Novel Probabilistic Reformulation Technique for Unconstrained Discrete RIS Optimization

Anish Pradhan, *Student Member, IEEE,* and Harpreet S. Dhillon, *Fellow, IEEE*
The Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, USA
Email:{pradhananish1, hdhillon}@vt.edu.

***Abstract*—Determining optimal phases for a discrete reconfigurable intelligent surface (RIS) in RIS-aided wireless systems is known to be a challenging problem. This paper develops a novel probabilistic reformulation technique to transform such discrete optimization problems into continuous domain problems. The idea is to treat optimization variables as a categorical random vector with independent but non-identically distributed (i.n.i.d.) entries and replace the objective function with its expectation. In the unconstrained case, we rigorously establish the equivalence between the original problem's unique optimal solution and the corresponding degenerate probability density function (PDF) of the transformed problem. Furthermore, we derive key analytical moments and gradients associated with the quadratic form and binary random vectors that are useful in the optimization of RIS-aided wireless systems. In order to concretely demonstrate the benefits of the proposed technique, we reformulate a canonical discrete RIS-aided signal-to-interference-plus-noise ratio (SINR) maximization problem and solve the reformulated problem with the gradient descent (GD) technique. Our solution includes an analytical approach that relies on closed-form approximations for the expectation, incorporating moment results, and a stochastic sampling method based on a log-derivative gradient estimator. Numerical results show that our expectation-based algorithms outperform state-of-the-art conventional algorithms, thereby demonstrating the effectiveness of our approach.**

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

I. Introduction

RISs have recently gained significant attention in the field of wireless communication systems, offering the potential to improve signal quality, coverage, and capacity. These surfaces are composed of electronically controllable passive elements that can adaptively manipulate the propagation of electromagnetic waves, providing a more efficient and flexible means of communication. However, owing to hardware constraints and the necessity for lower implementation complexity, RIS elements typically can only provide discrete phase-shifts, which in turn renders the optimization of RIS configurations a demanding discrete problem. Not surprisingly, existing discrete RIS optimization algorithms often face scalability issues or rely on rounding procedures after solving relaxed continuous optimization problems, leading to significant performance degradation. Despite these limitations, continuous optimization algorithms with rounding procedures, such as closest point projection (CPP), remain popular due to their speed and practical feasibility, even though this *two-fold approximation* lacks mathematical rigor in some cases [1]. A notable exception is the semidefinite relaxation (SDR) method combined with the Gaussian randomization procedure, which exhibits multiple approximation accuracy results [2]. However, the high computational complexity of SDR hinders its scalability.

Driven by the need for scalable and rigorous algorithms that deliver quality solutions for discrete optimization problems, we introduce a probabilistic reformulation technique that transforms a general unconstrained discrete optimization problem into an equivalent continuous stochastic optimization problem in terms of the optimal solution. While our reformulation could potentially be integrated with other algorithms, we concentrate on incorporating it with the GD algorithm to capitalize on its speed and convergence properties. As a result, we present both the analytical GD and the stochastic sampling approach, which outperform traditional algorithms for a typical SINR maximization subproblem in canonical discrete RIS-aided wireless scenarios. While our proposed reformulation technique is inspired by discrete RISs, it is expected to find applications in other domains because of its fundamental nature.

A. Related Work and Motivation

Considering the focus of this paper, the two research directions that are highly relevant to the discussion include optimizing discrete RISs and exploring the intersection of probability and optimization.

To begin with the first direction, the authors of [3] investigated a discrete RIS-aided OFDM system and optimized the element-wise exhaustive search in an alternating optimization framework. RIS phase-shifts in a multiple-input-single-output (MISO) wireless network were optimized using branch-and-bound methods that scale exponentially with the number of RIS elements in [4] and [5]. These solutions work with both continuous and discrete RIS phase-shifts. The authors in [6] relaxed the discrete RIS phase-shift constraint to a continuous one and solved the optimization problem using minorize-maximization (MM) and accelerated gradient methods for the spectral and energy efficiency trade-off in a multiple-user multiple-input-multiple-output (MU-MIMO) setup. However, these methods are either not scalable or have poor performance due to the two-fold approximation. Recent efforts [1], [7], [8] provide scalable optimal discrete RIS beamforming optimization for single-input-single-output (SISO) systems based on the fixed-rank result of [9]. However, these strategies are

The support of U.S. National Science Foundation (Grants ECCS-2030215 and CNS-2225511) is gratefully acknowledged.

specialized for single-antenna scenarios and are not applicable for multi-antenna scenarios.

