf{h}_{c_0})_n-\arg(h_{d_0})\right)},\quad\forall n=1,2,\dots,N, \tag{36}$$

where $(\mathbf{h}_{c_0})_n$ denotes the n-th element of $\mathbf{h}_{c_0}$ and project it to $\{-1,+1\}$ for a feasible $\mathbf{y}_{\text{init}}$. After the projected gradient descent, we sample G feasible solutions and choose the best one. The complete procedure is described in Algorithm 1. Note that the numerical results associated with the case studies will be discussed in the next section.

B. Overhead-aware Rate and EE maximization in an RIS-aided system

In order to tackle another canonical setting, we now focus on the RIS sub-problems where configuring the RIS to optimize the chosen performance metric requires finding the optimal number of reflecting elements N_{opt} first. However, this is only possible for simpler objective forms [46]. To address this limitation, we introduce a comprehensive stochastic sampling approach that optimizes more complex objective forms, including rate and EE while taking interference into account, circumventing the explicit calculation of N_{opt}.

1) System Model: Our system model is solely dictated by the signal model in (25). Along with that, we include the overhead and power consumption models developed in [46], [47]. We also assume that each RIS element has the ability to turn off or $\boldsymbol{\theta}\in\{-1,0,+1\}^N$. Note that, this allows us to avoid explicit derivation of N_{opt}. We also assume that the estimated channels are reliable. Now, we define the rate of the system considering interference and channel estimation overhead below:

$$R(\boldsymbol{\theta})=\left(1-\frac{T_E(\|\boldsymbol{\theta}\|_0)+T_F(\|\boldsymbol{\theta}\|_0)}{T}\right)\times$$
$$B\log_2\left(1+\frac{\beta_0|h_{d_0}+\mathbf{h}_0^H\mathrm{diag}(\boldsymbol{\theta})\mathbf{f}_0|^2}{\sum\limits_{i=1}^{N_I}\beta_i|h_{d_i}+\mathbf{h}_i^H\mathrm{diag}(\boldsymbol{\theta})\mathbf{f}_i|^2+\sigma_w^2}\right), \tag{37}$$

where the bandwidth is denoted by B, the noise variance is $\sigma_w^2=BN_0$, N_0 is the noise power spectral density, total duration of the time slot is denoted by T, $T_E(\|\boldsymbol{\theta}\|_0)$ denotes the time taken to estimate the channels, and $T_F(\|\boldsymbol{\theta}\|_0)$ is the feedback duration of the RIS configuration. The channel estimation time and feedback duration time are dependent on the number of RIS elements $\|\boldsymbol{\theta}\|_0$ and are expressed next,

$$T_E(\|\boldsymbol{\theta}\|_0)=T_0(\|\boldsymbol{\theta}\|_0+1),$$
$$T_F(\|\boldsymbol{\theta}\|_0)=\frac{\|\boldsymbol{\theta}\|_0 b_F}{B_F\log\left(1+p_F\left|h_F\right|^2/(N_0B_F)\right)}, \tag{38}$$

where T_0 is the duration of each pilot tone, and $b_F=2$ is the number of bits used to represent the states of each RIS element. Additionally, B_F, p_F, and h_F refer to the communication bandwidth, transmit power, and effective channel, respectively, in the feedback phase. Subsequently, the total power consumption can be expressed as

$$P_{tot}(\|\boldsymbol{\theta}\|_0)=P_E(\|\boldsymbol{\theta}\|_0)+\left(1-\frac{T_E(\|\boldsymbol{\theta}\|_0)}{T}\right)\mu p+$$
$$\frac{T_F(\|\boldsymbol{\theta}\|_0)}{T}\left(\mu_F p_F-\mu p\right)+\|\boldsymbol{\theta}\|_0P_{c,n}+P_{c,0}, \tag{39}$$

where $P_E(\|\boldsymbol{\theta}\|_0)=\frac{P_0T_E(\|\boldsymbol{\theta}\|_0)}{T}$ is the power consumption in the channel estimation phase, P_0 denotes the power of each pilot tone, p is the maximum transmit power in the data transmission phase, $P_{c,n}$ is the power required to operate each RIS element and $P_{c,0}$ is the static hardware power for the remaining system components. Additionally, $\frac{1}{\mu}$ and $\frac{1}{\mu_F}$ denote the transmit amplifier efficiency in the data transmission and feedback phase, respectively. Finally, the EE of the system is defined by,

$$EE(\boldsymbol{\theta})=\frac{R(\boldsymbol{\theta})}{P_{tot}(\|\boldsymbol{\theta}\|_0)}. \tag{40}$$

In this system model, the number of RIS elements is chosen to be $N=\min(N_{max},N_0)$, where N_{max} is a parameter and N_0 is the maximum integer for which $T_E(N_0)+T_F(N_0)<T$. This condition assures that the rate is realistic.

2) RIS optimization: The general optimization problem can be expressed as,

$$\min_{\boldsymbol{\theta}\in\{-1,0,+1\}^N}\quad\mathcal{J}(\boldsymbol{\theta}), \tag{41}$$

Algorithm 2: Stochastic sampling approach to optimize $\mathbf{r} \in \{\mathbf{p}, \mathbf{q}\}$ for the sub-problems in Algorithm 3.

```
Input:  System parameters and channels, $\mathbf{r}_{\text{init}}$, $t_{max}$,
        $\epsilon_t$, $\beta_s$, $N_e$, $b_m$
Output: $\mathbf{r}^*$
Initialize $t = 1$, $\delta_{SGD} = 1$, and $\mathbf{r}^{(t)} = \mathbf{r}_{\text{init}}$.
Define $\mathbf{s} = \begin{cases} \mathbf{p}, & \mathbf{r} = \mathbf{q} \\ \mathbf{q}, & \mathbf{r} = \mathbf{p} \end{cases}$
while $\delta_{SGD} \leq \epsilon_t$ and $t \leq t_{max}$
 do
    Calculate $\hat{b}_{\mathbf{r}}^*$ from Lemma 5.
    Calculate $\tilde{\mathbf{g}}_{\mathbf{r}}$ from (45).
    $\mathbf{r}^{(t+\frac{1}{2})} = \mathbf{r}^{(t)} - \beta_s \tilde{\mathbf{g}}_{\mathbf{r}}$.
    $\mathbf{r}^{(t+1)} \in \min_{\mathbf{v}_r \in [0,1]^N} \|\mathbf{v}_r - \mathbf{r}^{(t+\frac{1}{2})}\|_2$ such that
    $\mathbf{r} \leq \mathbf{1}_N - \mathbf{s}$.
    $t = t + 1$.
    $\delta_{SGD} = \|\mathbf{r}^{(t+1)} - \mathbf{r}^{(t)}\|_2^2$.
$\mathbf{r}^* = \mathbf{r}^{(t+1)}$.
```

Algorithm 3: SSA-T (Based on the BCD framework)

```
Input:  System parameters and channels, $\epsilon$, $G_s$
Output: $\boldsymbol{\theta}^*$
Initialize $\mathbf{q}^*$ with a random vector, $i = 0$, $\gamma_0 = 0$,
$\Delta = \epsilon + 1$, and $\bar{\boldsymbol{\theta}}_0$ with all zeros.
while $\Delta > \epsilon$ do
    Obtain $\mathbf{p}^*$ from Algorithm 2 with fixed $\mathbf{q}^*$.
    Obtain $\mathbf{q}^*$ from Algorithm 2 with fixed $\mathbf{p}^*$.
    Generate $G_s$ samples of $\boldsymbol{\theta}$ from the obtained $\mathbf{p}^*$
    and $\mathbf{q}^*$.
    Set $\gamma_{i+1} = \frac{1}{G_s} \sum_{g=1}^{G_s} \mathcal{J}(\boldsymbol{\theta}\{g\})$ and $\bar{\boldsymbol{\theta}}_{i+1} = \boldsymbol{\theta}\{g^*\}$,
    where $g^*$ is the index of the random sample that
    provides the best objective value.
    if $\gamma_{i+1} \leq \gamma_i$ then $\bar{\boldsymbol{\theta}}_{i+1} = \bar{\boldsymbol{\theta}}_i$ ;
    Evaluate $\Delta = |\gamma_{i+1} - \gamma_i| / \gamma_i$.
    $i = i + 1$.
$\boldsymbol{\theta}^* = \bar{\boldsymbol{\theta}}_{i-1}$.
```

where $\mathcal{J}(\boldsymbol{\theta})$ can be $-R(\boldsymbol{\theta})$ or $-EE(\boldsymbol{\theta})$. As the objective function is non-smooth and non-convex due to the presence of the interference term and L0 norm, we resort to the Lemma 1 and reformulate the problem below:

$$\begin{aligned} \min_{\mathbf{p}, \mathbf{q} \in (0,1)^N} \quad & \mathrm{E}_{\boldsymbol{\theta} \sim \mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})} \left[\mathcal{J}(\boldsymbol{\theta}) \right], \\ \text{s.t.} \quad & \mathbf{p} + \mathbf{q} \leq \mathbf{1}_N. \end{aligned} \tag{42}$$