Venturing beyond the existing discrete RIS literature, let us move on to the second direction. The authors of [2] establish that the SDR formulation is a stochastic version of the original non-convex quadratic program. The authors of [10], [11] propose a probabilistic data association algorithm to tackle binary quadratic programs iteratively, achieving near-optimal results. Additionally, [12] presents a stochastic learning framework for binary optimization problems, treating the optimization variable as a random variable and taking expectation on it. While these works motivate further exploration at the intersection of probability and optimization, they do not offer a technique for solving general unconstrained discrete optimization problems, especially in the context of discrete RIS.

To bridge this gap, we develop a rigorous probabilistic technique that converts discrete optimization problems into the continuous probability domain. We showcase the effectiveness of this technique in a canonical discrete RIS-aided wireless system for an SINR maximization problem.

B. Contributions

Our work presents a comprehensive probabilistic reformulation technique for general unconstrained discrete optimization problems. This technique replaces the objective function with its expectations by re-imagining the optimization variables as categorical random variables with i.n.i.d. entries, thereby transforming the discrete optimization problems into continuous domain problems. More importantly, we rigorously establish the equivalence between a discrete unconstrained problem with a unique optimal solution and the reformulated problem in terms of the optimal point. We also derive various analytical moments and their gradients associated with the quadratic form and binary random vectors that are essential intermediate results for our GD algorithms. We apply this approach to an SINR maximization problem as a canonical case study, and we propose a stochastic sampling and an analytical GD approach to solve the reformulated problem. The analytical GD algorithm utilizes the first and second-order Taylor approximations of the expectation of the SINR, while the stochastic approach uses an estimator of the gradient. The numerical results demonstrate that our expectation-based algorithms outperform the state-of-the-art conventional approaches evaluated, offering significant advantages over existing techniques. These results confirm that when calculating the analytical expectation is challenging, employing an analytical approximation still leads to an improvement in performance. In situations where the analytical expectation is relatively simple to compute or can be estimated accurately using simpler analytical expressions, the solutions are expected to be close to optimal.

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 $\mathbf{x}$, $\mathrm{diag}\,(\mathbf{x})$ denotes a diagonal matrix with the entries of $\mathbf{x}$ as its diagonal elements. For a matrix $\mathbf{X}$, $\mathbf{X}^H$, $\mathbf{X}^T$, $\mathrm{Re}\,(\mathbf{X})$, $\mathrm{Tr}\,(\mathbf{X})$, $\mathrm{diag}(\mathbf{X})$, and $\mathbf{X} \succeq 0$ denote its conjugate transpose, transpose, real part, trace, diagonal elements as a vector, and positive semidefiniteness, respectively. Additionally, $\mathbf{X}_{wd} = \mathbf{X} - \mathrm{diag}(\mathbf{X})$. The expectation operation is denoted by $\mathrm{E}[\cdot]$, $\mathrm{var}(\cdot)$ 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 L2 norm is denoted by $\|\cdot\|_2$. 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 Unconstrained Discrete Optimization

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 entry 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 $x_i \in \mathcal{C}$, 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 summarized 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* $\mathbf{x}_{\mathrm{opt}}$*, then* $\mathbf{P}_{\mathrm{opt}} = \mathrm{Degen}(\mathbf{x}_{\mathrm{opt}})$ *is the unique optimal solution of* (3)*, where the* $\mathbf{P} = \mathrm{Degen}(\mathbf{x})$ *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 $\mathbf{P} = \mathrm{Degen}(\mathbf{x})$ which is the parameter matrix of n degenerate categorical distributions. From these arguments, it follows that $\Omega_{\mathbf{x}} \subseteq \Omega_{\mathbf{P}}$.

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

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

where $\mathbf{x}\{k\}$ denotes the k-th combination out of possible b^n combinations of $\mathbf{x}$. This stems from the observation that the expectation is nothing but a convex combination of all the possible values of $f(\mathbf{x})$. 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,

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

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

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. ∎

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

Many discrete RIS applications focus on binary phase-shift RIS $\{-1, +1\}$ for its operational simplicity. Analytical moments and gradients of binary random vectors are derived next for use in expectation-based optimization. We begin the discussion with the covariance matrix in the next section.

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

$$\mathrm{E}[\mathbf{x}\mathbf{x}^T] = (\mathbf{y}\mathbf{y}^T) \odot \mathbf{E}_m + \mathbf{I}_N, \quad (7)$$

where $\mathbf{E}_m$ is the all-one matrix with a hollow diagonal and $\mathbf{p}$ is defined similarly to (19)*.*

Now, we state the first moment and its gradient in Lemma 2 without proof due to its trivial nature.

Lemma 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 expectation and the gradient of a sum between a quadratic form and a linear form are*

$$\mu_{qf}(\mathbf{G}, \mathbf{z}, \mathbf{y}) = \mathrm{E}[\mathbf{x}^T \mathbf{G} \mathbf{x} + \mathbf{z}^T \mathbf{x}] = \mathbf{y}^T \mathbf{G}_{wd} \mathbf{y} + \mathrm{Tr}(\mathbf{G}) + \mathbf{z}^T \mathbf{y}, \quad (8)$$

$$\vartheta_{qf}(\mathbf{G}, \mathbf{z}, \mathbf{y}) = \nabla_{\mathbf{y}} \mathrm{E}[\mathbf{x}^T \mathbf{G} \mathbf{x}] = (\mathbf{G}_{wd} + \mathbf{G}_{wd}^T)\mathbf{y} + \mathbf{z}. \quad (9)$$

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*

$$\mu_{ql}(\mathbf{G}, \mathbf{z}, \mathbf{y}) = \mathrm{E}[\mathbf{x}^T \mathbf{G} \mathbf{x} \mathbf{z}^T \mathbf{x}] = 2\mathbf{y}^T \mathbf{G}_{wd} \mathbf{z} + \mathbf{z}^T \mathbf{y} \mathrm{Tr}(\mathbf{G}) + \mathbf{1}^T \{(\mathbf{G}_{wd} \mathbf{Y}_{wd}) \odot \mathbf{Y}_{wd}\}(\mathbf{y} \odot \mathbf{z}), \quad (10)$$

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

Proof: The proof begins by transforming the matrix expressions into a series of summations, taking into account various scenarios involving the relationships between the indices in the sums, such as when they are equal or distinct from one another. It then primarily leverages the properties of x_i^2 equaling 1. Subsequently, the expression is simplified and rewritten using a convenient matrix form. For detailed step-by-step proof, please refer to the journal version available as an arxiv preprint in [13]. ∎

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:*

$$\vartheta_{ql}(\mathbf{G}, \mathbf{z}, \mathbf{y}) = 2\mathbf{G}_{wd}\mathbf{z} + \mathbf{z}\mathrm{Tr}(\mathbf{G}) + ((\mathbf{G}_T^T \odot \mathbf{E}_m)\mathbf{y}) \odot \mathbf{z} + \mathrm{diag}(\mathbf{G}_T \mathrm{diag}(\mathbf{y} \odot \mathbf{z})\mathbf{E}_m) + (\mathbf{G}_T \odot \mathbf{E}_m)(\mathbf{y} \odot \mathbf{z}), \quad (11)$$

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*

$$\mu_{qs}(\mathbf{G}, \mathbf{y}) = \mathrm{E}[(\mathbf{x}^T \mathbf{G} \mathbf{x})^2] = \mathbf{y}^T (\mathbf{G}_s - \mathbf{F}(\mathbf{y})) \mathbf{y} + \mathrm{Tr}(\mathbf{G})^2 + 2\mathrm{Tr}(\mathbf{Z}) + (\mathbf{y}^T \mathbf{G} \mathbf{y})^2 - \mathbf{d}^T \mathbf{G}_g \mathbf{d}, \quad (12)$$

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

$$\mathbf{B} = \begin{bmatrix} \mathbf{b}_{1,1}, \ldots, \mathbf{b}_{1,N} \\ \cdots, \cdots, \cdots, \\ \mathbf{b}_{N,1}, \ldots, \mathbf{b}_{N,N}. \end{bmatrix}, \quad (13)$$

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

Proof: The proof involves expanding $(\mathbf{x}^T \mathbf{G} \mathbf{x})^2$ into a series of summations considering different cases of the sum indices. Most of these sums can be rearranged in a matrix form except for the case when all indices are different. This term can be found by investigating the term $(\mathbf{y}^T \mathbf{G} \mathbf{y})^2$. A detailed proof is available in the journal version available as an arxiv preprint in [13]. ∎