We assume that the n-th element of $\boldsymbol{\theta}$ or θ_n is an independent categorical random variable and can take the value $+1$ with probability q_n that denotes the n-th entry of $\mathbf{q}$ and -1 with probability p_n that denotes the n-th entry of $\mathbf{p}$. The joint PDF for the ternary random vector can be expressed as:

$$\mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q}) = \prod_{n=1}^{N} p_n^{\frac{\theta_n(\theta_n - 1)}{2}} q_n^{\frac{\theta_n(\theta_n + 1)}{2}} (1 - p_n - q_n)^{1 - \theta_n^2}. \tag{43}$$

The presence of coupled optimization variables in (42) complicates the problem. To circumvent this, we implement the block coordinate descent (BCD) framework. This method decouples the problem with two coupled variable sets into two tractable sub-problems, each addressing a single set of variables while considering the other fixed. Not only does this approach simplify the problems, but it is also inspired by the ties between Dykstra's algorithm for projections onto intersections of convex sets and BCD [48]. Ultimately, the SSA is used to resolve the sub-problems emerging from the BCD structure, as illustrated in the previous case study.

We start by taking the gradient of the objective function in (41) assuming $\mathbf{q}$ is fixed. Note that, all the following derivations can easily be derived when $\mathbf{p}$ is fixed by substituting $\nabla_{\mathbf{p}}$ with $\nabla_{\mathbf{q}}$ and are not derived explicitly. However, we will provide those results in the appropriate lemmas.

$$\begin{aligned} \mathbf{g}_{\mathbf{p}} =& \nabla_{\mathbf{p}} \mathrm{E}\left[\mathcal{J}(\boldsymbol{\theta})\right] \overset{(a)}{=} \sum_{k=1}^{3^N} \mathcal{J}(\boldsymbol{\theta}\{k\}) \nabla_{\mathbf{p}} \mathbb{P}_E(\boldsymbol{\theta}\{k\}|\mathbf{p},\mathbf{q}) \\ &\overset{(b)}{=} \sum_{k=1}^{3^N} \left(\mathcal{J}(\boldsymbol{\theta}\{k\}) \nabla_{\mathbf{p}} \log \mathbb{P}_E(\boldsymbol{\theta}\{k\}|\mathbf{p},\mathbf{q})\right) \mathbb{P}_E(\boldsymbol{\theta}\{k\}|\mathbf{p},\mathbf{q}) \\ &\overset{(c)}{=} \mathrm{E}\left[\mathcal{J}(\boldsymbol{\theta}) \nabla_{\mathbf{p}} \log \mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})\right], \end{aligned} \tag{44}$$

where (a) comes from the definition of expectation, $\boldsymbol{\theta}\{k\}$ denotes the k-th possible combination out of the possible 3^N in an arbitrary indexing order, (b) comes from the identity $\nabla_{\mathbf{p}} \log \mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q}) = \frac{1}{\mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})} \nabla_{\mathbf{p}} \mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})$, and (c) converts the summation into expectation. The MC approximation of this gradient for a stochastic optimization approach is:

$$\hat{\mathbf{g}}_{\mathbf{p}} = \frac{1}{N_e} \sum_{j=1}^{N_e} \mathcal{J}(\boldsymbol{\theta}\{j\}) \nabla_{\mathbf{p}} \log \mathbb{P}_E(\boldsymbol{\theta}\{j\}|\mathbf{p},\mathbf{q}), \tag{45}$$

where N_e is the number of samples used. For completeness and to reduce the variance of this estimator without increasing N_e drastically, we introduce a baseline $b_{\mathbf{p}}$ in the objective function. Such an estimator has the following form,

$$\begin{aligned} \tilde{\mathbf{g}}_{\mathbf{p}} =& \frac{1}{N_e} \sum_{j=1}^{N_e} \left(\mathcal{J}(\boldsymbol{\theta}\{j\}) - b_{\mathbf{p}}\right) \nabla_{\mathbf{p}} \log \mathbb{P}_E(\boldsymbol{\theta}\{j\}|\mathbf{p},\mathbf{q}) \\ =& \hat{\mathbf{g}}_{\mathbf{p}} - b_{\mathbf{p}} \mathbf{d}_{\mathbf{p}}, \end{aligned} \tag{46}$$

where $\mathbf{d}_{\mathbf{p}} = \frac{1}{N_e} \sum_{j=1}^{N_e} \nabla_{\mathbf{p}} \log \mathbb{P}_E(\boldsymbol{\theta}\{j\}|\mathbf{p},\mathbf{q})$. Note that $\mathrm{E}[\mathbf{d}_{\mathbf{p}}] = 0$ as the following results stands:

$$\begin{aligned} \mathrm{E}[\nabla_{\mathbf{p}} \log \mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})] =& \mathrm{E}\left[\frac{1}{\mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})} \nabla_{\mathbf{p}} \mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})\right] \\ &\overset{(a)}{=} \nabla_{\mathbf{p}} \sum_{k=1}^{3^N} \mathbb{P}_E(\boldsymbol{\theta}\{k\}|\mathbf{p},\mathbf{q}) = 0, \end{aligned} \tag{47}$$

where (a) comes from writing out the expectation in a summation. Using this result, we can also show that both estimators are also unbiased and $\mathrm{E}[\tilde{\mathbf{g}}_{\mathbf{p}}] = \mathrm{E}[\hat{\mathbf{g}}_{\mathbf{p}}] = \mathbf{g}_{\mathbf{p}}$. In the next lemma, we include the key gradient results for the ternary random vector that is instrumental in the stochastic sampling approach.

Lemma 4. *The gradient of the log of joint PDF with respect* *to* $\mathbf{p}$ *and* $\mathbf{q}$ *are:*

$$\begin{aligned}&\nabla_{\mathbf{p}} \log \mathbb{P}_E(\boldsymbol{\theta}\{j\}|\mathbf{p},\mathbf{q})\\ &= \sum_{n=1}^{N}\left(\frac{\theta_n\{j\}(\theta_n\{j\}-1)}{2p_n} + \frac{(\theta_n\{j\}^2-1)}{1-p_n-q_n}\right)\mathbf{e}_n,\end{aligned} \tag{48}$$

$$\begin{aligned}&\nabla_{\mathbf{q}} \log \mathbb{P}_E(\boldsymbol{\theta}\{j\}|\mathbf{p},\mathbf{q})\\ &= \sum_{n=1}^{N}\left(\frac{\theta_n\{j\}(\theta_n\{j\}+1)}{2q_n} + \frac{(\theta_n\{j\}^2-1)}{1-p_n-q_n}\right)\mathbf{e}_n,\end{aligned} \tag{49}$$

where $\mathbf{e}_n$ *is the* n*-th unit vector of length* N*.*

Proof: We start by substituting the joint PDF:

$$\begin{aligned}&\nabla_{\mathbf{p}} \log \mathbb{P}_E(\boldsymbol{\theta}\{j\}|\mathbf{p},\mathbf{q})\\ &= \sum_{n=1}^{N}\frac{\theta_n\{j\}(\theta_n\{j\}-1)}{2}\nabla_{\mathbf{p}}\log p_n\\ &+ (1-\theta_n\{j\}^2)\nabla_{\mathbf{p}}\log(1-p_n-q_n)\\ &= \sum_{n=1}^{N}\left(\frac{\theta_n\{j\}(\theta_n\{j\}-1)}{2p_n} + \frac{(\theta_n\{j\}^2-1)}{1-p_n-q_n}\right)\mathbf{e}_n.\end{aligned} \tag{50}$$

Similarly, the gradient $\nabla_{\mathbf{q}} \log \mathbb{P}_E(\boldsymbol{\theta}\{j\}|\mathbf{p},\mathbf{q})$ can be derived with ease. ■

With these important gradients available, we find the optimal baseline for the estimator defined in (46) in the next lemma.