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:*

$$\vartheta_{qs}(\mathbf{G}, \mathbf{y}) = (\mathbf{G}_s + \mathbf{G}_s^T)\mathbf{y} + 2\mathbf{y}^T \mathbf{G} \mathbf{y} (\mathbf{G} + \mathbf{G}^T)\mathbf{y} - 2\mathbf{y}^T (\mathbf{G} + \mathbf{G}_{wd})\mathbf{y} (\mathrm{diag}(\mathbf{G}) \odot \mathbf{y}) - \mathbf{d}^T \mathrm{diag}(\mathbf{G})(\mathbf{G} + \mathbf{G}_{wd})\mathbf{y} - \mathrm{diag}(\mathbf{G})^T \mathbf{d} (\mathbf{G} + \mathbf{G}_{wd})^T \mathbf{y} - 2((\mathbf{G}_g + \mathbf{G}_g^T)\mathbf{d}) \odot \mathbf{y} - 8\mathbf{y} \odot \mathbf{b}_s - 4(\mathbf{U}_{wd} + \mathbf{U}_{wd}^T)\mathbf{y}, \quad (14)$$

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: The proof begins with utilizing the chain rule and writing out the gradient entry-wise. Then it primarily uses the block structure of $\mathbf{U}$ for the entries in the gradient vector. A complete proof can be found in the journal version available as an arxiv preprint in [13]. ∎

III. SINR Maximization with RIS Optimization

1) Signal model: Optimizing RIS phase-shifts in MIMO communication is challenging, especially for discrete RISs. To address this, we break down the problem into smaller, manageable sub-problems and focus on canonical forms found in the literature. We use a unified signal model next to represent various RIS-aided scenarios and sub-problems, such as device-to-device communication, cellular networks with antenna selection, and wireless communication with fixed receive beamformer vector [14]:

$$y_r = (h_{d_0} + \mathbf{h}_0^H \mathrm{diag}(\boldsymbol{\theta})\mathbf{f}_0)x_{s,0} + \sum_{i=1}^{N_I}(h_{d_i} + \mathbf{h}_i^H \mathrm{diag}(\boldsymbol{\theta})\mathbf{f}_i)x_{s,i} + w, \quad (15)$$

where y_r is the received signal from the transmitter (Tx) of interest (denoted by $i = 0$), h_{d_i} denotes the direct channel between the i-th Tx and receiver (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. 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.

2) System model: We consider a generic system model dictated by the signal model (15). We consider the RIS phase vector $\boldsymbol{\theta} = [\theta_1 \quad \theta_2 \ldots \theta_n \ldots \theta_N]^T$ with $\theta_n \in \{-1, +1\}$. 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,

$$\gamma = \frac{\beta_0 |h_{d_0} + \mathbf{h}_{c_0}^H \boldsymbol{\theta}|^2}{\sum_{i=1}^{N_I} \beta_i |h_{d_i} + \mathbf{h}_{c_i}^H \boldsymbol{\theta}|^2 + \sigma_w^2} = \frac{f_s(\boldsymbol{\theta})}{f_I(\boldsymbol{\theta})} = \frac{\boldsymbol{\theta}^T \mathbf{R}_0 \boldsymbol{\theta} + \mathbf{c}_0^T \boldsymbol{\theta}}{\boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta} + \mathbf{s}^T \boldsymbol{\theta}}, \quad (16)$$

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 $\mathbf{s} = \sum_{i=1}^{N_I} \mathbf{c}_i$.

3) RIS optimization: In this subsection, our objective is to maximize the SINR given in (16) while the RIS elements are discrete in nature. The optimization problem is described below:

$$\min_{\boldsymbol{\theta} \in \{-1,+1\}^N} \quad -\frac{f_s(\boldsymbol{\theta})}{f_I(\boldsymbol{\theta})}. \quad (17)$$

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 [14]. 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:

$$\min_{\mathbf{y}_s \in [-1,+1]^N} \quad -\mathrm{E}_{\boldsymbol{\theta} \sim \mathbb{P}_B(\boldsymbol{\theta}|\mathbf{p}_s)} \left[\frac{f_s(\boldsymbol{\theta})}{f_I(\boldsymbol{\theta})}\right], \quad (18)$$