Lemma 5. *The optimal baselines when with respect to* $\mathbf{p}$ *and* $\mathbf{q}$ *are*

$$\begin{aligned}b_{\mathbf{p}}^* &= \frac{N_e}{\sum\limits_{n=1}^{N}\frac{1}{p_n}+\frac{1}{1-p_n-q_n}}\mathrm{E}[\hat{\mathbf{g}}_{\mathbf{p}}^T\mathbf{d}_{\mathbf{p}}],\\ b_{\mathbf{q}}^* &= \frac{N_e}{\sum\limits_{n=1}^{N}\frac{1}{q_n}+\frac{1}{1-p_n-q_n}}\mathrm{E}[\hat{\mathbf{g}}_{\mathbf{q}}^T\mathbf{d}_{\mathbf{q}}],\end{aligned} \tag{51}$$

where $\hat{\mathbf{g}}_{\mathbf{q}}$ *and* $\mathbf{d}_{\mathbf{q}}$ *can be found by replacing the* $\nabla_{\mathbf{q}}$ *in place of* $\nabla_{\mathbf{p}}$ *in* (45) *and* (46)*, respectively.*

Proof: See Appendix D. ■

Remark 3. *We can also approximate* $\mathrm{E}[\hat{\mathbf{g}}_{\mathbf{p}}^T\mathbf{d}_{\mathbf{p}}]$ *by taking* b_m *batches of* N_e *data points and average them to get* $\hat{b}_{\mathbf{p}}^*$ *to use in the algorithm.*

Now we have all the information to develop the stochastic sampling approach for ternary random variables. The algorithm to solve the sub-problems is demonstrated in Algorithm 2 and the BCD architecture is illustrated in Algorithm 3. In the Algorithm 2, the entries of $\mathbf{r}_{\text{init}}$ are independent and identically distributed (i.i.d) with uniform distribution $\mathcal{U}(0, r_{max})$, where $0 < r_{max} \leq 1$. By choosing a small r_{max}, we control the initial sparsity of the solution.

C. Worst-case computational complexity discussion

In this subsection, we derive the worst-case computational complexities for the proposed algorithms in terms of big-O notation. However, we would like to note that the complexity of gradient descent-based algorithms cannot be trivially expressed in the big-O notation, as the number of iterations for convergence heavily depends on the initial point and cannot be precisely determined [49]. In the literature, the number of iterations is regarded as a parameter, and subsequently, the complexity is represented using big-O notation [50], [51]. In this subsection, we follow the same approach, while also preserving more terms in the big-O expression for a better comparison among the proposed algorithms. Building upon the previous discussion, The algorithms are based on five fundamental operations: gradient calculation, descent-projection, inner looping, outer looping, and sampling. The descent-projection operation has a complexity of $O(N)$ across all algorithms. Regarding the inner looping operation, we need I_{E_1}, I_{E_2}, I_1, and I_2 iterations for the gradient descent algorithms to converge for E-GD using first and second-order Taylor approximations, and for the stochastic sampling approaches with binary and ternary variables respectively. There is typically no need for outer looping iterations except for the ternary variable stochastic sampling method due to the BCD framework. In this case, we assume that I_{BCD} iterations are needed to achieve convergence.

1) E-GD with first-order Taylor series approximation: For this algorithm, the gradient calculation step is primarily dictated by the matrix multiplications inherent in the quadratic forms of (32), with a complexity of $O(N^2)$. The sampling step adds an $O(GN^2)$ complexity due to G evaluations of the objective function, making the total complexity $O(I_{E_1}(N^2 + N) + GN^2)$.

2) E-GD with second-order Taylor series approximation: For this variant of the algorithm, the gradient calculation step is primarily affected by the matrix multiplications required for computing the matrix $\mathbf{U}$ as per Theorem 2, and has a complexity of $O(N^4)$. The other steps share the same complexities as the first-order version, yielding a total complexity of $O(I_{E_2}(N^4 + N) + GN^2)$.

3) Stochastic sampling for binary variables: The first step of this algorithm, the gradient estimator calculation, is dominated by the N_{ens} objective evaluations resulting in a complexity of $O(N_{ens}N^2)$. The sampling step carries a complexity of $O(G_sN^2)$ due to G_s objective function evaluations, which makes the overall complexity $O(I_1(N_{ens}N^2 + N) + G_sN^2)$.

4) Stochastic sampling for ternary variables: The main differences between this algorithm and the binary variant lie in the objective function evaluation, which has a complexity of $O(N^2+N)$ due to the additional L0 norm calculation. This yields a total complexity of $O(I_{BCD}I_2(N_{ens}(N^2+N)+N)+G_sN^2)$.

IV. Simulation Results

For our simulation results, we focus on a canonical (and perhaps most practically relevant) RIS scenario where RIS can significantly enhance performance: the creation of virtual line-of-sight (LoS) links when direct paths are obstructed, as highlighted in [14], [21], [52]. We maintain this assumption throughout our simulation. We also consider that we operate in a high interference regime where one interferer exists with average power similar to our user. This also highlights the ability of our developed algorithms to cope with high interference. The common simulation parameters used in both the cases are $\beta_i = p\delta_{PL}$, p is the transmit power, $\delta_{PL} = -110$ dB, $B = 5$ MHz, and $N_0 = -174$ dBm/Hz [46]. Additionally, all the channels are Rician distributed with the Rician factor of 4 while all the results in this section are averaged over 1000 independent channel realizations.

A. SINR maximization with RIS optimization

In this application, we compare the achievable capacity $C_{cap} = \log_2(1+\gamma)$ of our developed algorithms with the popular SDR method and the CPP methods. The transmit power p is considered to be 0 dBm. The algorithm parameters are $\varrho = 0.5$,, $\varepsilon = 0.0005$, $\epsilon_{th} = 10^{-2}$, $\beta_{\text{init}} = 0.01$, and $G = 100$. Our proposed first-order and second-order analytical GD algorithms are denoted by E-GD-1 and E-GD-2, respectively while our proposed stochastic sampling approach is denoted by SSA-B. The solution of the GD algorithm developed in [21] for continuous phase shifts projected to the discrete phase-shifts also acts like a baseline and is denoted by CPP-1. The CPP of the solution of (27) when the constraint is relaxed to be continuous is denoted by CPP-2. Note that, the only difference between E-GD-1 and CPP-2 is the final sampling step as the former treats the solution as a probability vector, and the latter projects it to $\{-1,+1\}$ for a solution. The CPP of the simple signal alignment scheme in (36) is denoted by SA. CPP methods are considered comparison baselines as they are more practical in terms of speed and are often used in the literature over the traditional branch-and-bound methods that do not scale well with the number of elements.

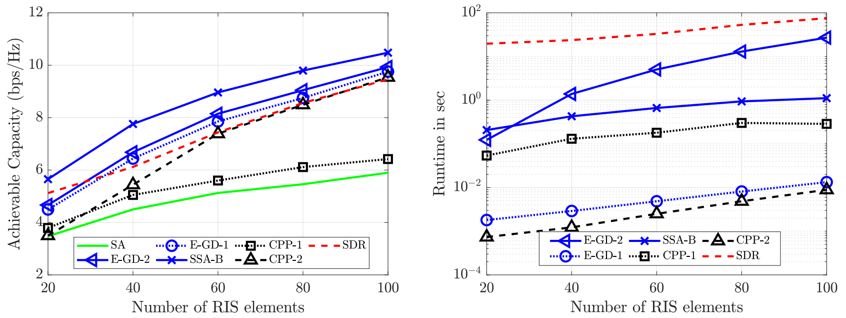
In Fig. 1a, we can observe that all the expectation-based algorithms perform better than the CPP algorithms, for all N, and the SDR for $N > 20$. Along with that, SSA-B outshines the expectation-based EGD algorithms that utilize approximations for expectation computation. The edge of SSA-B lies in its robust gradient estimates, derived without reliance on Taylor series approximations, hence providing more precise results. Moreover, the accuracy of SSA-B's gradient estimates can be enhanced by increasing the sample size, although this incurs a higher computational cost. Moreover, the scheme CPP-1 performs worse compared to CPP-2 due to its design for continuous RIS phase-shifts with unit-modulus constraints, which means its RIS optimization variable domain spans all angles from 0 to 2π corresponding to the set of all unit-modulus complex gains. In contrast, the domain of CPP-2 ranges from -1 to 1, making it closer to the original domain of $\{-1,+1\}$. This difference gets more prominent as the number of RIS elements grows and CPP-2 provides a sharper increase in achievable capacity than CPP-1.

In Fig. 1b, we plot the run-time for a single iteration of all the algorithms with varying numbers of RIS elements. These results are taken from the simulations needed to create Fig. 1a on a 3.6GHz Intel Core i7-4790 8-CPU system with 16GB RAM. From this plot, we note that the runtime of SSA-B is between the E-GD-1 and E-GD-2 methods while SDR is prohibitively slow. The runtime of our proposed E-GD-2 method is better than SDR but still slower than its first-order counterpart due to the complex gradient calculation. The overall performance of our analytical GD algorithms is dependent on the trade-off between the complexity of the gradient and the accuracy of the approximation for the expectation. These simulation results demonstrate the superiority of the expectation-based algorithms in discrete optimization problems providing important insights into such analytical expectation derivation.

B. Overhead-aware rate and EE maximization in an RIS-aided system

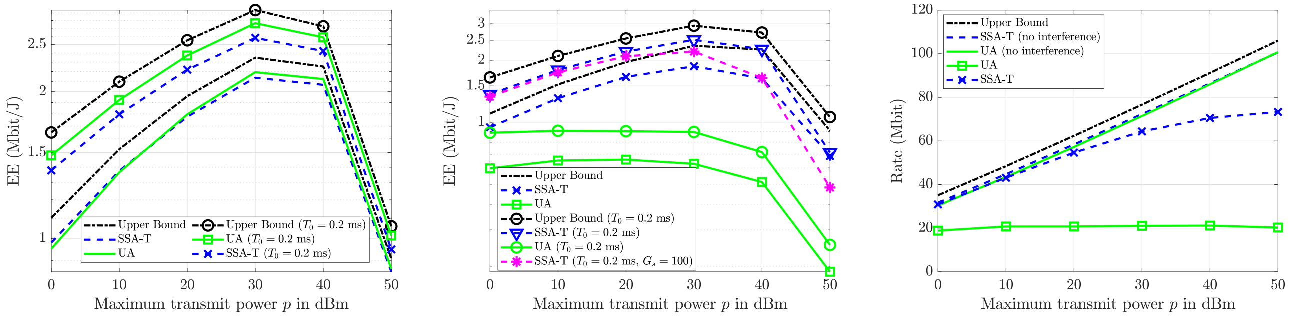
In this application, we maximize the rate and the EE of the system with our stochastic sampling approach. As a baseline, we compare it with the solution in [46] without interference projected to the discrete RIS phase-shifts. It should be noted that when interference exists, this baseline is no longer relevant because the unimodality required to compute the optimal number of RIS elements is dependent on the simple objective structure without interference. The simulation parameters are set according to [46]: $B_F = 1$ MHz, $P_{c,0} = 45$ dBm, $P_{c,n} = 10$ dBm, $\mu = \mu_F = 1$, $T_0 = 1$ ms, $p_F = 30$ dBm, $P_0 = 10$ dBm, $T = 100$ ms, and $N_{max} = 300$. The optimization algorithm parameters are, $\epsilon = 10^{-6}$, $N_e = 200$, $b_m = 10$, $r_{max} = 0.1$, $t_{max} = 300$, $G_s = 10000$, $\epsilon_t = 10^{-8}$, $\beta_s = 0.5$ for EE and $\beta_s = 0.01$ for rate optimization. In the simulation figures, the upper bound is calculated with the optimum continuous phase shifts without interference through the unimodal approach (UA) devised in [46]. The CPP of this approach also acts as a baseline and is denoted by UA while our algorithm is denoted by SSA-T or stochastic sampling approach for the ternary variable.

In Fig. 2a, we plot the average EE achieved with the transmit power when interference is not present. For $T_0 = 1$ ms, our algorithm performs very similarly to the unimodal approach. However, for $T_0 = 0.2$ ms, our algorithm achieves an EE that is 0.18 Mbit/J less at $p = 30$ dBm than the UA. While the UA method offers optimal results in the continuous RIS case where no interference is present - a scenario that can be viewed as a special instance of the general formulation with zero interference - it naturally extends well to the discrete case as well. In contrast, our proposed algorithm has a broader scope, demonstrating its capability to handle any form of objective function. Despite this versatility, the trade-off is a guarantee of optimality, hence the observed performance is completely expected. Our algorithm uniquely excels in managing interference and can adapt to any general objective form, an area where the UA method notably underperforms. Therefore, the superior performance of the UA approach in this specific case is anticipated, as it was designed precisely for such interference-free scenarios. This distinction underscores the unique use case of the stochastic sampling approach: when a reliable solution for the continuous problem exists, discrete projection may be sufficient. However, when the objective function becomes complex, even in its continuous form, our proposed algorithm shines, providing high-quality solutions where other methods might fall short. We can also observe that at the high transmit power region, the EE drops as power consumption dominates and our algorithm approaches the unimodal approach and the upper bound.



(a) Achievable capacity. **(b)** Runtime of algorithms.

Fig. 1: Comparison of the algorithms developed.



(a) EE vs p without interference. **(b)** EE vs p with interference. **(c)** Rate vs p.

Fig. 2: Performance with transmit power p in dBm.

	0 dBm	10 dBm	20 dBm	30 dBm	40 dBm	50 dBm
UA ($T_0 = 0.2$ms)	70.8	59.7	51.7	46.4	49.2	94.5
UA ($T_0 = 1$ms)	21.1	17.3	14.64	13	13.4	23.6
SSA-T ($T_0 = 0.2$ms)	52.7	47.4	46	46.1	47.9	40.7
SSA-T ($T_0 = 1$ms)	17.4	14.6	13.5	13.4	13.2	14.7

TABLE I: Average number of RIS elements to maximize EE in presence of interference with varying p.

In Fig. 2b, we plot the average EE achieved with the transmit power when interference is present. In situations where interference is present, as noted in Section III, the transmit power displayed on the x-axis represents the maximum transmit powers of both the user and the interferer. This mimics a non-cooperative scenario where each entity maximizes its own performance by transmitting at peak power simultaneously. We observe that our algorithms continue to perform close to the upper bound whereas the unimodal approach fails as expected. We can also observe the general trend of better performance with decreasing T_0. The reasoning is two-fold: a) we have more time to transmit data due to lower channel estimation time, and b) more overhead for RIS elements can be supported, resulting in the utilization of more RIS elements. This can be verified in Table I where we report the average number of RIS elements to maximize EE in presence of interference varying with the transmit power. We can also observe the effect of G_s in this figure. With a lower $G_s = 100$, the EE achieved is around 0.62 Mbit/J less than the default parameter $G_s = 10000$ case at $p = 30$ dBm. Finally, in Fig. 2c, we observe that in the rate maximization problem, the proposed algorithm performs closer to the upper bound than the UA approach, irrespective of interference. Without interference, the performance of SSA-T marginally exceeds that of UA. However, in the presence of interference, the advantage of SSA-T over UA becomes substantial. Furthermore, as we increase the maximum transmit power, it becomes clear that the achievable rate via our proposed approach reaches a saturation limit. This limit is imposed by the proportional increase in interference power which is not completely suppressed by the discrete RIS along with user transmit power.