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

$$\mathbb{P}_B(\boldsymbol{\theta}|\mathbf{p}_s) = \prod_{n=1}^{N} \left(\delta(\theta_n - 1)p_{s,n} + \delta(\theta_n + 1)(1 - p_{s,n})\right), \quad (19)$$

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 (18): a) stochastic sampling approach, and b) analytical GD approach. The former approach generally does not require an explicit expression of the gradient whereas the latter does. In the stochastic sampling approach, generally, an estimator of the gradient is used in the GD algorithm. Such a procedure has appeared in [12] for a binary random vector in the context of the Bayesian optimal design of experiments utilizing the standard combination of log-derivative trick and Monte Carlo (MC) sampling [15]. To our knowledge, the current paper is the first work to use this reformulation and the stochastic approach in the RIS context. So, we only delve into the analytical optimization approach in this case. In this analytical GD approach, calculating the direct expectation of a ratio of correlated random variables is difficult. So, we consider the Taylor series approximations of

Algorithm 1: E-GD

Input: $\mathbf{R}_i, \mathbf{c}_i, \varrho, \varepsilon, \epsilon_{th}, \beta_{\mathrm{init}}, G \ \forall i$
Output: $\bar{\boldsymbol{\theta}}_{i+1}$
Initialize $t = 1$, $\delta_{GD} = 1$, and $\mathbf{y}_s^{(t)} = \mathbf{y}_{\mathrm{init}}$.
while $\delta_{GD} \leq \epsilon_{th}$
do
  Initialize $\beta^{(1)} = \beta_{\mathrm{init}}, d_f = -1$.
  Calculate $\nabla_{\mathbf{y}_s} \mathcal{J}_l(\mathbf{y}_s^{(t)})$ from (22) or (23).
  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 $[-\mathbf{1}_N, +\mathbf{1}_N]$.
    $d_f = -\mathcal{J}_l(\mathbf{y}_s^{(t)}) - \varepsilon\beta^{(t)}\|\nabla_{\mathbf{y}_s}\mathcal{J}_l(\mathbf{y}_s^{(t)})\|_2^2 + \mathcal{J}_l(\mathbf{y}_{proj})$.
    $\beta^{(t)} = \varrho\beta^{(t)}$.
  $\mathbf{y}_s^{(t+\frac{1}{2})} = \mathbf{y}_s^{(t)} - \beta^{(t)} \nabla_{\mathbf{y}_s} \mathcal{J}_l(\mathbf{y}_s^{(t)})$.
  $\mathbf{y}_s^{(t+1)} \in \min_{\mathbf{v}_y \in [-1,+1]^N} \|\mathbf{v}_y - \mathbf{y}_s^{(t+\frac{1}{2})}\|_2$.
  $t = t + 1$.
  $\delta_{GD} = \|\mathbf{y}_s^{(t+1)} - \mathbf{y}_s^{(t)}\|_2^2$.
$\mathbf{p}_s = \frac{\mathbf{y}_s^{(t)}+1}{2}$.
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.
$\bar{\boldsymbol{\theta}}_{i+1} = \boldsymbol{\theta}_{best}$.

such an expectation [16]. Both the first-order approximation $\mathcal{J}_1(\mathbf{y}_s)$ and second-order approximation $\mathcal{J}_2(\mathbf{y}_s)$ are stated below:

$$\mathcal{J}_1(\mathbf{y}_s) = \frac{\mathrm{E}[f_s(\boldsymbol{\theta})]}{\mathrm{E}[f_I(\boldsymbol{\theta})]} = \frac{\mu_{qf}(\mathbf{R}_0, \mathbf{c}_0, \mathbf{y}_s)}{\mu_{qf}(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)},$$
$$\mathcal{J}_2(\mathbf{y}_s) = \mathcal{J}_1(\mathbf{y}_s) - \frac{\mathrm{E}[f_s(\boldsymbol{\theta}) f_I(\boldsymbol{\theta})]}{\mathrm{E}^2[f_I(\boldsymbol{\theta})]} + \frac{\mathrm{E}[f_I^2(\boldsymbol{\theta})]\mathrm{E}[f_s(\boldsymbol{\theta})]}{\mathrm{E}^3[f_I(\boldsymbol{\theta})]}. \tag{20}$$