V. Conclusion

In this paper, we developed a novel probabilistic reformulation technique for general discrete optimization problems. In particular, we interpret the discrete optimization variable as a categorical random vector and take expectations on the objective function along with any constraints present. We provide rigorous mathematical justification that the corresponding degenerate PDF of the unique optimal solution of an unconstrained problem is the unique optimal solution of the transformed problem and for a constrained problem, the primal solution of the transformed problem is bounded between the dual and primal solutions of the original problem implying that it is a relaxation of the original problem. However, if strong duality holds, the transformed problem provides the same objective value as the original constrained problem. We also explored a simple two-way partitioning problem to gain more insights into our reformulation such as its similarity to SDR, and capability to change the problem structure. We ultimately used this technique to tackle two canonical discrete RIS applications: a) SINR maximization, and b) overhead-aware rate and EE maximization. As demonstrated in our RIS applications, the reformulation allows for both stochastic and analytical interpretations of the original problems. For the SINR maximization problem, an analytical GD technique based on closed-form approximations for the expectation is proposed, while a stochastic sampling approach is proposed for both applications. The numerical results reveal that there is a fundamental trade-off between the complexity of the gradient and the accuracy of the approximation in our proposed analytical GD methods, and the expectation-based algorithms outperform the other algorithms evaluated. The simulation results also demonstrate that our proposed framework is very general and performs well for both rate and EE maximization problems without much change in the algorithm. In particular, we show that it performs at par with the algorithm specifically developed for the interference-free case when interference is not present and keeps performing well even when interference exists. We also explicitly calculate the worst-case computational complexities for our proposed algorithms. As the scope of this technique is very general, utilizing this technique to develop a more sophisticated projected gradient descent framework and a general methodology to deal with constrained problems are left as future work.

Appendix

A. Proof of Theorem 1

The expectation can be calculated by converting the matrix expressions into series sums as shown below,

$$\begin{aligned}\mathrm{E}[\mathbf{x}^T\mathbf{G}\mathbf{x}\mathbf{z}^T\mathbf{x}] = \quad & 2\sum_{i=1,i\neq j}^{n}\sum_{j=1,k=j}^{n}\mathrm{E}[x_i]G_{ij}z_j+ \\ & \mathrm{E}\left[\sum_{k=1}^{n}z_kx_k\sum_{i=1,i=j}^{n}x_i^2G_{ii}\right]+\mathrm{E}\left[\sum_{i\neq j\neq k,k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}x_ix_jx_kG_{ij}z_k\right] \\ & \overset{(a)}{=} 2\mathbf{y}^T\mathbf{G}_{wd}\mathbf{z}+\mathbf{z}^T\mathbf{y}\mathrm{Tr}(\mathbf{G})+\sum_{i\neq j\neq k,k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}y_iy_jy_kG_{ij}z_k.\end{aligned} \tag{52}$$

In step (a), we use the fact that $\mathbf{G}$ is real symmetric, $x_i^2 = 1$, and $\mathrm{E}[x_i] = y_i$. The third term can be expressed in the following form:

$$\sum_{i\neq j\neq k,k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}y_iy_jy_kG_{ij}z_k=\begin{bmatrix}y_1z_1\\y_2z_2\\\vdots\\y_nz_n\end{bmatrix}^T\begin{bmatrix}\mathbf{y}^T\mathbf{G}_1\mathbf{y}\\\mathbf{y}^T\mathbf{G}_2\mathbf{y}\\\vdots\\\mathbf{y}^T\mathbf{G}_n\mathbf{y}\end{bmatrix}, \tag{53}$$

where $\mathbf{G}_i$ denotes the matrix $\mathbf{G}$ with the i-th row and column set to zeros. In matrix form, this can be expressed as $\mathbf{1}^T\{(\mathbf{G}_{wd}\mathbf{Y}_{wd})\odot\mathbf{Y}_{wd}\}(\mathbf{y}\odot\mathbf{z})$, where $\mathbf{Y} = \mathbf{y}\mathbf{1}^T$. This completes the proof.

B. Proof of Theorem 2

We start this proof by expanding a generic quadratic term in (54). In (54), the matrix $\mathbf{U} = [\mathbf{I}_N \otimes (\mathbf{y}\odot\mathbf{y})^T]\mathbf{B}$, and the matrix $\mathbf{B}$ is defined through blocks as

$$\mathbf{B} = \begin{bmatrix}\mathbf{b}_{1,1},\dots,\mathbf{b}_{1,N}\\ \cdots,\cdots,\cdots,\\ \mathbf{b}_{N,1},\dots,\mathbf{b}_{N,N}.\end{bmatrix}, \tag{55}$$

where the i-th element of $\mathbf{b}_{k,j}$ is $\mathbf{b}_{k,j}^i = G_{wd_{ij}}G_{wd_{ki}}$.

Considering that the d-th term in the final expression without the numeric coefficient is denoted by $S_d(\mathbf{x})$, the above expression can be expressed as

$$q_s(\mathbf{x}) = S_1(\mathbf{x}) + S_2(\mathbf{x}) + 2S_3(\mathbf{x}) + 4S_4(\mathbf{x}) + S_5(\mathbf{x}). \tag{56}$$

Considering $x_i^2 = 1$, the second moment of a quadratic form can be expressed as,

$$\begin{aligned}&\mathrm{E}[q_s(\mathbf{x})] = \mathrm{Tr}(\mathbf{G})(\mathbf{y}^T\mathbf{G}_{wd}\mathbf{y}+\mathrm{Tr}(\mathbf{G}))+\mathrm{Tr}(\mathbf{G})\mathbf{y}^T\mathbf{G}_{wd}\mathbf{y}+ \\ &2\mathrm{Tr}(\mathbf{Z})+4\mathbf{y}^T\mathbf{Z}_{wd}\mathbf{y}+\sum_{l=1,i\neq j\neq k\neq l}^{N}\sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{N}y_iy_jy_ky_lG_{ij}G_{kl},\end{aligned} \tag{57}$$

where $\mathbf{Z} = \mathbf{G}_{wd}\mathbf{G}_{wd}^T$. This is readily found by taking expectation on (54). Note that the final term or $S_5(\mathbf{y})$ can be found from the following observation:

$$S_5(\mathbf{y}) = q_s(\mathbf{y}) - (S_1(\mathbf{y}) + S_2(\mathbf{y}) + 2S_3(\mathbf{y}) + 4S_4(\mathbf{y})). \tag{58}$$

The theorem is proved by combining the final two expressions.

C. Proof of Corollary 2

As the gradient of most of the terms can be trivially calculated [53], [54], we focus on the nontrivial gradient calculations here. In particular, the gradient of $\mathbf{y}^T\mathbf{U}_{wd}\mathbf{y}$ with respect to $\mathbf{y}$ is derived next. We can write the following expression due to the chain rule:

$$\nabla_{\mathbf{y}}\left(\mathbf{y}^T\mathbf{U}_{wd}\mathbf{y}\right)=\begin{bmatrix}\sum_k\sum_j y_jy_k\frac{\partial}{\partial y_1}\left((\mathbf{U}_{wd})_{jk}\right)\\ \sum_k\sum_j y_jy_k\frac{\partial}{\partial y_2}\left((\mathbf{U}_{wd})_{jk}\right)\\ \vdots\\ \sum_k\sum_j y_jy_k\frac{\partial}{\partial y_N}\left((\mathbf{U}_{wd})_{jk}\right)\end{bmatrix}+ (\mathbf{U}_{wd}+\mathbf{U}_{wd}^T)\mathbf{y}, \tag{59}$$

where $(\mathbf{U}_{wd})_{jk}$ is (j,k)-th element of the matrix $\mathbf{U}_{wd}$ and the matrix $\mathbf{U}$ can be expressed as

$$\mathbf{U} = \begin{bmatrix}(\mathbf{y}\odot\mathbf{y})^T\mathbf{b}_{1,1},\dots,(\mathbf{y}\odot\mathbf{y})^T\mathbf{b}_{1,N}\\ \cdots,\cdots,\cdots,\\ (\mathbf{y}\odot\mathbf{y})^T\mathbf{b}_{N,1},\dots,(\mathbf{y}\odot\mathbf{y})^T\mathbf{b}_{N,N}\end{bmatrix}.$$

With this formulation, the inner derivative is $\frac{\partial}{\partial y_i}\left((\mathbf{U}_{wd})_{jk}\right) = 2y_i(\mathbf{b}_{j,k})_i \quad \forall j \neq k$, where $(\mathbf{b}_{j,k})_i$ is the i-th element of the

$$
\begin{aligned}
q_s(\mathbf{x}) = (\mathbf{x}^T\mathbf{G}\mathbf{x})^2 = \sum_{l=1}^{N}\sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{N} x_i x_j x_k x_l G_{ij} G_{kl} \\
=&\sum_{k=1,k=l}^{N}\ \sum_{j=1}^{N}\sum_{i=1}^{N} x_i x_j x_k^2 G_{ij} G_{kk} + \sum_{l=1,k\neq l}^{N}\ \sum_{k=1}^{N}\ \sum_{i=1,j=i}^{N} x_i^2 x_k x_l G_{ii} G_{kl} + 2 \sum_{l=1,k\neq l}^{N}\ \sum_{k=1}^{N} x_k^2 x_l^2 G_{lk} G_{kl} \\
&+ 4 \sum_{k=1,i\neq j\neq k}^{N}\ \sum_{j=1}^{N}\ \sum_{i=1,i=l}^{N} x_i^2 x_j x_k G_{ij} G_{ki} + \sum_{l=1,i\neq j\neq k\neq l}^{N}\ \sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{N} x_i x_j x_k x_l G_{ij} G_{kl} \\
=&(\mathbf{x}\odot\mathbf{x})^T \mathrm{diag}(\mathbf{G})\mathbf{x}^T\mathbf{G}\mathbf{x} + (\mathbf{x}\odot\mathbf{x})^T \mathrm{diag}(\mathbf{G})\mathbf{x}^T\mathbf{G}_{wd}\mathbf{x} + 2(\mathbf{x}\odot\mathbf{x})^T\mathbf{G}_{wd}\odot\mathbf{G}_{wd}(\mathbf{x}\odot\mathbf{x}) \quad + 4\mathbf{x}^T\mathbf{U}_{wd}\mathbf{x} + \\
&\sum_{l=1,i\neq j\neq k\neq l}^{N}\ \sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{N} x_i x_j x_k x_l G_{ij} G_{kl}, \qquad (54)
\end{aligned}
$$

vector $\mathbf{b}_{j,k}$. Finally, the gradient can be written as,

$$
\nabla_{\mathbf{y}}\left(\mathbf{y}^T\mathbf{U}_{wd}\mathbf{y}\right) = 2\mathbf{y}\odot\begin{bmatrix} \sum\limits_{k\neq j}\sum\limits_{j} y_j y_k (\mathbf{b}_{j,k})_1 \\ \sum\limits_{k\neq j}\sum\limits_{j} y_j y_k (\mathbf{b}_{j,k})_2 \\ \vdots \\ \sum\limits_{k\neq j}\sum\limits_{j} y_j y_k \frac{\partial}{\partial y_N}(\mathbf{b}_{j,k})_N \end{bmatrix} +
$$

$$
(\mathbf{U}_{wd}+\mathbf{U}_{wd}^T)\mathbf{y} \overset{(a)}{=} 2\mathbf{y}\odot\mathbf{b}_s + (\mathbf{U}_{wd}+\mathbf{U}_{wd}^T)\mathbf{y}, \qquad (60)
$$

where (a) is obtained by some matrix manipulations and the definition of $\mathbf{b}_{j,k}$ vectors.

D. Proof of Lemma 5

We begin by calculating the total variance of the estimator $\tilde{\mathbf{g}}_{\mathbf{p}}$ below.

$$
\begin{aligned}
&\mathrm{var}(\tilde{\mathbf{g}}_{\mathbf{p}}) = \mathrm{E}[\tilde{\mathbf{g}}_{\mathbf{p}}^T\tilde{\mathbf{g}}_{\mathbf{p}}] - \mathrm{E}[\tilde{\mathbf{g}}_{\mathbf{p}}]^T\mathrm{E}[\tilde{\mathbf{g}}_{\mathbf{p}}] \\
&\overset{(a)}{=}\mathrm{E}[\hat{\mathbf{g}}_{\mathbf{p}}^T\hat{\mathbf{g}}_{\mathbf{p}}] - \mathrm{E}[\tilde{\mathbf{g}}_{\mathbf{p}}]^T\mathrm{E}[\tilde{\mathbf{g}}_{\mathbf{p}}] - 2b_{\mathbf{p}}\mathrm{E}[\hat{\mathbf{g}}_{\mathbf{p}}^T\mathbf{d}_{\mathbf{p}}] + b_{\mathbf{p}}^2\mathrm{E}[\mathbf{d}_{\mathbf{p}}^T\mathbf{d}_{\mathbf{p}}] \\
&\overset{(b)}{=}\mathrm{var}(\hat{\mathbf{g}}_{\mathbf{p}}) - 2b_{\mathbf{p}}\mathrm{E}[\hat{\mathbf{g}}_{\mathbf{p}}^T\mathbf{d}_{\mathbf{p}}] + b_{\mathbf{p}}^2\mathrm{var}(\mathbf{d}_{\mathbf{p}}), \qquad (61)
\end{aligned}
$$

where $\mathrm{var}(\mathbf{d}_{\mathbf{p}}) = \frac{1}{N_e^2}\sum\limits_{j=1}^{N_e}\mathrm{var}\left(\nabla_{\mathbf{p}}\log\mathbb{P}_E(\boldsymbol{\theta}\{j\}|\mathbf{p},\mathbf{q})\right)$ and as the variance does not depend on the j-th index, we can calculate the variance of the inner quantity next ignoring the index.

$$
\begin{aligned}
&\mathrm{var}\left(\nabla_{\mathbf{p}}\log\mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})\right) \\
&= \mathrm{E}\left[\left(\nabla_{\mathbf{p}}\log\mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})\right)^T\left(\nabla_{\mathbf{p}}\log\mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})\right)\right] \\
&= \mathrm{E}\left[\sum_{n=1}^{N}\left(\frac{\partial\log\mathbb{P}_E(\boldsymbol{\theta}|\mathbf{p},\mathbf{q})}{\partial p_n}\right)^2\right] \\
&= \sum_{n=1}^{N}\mathrm{E}\left[\left(\frac{\theta_n(\theta_n-1)}{2p_n} + \frac{(\theta_n^2-1)}{1-p_n-q_n}\right)^2\right] \\
&\overset{(a)}{=} \sum_{n=1}^{N}\left(\frac{1}{p_n} + \frac{1}{1-p_n-q_n}\right). \qquad (62)
\end{aligned}
$$

Using (62) in (61), we can write that,

$$
\mathrm{var}(\tilde{\mathbf{g}}_{\mathbf{p}}) = \mathrm{var}(\hat{\mathbf{g}}_{\mathbf{p}}) - 2b_{\mathbf{p}}\mathrm{E}[\hat{\mathbf{g}}_{\mathbf{p}}^T\mathbf{d}_{\mathbf{p}}] +
$$

$$
\frac{b_{\mathbf{p}}^2}{N_e}\sum_{n=1}^{N}\left(\frac{1}{p_n} + \frac{1}{1-p_n-q_n}\right), \qquad (63)
$$

where (a) is a result of the following observations: $\mathrm{E}[\theta_n^{2k_0}] = q_n + p_n$ and $\mathrm{E}[\theta_n^{2k_0+1}] = q_n - p_n$, where k_0 is a non-negative integer. Note that this is a convex quadratic expression in $b_{\mathbf{p}}$ and we can find the minimum by equating the derivative of this variance equal to zero. The optimal baseline is $b_{\mathbf{p}}^* = \frac{N_e}{\sum\limits_{n=1}^{N}\frac{1}{p_n}+\frac{1}{1-p_n-q_n}}\mathrm{E}[\hat{\mathbf{g}}_{\mathbf{p}}^T\mathbf{d}_{\mathbf{p}}]$ and the lemma is proved.