The second-order approximation requires two additional expectations that are derived along with their gradients in (21). Using the definitions in (21), we can express the gradients of the Taylor series approximations as follows:

$$\nabla_{\mathbf{y}_s} \mathcal{J}_1(\mathbf{y}_s) = \frac{\vartheta_{qf}(\mathbf{R}_0, \mathbf{c}_0, \mathbf{y}_s) - \mathcal{J}_1(\mathbf{y}_s)\vartheta_{qf}(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}{\mu_{qf}(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}, \tag{22}$$

$$\nabla_{\mathbf{y}_s} \mathcal{J}_2(\mathbf{y}_s) = \nabla_{\mathbf{y}_s} \mathcal{J}_1(\mathbf{y}_s) - \frac{\vartheta_{cv}}{\mu_{qf}^2(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)} +$$
$$\mu_{qf}(\mathbf{R}_0, \mathbf{c}_0, \mathbf{y}_s) \left(\frac{\vartheta_v}{\mu_{qf}^3(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)} - \frac{3v(\mathbf{y}_s)\vartheta_{qf}(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}{\mu_{qf}^4(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)} \right) +$$
$$\frac{2c_v(\mathbf{y}_s)\vartheta_{qf}(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}{\mu_{qf}^3(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)} + \frac{v(\mathbf{y}_s)\vartheta_{qf}(\mathbf{R}_0, \mathbf{c}_0, \mathbf{y}_s)}{\mu_{qf}^3(\mathbf{K}, \mathbf{s}, \mathbf{y}_s)}. \tag{23}$$

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 GD algorithm next:

$$\mathbf{y}_s^{(t+\frac{1}{2})} = \mathbf{y}_s^{(t)} - \beta^{(t)} \nabla_{\mathbf{y}_s} \mathcal{J}_l(\mathbf{y}_s^{(t)}), \tag{24}$$

$$\mathbf{y}_s^{(t+1)} \in \min_{\mathbf{v}_y \in [-1,+1]^N} \|\mathbf{v}_y - \mathbf{y}_s^{(t+\frac{1}{2})}\|_2, \tag{25}$$

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 (24) and the (25) 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 [17] to find a good step-size while avoiding saddle points due to its diminishing nature [18]. Complete details of the GD approach are shown in Algorithm 1. Note that 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:

$$\varphi_n^{\text{init}} = e^{-j\left(\arg(\mathbf{h}_{c_0})_n - \arg(h_{d_0})\right)}, \quad \forall n = 1, 2, \ldots, N, \tag{26}$$

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 GD, 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.

IV. Simulation Results

In this section, we compare the performance of our proposed algorithms to the widely used SDR and CPP methods. The primary metrics for comparison are the achievable capacity, denoted as $C_{cap} = \log_2(1 + \gamma)$, and the runtimes of each algorithm. We set up a simulation scenario where a Tx communicates with an Rx, aided by an RIS. This scenario is inspired by the signal model (15) discussed in Section III. It is important to reiterate that although the Tx and Rxs may have multiple antennas, our focus on the RIS subproblem causes the mathematical formulation to resemble a point-to-point link. We concentrate on the scenario where direct paths are blocked, as this is when the RIS proves most useful. Moreover, we operate in a high interference regime, where an interferer with average power similar to our user is present. This also highlights the ability of our developed algorithms to cope with high interference. The simulation parameters used are $\beta_i = p\delta_{PL}$, $p = 0$ dBm is the transmit power, $\delta_{PL} = -110$ dB, $B = 5$ MHz, and $N_0 = -174$ dBm/Hz [19]. The algorithm parameters are $\varrho = 0.5$, $\varepsilon = 0.0005$, $\epsilon_{th} = 10^{-2}$, $\beta_{\text{init}} = 0.01$, and $G = 100$. 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. 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 [14] 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 (17) 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 (26) 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. 1, we can observe that all the proposed expectation-based algorithms perform better than the CPP algorithms, for all N, and the SDR for $N > 20$. In particular, the proposed stochastic sampling approach 'SSA-B' performs the best, while our proposed analytical GD algorithms dependent on the approximations of expectation perform slightly worse. The scheme 'CPP-1' performs worse than 'CPP-2' because the former was developed for continuous RIS phase-shifts with unit-modulus constraints whereas the domain of the latter is much smaller and closer to the original problem.