REFERENCES

[1] A. Pradhan and H. S. Dhillon, "Novel Probabilistic Reformulation Technique for Unconstrained Discrete RIS Optimization," in *Proc., IEEE PIMRC*, 2023, to appear.

[2] A. Pradhan, J. K. Devineni, H. S. Dhillon, and A. F. Molisch, "Intelligent Surface Optimization in Terahertz under Two Manifestations of Molecular Re-radiation," in *Proc., IEEE Globecom*, Dec. 2021.

[3] Ö. Özdogan, E. BjÖrnson, and E. G. Larsson, "Using Intelligent Reflecting Surfaces for Rank Improvement in MIMO Communications," in *Proc., IEEE Intl. Conf. on Acoustics, Speech, and Sig. Proc. (ICASSP)*, May 2020.

[4] Z. Huang, B. Zheng, and R. Zhang, "Transforming Fading Channel from Fast to Slow: IRS-Assisted High-Mobility Communication," in *Proc., IEEE Intl. Conf. on Commun. (ICC)*, June 2021.

[5] T. Jiang and W. Yu, "Interference Nulling Using Reconfigurable Intelligent Surface," *IEEE Journal on Sel. Areas in Commun.*, vol. 40, no. 5, pp. 1392–1406, Jan. 2022.

[6] A. Elzanaty, A. Guerra, F. Guidi, and M.-S. Alouini, "Reconfigurable Intelligent Surfaces for Localization: Position and Orientation Error Bounds," *IEEE Trans. on Signal Processing*, vol. 69, pp. 5386–5402, Aug. 2021.

[7] D.-R. Emenonye, H. S. Dhillon, and R. M. Buehrer, "RIS-Aided Localization under Position and Orientation Offsets in the Near and Far Field," *arXiv:2210.03599*, 2022.

[8] ——, "Fundamentals of RIS-aided Localization in the Far-field," *arXiv:2206.01652*, 2022.

[9] Y. Zhang, K. Shen, S. Ren, X. Li, X. Chen, and Z.-Q. Luo, "Configuring Intelligent Reflecting Surface With Performance Guarantees: Optimal Beamforming," *IEEE Journal of Sel. Topics in Signal Processing*, vol. 16, no. 5, pp. 967–979, May 2022.

[10] Z.-q. Luo, W.-k. Ma, A. M.-c. So, Y. Ye, and S. Zhang, "Semidefinite Relaxation of Quadratic Optimization Problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, April 2010.

[11] A. Attia, S. Leyffer, and T. S. Munson, "Stochastic Learning Approach for Binary Optimization: Application to Bayesian Optimal Design of Experiments," *SIAM Journal on Scientific Computing*, vol. 44, no. 2, pp. B395–B427, April 2022.

[12] J. Yuan, Y.-C. Liang, J. Joung, G. Feng, and E. G. Larsson, "Intelligent Reflecting Surface-Assisted Cognitive Radio System," *IEEE Trans. on Commun.*, vol. 69, no. 1, pp. 675–687, Oct. 2020.

[13] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, “Robust and Secure Wireless Communications via Intelligent Reflecting Surfaces,” *IEEE Journal on Sel. Areas in Commun.*, vol. 38, no. 11, pp. 2637–2652, July 2020.

[14] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, “Reconfigurable Intelligent Surfaces for Energy Efficiency in Wireless Communication,” *IEEE Trans. on Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, June 2019.

[15] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, “Multicell MIMO Communications Relying on Intelligent Reflecting Surfaces,” *IEEE Trans. on Wireless Commun.*, vol. 19, no. 8, pp. 5218–5233, May 2020.

[16] H. Ma, H. Zhang, N. Zhang, J. Wang, N. Wang, and V. C. M. Leung, “Reconfigurable Intelligent Surface With Energy Harvesting Assisted Cooperative Ambient Backscatter Communications,” *IEEE Wireless Commun. Letters*, vol. 11, no. 6, pp. 1283–1287, April 2022.

[17] S. Huang, Y. Ye, M. Xiao, H. V. Poor, and M. Skoglund, “Decentralized Beamforming Design for Intelligent Reflecting Surface-Enhanced Cell-Free Networks,” *IEEE Wireless Commun. Letters*, vol. 10, no. 3, pp. 673–677, Dec. 2020.

[18] Z. Yang, M. Chen, W. Saad, W. Xu, M. Shikh-Bahaei, H. V. Poor, and S. Cui, “Energy-Efficient Wireless Communications With Distributed Reconfigurable Intelligent Surfaces,” *IEEE Trans. on Wireless Commun.*, vol. 21, no. 1, pp. 665–679, July 2021.

[19] K. Feng, X. Li, Y. Han, S. Jin, and Y. Chen, “Physical Layer Security Enhancement Exploiting Intelligent Reflecting Surface,” *IEEE Commun. Letters*, vol. 25, no. 3, pp. 734–738, Dec. 2020.

[20] Y. Chen, M. Wen, E. Basar, Y.-C. Wu, L. Wang, and W. Liu, “Exploiting Reconfigurable Intelligent Surfaces in Edge Caching: Joint Hybrid Beamforming and Content Placement Optimization,” *IEEE Trans. on Wireless Commun.*, vol. 20, no. 12, pp. 7799–7812, June 2021.

[21] A. Pradhan, M. A. Abd-Elmagid, H. S. Dhillon, and A. F. Molisch, “Robust Optimization of RIS in Terahertz under Extreme Molecular Re-radiation Manifestations,” *IEEE Trans. on Wireless Commun.*, 2023, to appear.

[22] Y. Ma, Y. Shen, X. Yu, J. Zhang, S. Song, and K. B. Letaief, “A Low-Complexity Algorithmic Framework for Large-Scale IRS-Assisted Wireless Systems,” in *Proc., IEEE Globecom Workshops*, Dec. 2020.

[23] W. Cai, H. Li, M. Li, and Q. Liu, “Practical Modeling and Beamforming for Intelligent Reflecting Surface Aided Wideband Systems,” *IEEE Commun. Letters*, vol. 24, no. 7, pp. 1568–1571, April 2020.

[24] Q. Wu and R. Zhang, “Beamforming Optimization for Wireless Network Aided by Intelligent Reflecting Surface With Discrete Phase Shifts,” *IEEE Trans. on Commun.*, vol. 68, no. 3, pp. 1838–1851, Dec. 2019.

[25] X. Yu, D. Xu, and R. Schober, “Optimal Beamforming for MISO Communications via Intelligent Reflecting Surfaces,” in *Proc., IEEE SPAWC*, May 2020.

[26] J. Dai, Y. Wang, C. Pan, K. Zhi, H. Ren, and K. Wang, “Reconfigurable Intelligent Surface Aided Massive MIMO Systems With Low-Resolution DACs,” *IEEE Commun. Letters*, vol. 25, no. 9, pp. 3124–3128, July 2021.

[27] N. S. Perović, L.-N. Tran, M. Di Renzo, and M. F. Flanagan, “Achievable Rate Optimization for MIMO Systems With Reconfigurable Intelligent Surfaces,” *IEEE Trans. on Wireless Commun.*, vol. 20, no. 6, pp. 3865–3882, Feb. 2021.

[28] L. You, J. Xiong, D. W. K. Ng, C. Yuen, W. Wang, and X. Gao, “Energy Efficiency and Spectral Efficiency Tradeoff in RIS-Aided Multiuser MIMO Uplink Transmission,” *IEEE Trans. on Signal Processing*, vol. 69, pp. 1407–1421, Dec. 2020.