In Fig. 2, 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. 1 on a 3.6GHz Intel Core i7-4790 8-CPU system with 16GB RAM. From this plot, we note that the runtime of our proposed '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

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

$$\vartheta_{cv}=\nabla_{\mathbf{y}_s}c_v(\mathbf{y}_s)=\frac{\vartheta_{qs}(\mathbf{R}_0+\mathbf{K},\mathbf{y}_s)-\vartheta_{qs}(\mathbf{R}_0+\mathbf{K},\mathbf{y}_s)}{4}+\vartheta_{ql}(\mathbf{R}_0,\mathbf{s},\mathbf{y}_s)+\vartheta_{ql}(\mathbf{K},\mathbf{c}_0,\mathbf{y}_s)+\mathbf{s}\odot(\mathbf{E}_m(\mathbf{c}_0\odot\mathbf{y}_s))+$$

$$\mathbf{c}_0\odot(\mathbf{E}_m(\mathbf{s}\odot\mathbf{y}_s)),\quad v(\mathbf{y}_s)=\mathrm{E}[f_I^2(\boldsymbol{\theta})]=\mu_{qs}(\mathbf{K},\mathbf{y}_s)+\mathbf{s}^T\left((\mathbf{y}_s\mathbf{y}_s^T)\odot\mathbf{E}_m+\mathbf{I}_N\right)\mathbf{s}+2\mu_{ql}(\mathbf{K},\mathbf{s},\mathbf{y}_s),$$

$$\vartheta_v=\nabla_{\mathbf{y}_s}v(\mathbf{y}_s)=\vartheta_{qs}(\mathbf{K},\mathbf{y}_s)+2\mathbf{s}\odot(\mathbf{E}_m(\mathbf{s}\odot\mathbf{y}_s))+2\vartheta_{ql}(\mathbf{K},\mathbf{s},\mathbf{y}_s).\tag{21}$$

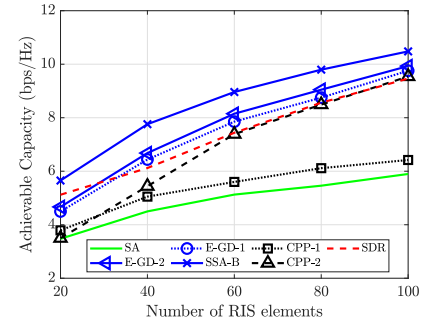

Fig. 1: Achievable capacity.

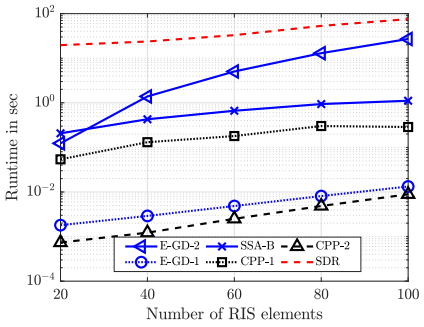

Fig. 2: Runtime of algorithms.

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 proposed expectation-based algorithms in discrete optimization problems.

V. Conclusion

In this paper, we proposed a novel probabilistic reformulation technique to tackle general unconstrained discrete optimization problems. Our approach treated the discrete optimization variable as a categorical random vector with i.n.i.d. entries and substituted the objective function with its expectation. We provided a rigorous mathematical proof demonstrating the equivalence between the unique optimal solution of an unconstrained problem and the corresponding degenerate PDF of the transformed problem. Moreover, we derived analytical moments and gradients associated with the quadratic form and binary random vectors, which served as essential intermediate results. We applied our method to a canonical discrete RIS-aided SINR maximization problem. To solve the reformulated problem, we proposed an analytical GD technique, based on closed-form Taylor series-based approximations for the expectation, and a stochastic sampling approach. Numerical results revealed that our expectation-based algorithms outperform other state-of-the-art conventional algorithms, highlighting the advantages of our approach. Although our focus has been on RIS applications, the proposed method is applicable to a wide range of other domains. Future research directions include developing a more advanced projected GD framework and investigating alternative gradient estimators to further enhance the performance and applicability of our technique in solving diverse discrete optimization problems.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] A. Pradhan and H. S. Dhillon, "A Probabilistic Reformulation Technique for Discrete RIS Optimization in Wireless Systems," *arXiv:2303.00182*, 2023.

[14] 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," *arXiv:2210.00570*, 2022.

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

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

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

[18] 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.

[19] 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.