[29] J. Sanchez, E. Bengtsson, F. Rusek, J. Flordelis, K. Zhao, and F. Tufvesson, “Optimal, Low-Complexity Beamforming for Discrete Phase Reconfigurable Intelligent Surfaces,” in *Proc., IEEE Globecom*, Dec. 2021.

[30] R. Xiong, X. Dong, T. Mi, and R. C. Qiu, “Optimal Discrete Beamforming of Reconfigurable Intelligent Surface,” *arXiv:2211.04167*, 2022.

[31] K. Allemand, K. Fukuda, T. M. Liebling, and E. Steiner, “A Polynomial Case of Unconstrained Zero-one Quadratic Optimization,” *Mathematical Programming*, vol. 91, no. 1, pp. 49–52, May 2001.

[32] J. Luo, K. Pattipati, P. Willett, and F. Hasegawa, “Near-optimal Multiuser Detection in Synchronous CDMA using Probabilistic Data Association,” *IEEE Commun. Letters*, vol. 5, no. 9, pp. 361–363, Sep. 2001.

[33] A. Yellepeddi, K. J. Kim, C. Duan, and P. Orlik, “On Probabilistic Data Association for Achieving Near-exponential Diversity over Fading Channels,” in *Proc., IEEE Intl. Conf. on Commun. (ICC)*, June 2013.

[34] J. Staines and D. Barber, “Variational Optimization,” *arXiv:1212.4507*, 2012.

[35] R. T. Rockafellar, *Convex Analysis*. Princeton university press, 1997, vol. 11.

[36] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004.

[37] Z.-Q. Luo and W. Yu, “An Introduction to Convex Optimization for Communications and Signal Processing,” *IEEE Journal on Sel. Areas in Commun.*, vol. 24, no. 8, pp. 1426–1438, Aug. 2006.

[38] I. G. Rosenberg, “Brèves communications. 0-1 optimization and non-linear programming,” *RAIRO - Operations Research - Recherche Opérationnelle*, vol. 6, no. V2, pp. 95–97, 1972.

[39] Y. Xia, X. Sun, D. Li, and X. Zheng, “On The Reduction of Duality Gap in Box Constrained Nonconvex Quadratic Program,” *SIAM Journal on Optimization*, vol. 21, no. 3, pp. 706–729, 2011.

[40] M. X. Goemans and D. P. Williamson, “Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming,” *Journal of the ACM (JACM)*, vol. 42, no. 6, pp. 1115–1145, Nov. 1995.

[41] H. Kamoda, T. Iwasaki, J. Tsumochi, T. Kuki, and O. Hashimoto, “60-GHz Electronically Reconfigurable Large Reflectarray Using Single-Bit Phase Shifters,” *IEEE Trans. on Antennas and Propagation*, vol. 59, no. 7, pp. 2524–2531, May 2011.

[42] R. J. Williams, “Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning,” *Machine Learning*, vol. 8, no. 3–4, p. 229–256, May 1992.

[43] A. Stuart and K. Ord, *Kendall’s Advanced Theory of Statistics, Distribution Theory*. John Wiley & Sons, 2010, vol. 1.

[44] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, 1999.

[45] I. Panageas, G. Piliouras, and X. Wang, “First-order Methods Almost Always Avoid Saddle Points: The Case of Vanishing Step-sizes,” *Advances in Neural Info. Processing Systems*, vol. 32, 2019.

[46] A. Zappone, M. Di Renzo, X. Xi, and M. Debbah, “On the Optimal Number of Reflecting Elements for Reconfigurable Intelligent Surfaces,” *IEEE Wireless Commun. Letters*, vol. 10, no. 3, pp. 464–468, Oct. 2020.

[47] A. Zappone, M. Di Renzo, F. Shams, X. Qian, and M. Debbah, “Overhead-Aware Design of Reconfigurable Intelligent Surfaces in Smart Radio Environments,” *IEEE Trans. on Wireless Commun.*, vol. 20, no. 1, pp. 126–141, Sep. 2020.

[48] R. J. Tibshirani, “Dykstra’s Algorithm, ADMM, and Coordinate Descent: Connections, Insights, and Extensions,” *Advances in Neural Info. Processing Systems*, vol. 30, 2017.

[49] J. You, S. Jung, J. Seo, and J. Kang, “Energy-Efficient 3-D Placement of an Unmanned Aerial Vehicle Base Station With Antenna Tilting,” *IEEE Commun. Letters*, vol. 24, no. 6, pp. 1323–1327, June 2020.

[50] Z. Yang, J.-Y. Xia, J. Luo, S. Zhang, and D. Gündüz, “A Learning-Aided Flexible Gradient Descent Approach to MISO Beamforming,” *IEEE Wireless Commun. Letters*, vol. 11, no. 9, pp. 1895–1899, Sep. 2022.

[51] J.-C. Chen and Y.-C. Lin, “A Projected Gradient Descent Algorithm for Designing Low-Resolution Finite-Alphabet Equalizers in All-Digital Massive MU-MIMO Communication Systems,” *IEEE Access*, vol. 11, pp. 50 744–50 751, May 2023.

[52] A. Pradhan, J. K. Devineni, H. S. Dhillon, and A. F. Molisch, “Intelligent Surface Optimization in Terahertz under Two Manifestations of Molecular Re-radiation,” in *Proc., IEEE Globecom*, Dec. 2021.

[53] S. Laue, M. Mitterreiter, and J. Giesen, “Computing Higher Order Derivatives of Matrix and Tensor Expressions,” *Advances in Neural Info. Processing Systems*, vol. 31, 2018.

[54] ——, “A Simple and Efficient Tensor Calculus,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4527–4534.

Anish Pradhan (Graduate Student Member, IEEE) received the B.Tech degree in Electronics and Communication Engineering from the National Institute of Technology Durgapur, India, in 2018. He is currently pursuing a Ph.D. degree at the Bradley Department of Electrical and Computer Engineering, Virginia Tech, USA. His current research interests include MIMO, optimization theory, and reconfigurable intelligent surfaces.

Harpreet S. Dhillon (Fellow, IEEE) received the B.Tech. degree in electronics and communication engineering from IIT Guwahati in 2008, the M.S. degree in electrical engineering from Virginia Tech in 2010, and the Ph.D. degree in electrical engineering from the University of Texas at Austin in 2013.

After serving as a Viterbi Postdoctoral Fellow at the University of Southern California for a year, he joined Virginia Tech in 2014, where he is currently a Professor of electrical and computer engineering, the chair of the communications area, the Associate Director of Wireless@VT research group, and the Elizabeth and James E. Turner Jr. '56 Faculty Fellow. His research interests include communication theory, wireless networks, geolocation, and stochastic geometry. He is a fellow of AAIA and a Clarivate Analytics (Web of Science) Highly Cited Researcher. He has received six best paper awards including the 2014 IEEE Leonard G. Abraham Prize, the 2015 IEEE ComSoc Young Author Best Paper Award, and the 2016 IEEE Heinrich Hertz Award. He has also received Early Achievement Awards from three IEEE ComSoc Technical Committees, namely, the Communication Theory Technical Committee (CTTC) in 2020, the Radio Communications Committee (RCC) in 2020, and the Wireless Communications Technical Committee (WTC) in 2021. He was named the 2017 Outstanding New Assistant Professor, the 2018 Steven O. Lane Junior Faculty Fellow, the 2018 College of Engineering Faculty Fellow, and the recipient of the 2020 Dean's Award for Excellence in Research by Virginia Tech. His other academic honors include the 2008 Agilent Engineering and Technology Award, the UT Austin MCD Fellowship, the 2013 UT Austin WNCG leadership award, and the inaugural IIT Guwahati Young Alumni Achiever Award 2020. He has served as the TPC Co-chair for IEEE WCNC 2022 and as a symposium TPC Co-chair for many IEEE conferences. He has also served on the Editorial boards of several IEEE journals with his current appointments being on the Executive Editorial Committee for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and as a Senior Editor for IEEE WIRELESS COMMUNICATIONS LETTERS.