Distributionally Robust Optimization for Nonconvex QCQPs with Stochastic Constraints

Eli Brock* EECS, UC Berkeley
Haixiang Zhang* Math, UC Berkeley
Julie Mulvaney Kemp IEOR, UC Berkeley
Javad Lavaei IEOR, UC Berkeley
Somayeh Sojoudi EECS, UC Berkeley

***Abstract*— The quadratically constrained quadratic program (QCQP) with stochastic constraints appears in a wide range of real-world problems, including but not limited to the control of power systems. The randomness in the constraints prohibits the application of classic stochastic optimization algorithms. In this work, we utilize the techniques from the distributionally robust optimization (DRO) and propose two novel optimization formulations to solve the QCQP problems under strong duality. The proposed formulations do not contain stochastic constraints. The solutions to the optimization formulations attain the optimal objective value among all solutions that satisfy the stochastic constraints with high probability under the data-generating distribution, even when only a few samples from the distribution are available. We design corresponding algorithms to solve the optimization problems under both formulations. Numerical experiments are conducted to verify the theory and illustrate the empirical performance of the proposed algorithm. This work provides the first results on the application of DRO techniques to non-convex optimization problems with stochastic constraints and the approach can be extended to a broad class of optimization problems.**

I. Introduction

In a wide range of real-world applications, one needs to solve the quadratically constrained quadratic programs (QCQP) with stochastic constraints:

$$\min_{x\in\mathbb{R}^n} x^T M_0 x \quad \text{s.t. } x^T M_i x \geq \xi_i, \quad \forall i \in [m], \tag{1}$$

where $M_i \in \mathbb{R}^{n\times n}$ are symmetric matrices, $\xi \in \Xi \subset \mathbb{R}^m$ is a random vector and $[m] := \{1,\dots,m\}$ for positive integer m. The distribution of ξ is usually unknown and only a few samples $\xi^1,\dots,\xi^S$, which are generated from the distribution, are available.

In general, the QCQPs are nonconvex and are $\mathcal{NP}$-hard to solve in the worst case [1]. However, real-world optimization problems are usually highly structured and it is possible to reduce the computational complexity by utilizing their structures. Consider, for example, the optimal power flow (OPF) problem, which is similar to (1) in that power flow constraints are nonconvex quadratic functions of bus voltages. Moreover, such constraints are often stochastic in nature, as they reflect uncertain variables such as the power demand and the renewable generation. Many practical power circuits exhibit zero duality gap as a consequence of their network structures and admit an exact relaxation [2]. More generally, problems with specific graph structures are distinguished from abstract optimization problems and several relaxation approaches are proposed to transform the nonconvex problem to an equivalent convex problem; see [3]–[5]. One common relaxation approach used in OPF and other problems is to transform problem (1) to a semi-definite program (SDP):

$$\min_{X\in\mathbb{R}^{n\times n}} \langle M_0, X\rangle \text{ s.t. } X \succeq 0,\ \langle M_i, X\rangle \geq \xi_i,\ \forall i \in [m]. \tag{2}$$

Under suitable conditions on $M_0,\dots,M_m$, problems (1) and (2) are equivalent [3] (i.e., the relaxation is tight). In this work, we make the following assumptions, including that such suitable conditions are present.

Assumption 1. *Problem (1) is feasible and has a finite optimal value for all $\xi \in \Xi$. In addition, the SDP relaxation of problem (1) is tight. For all $\xi \in \Xi$, Slater's condition [6] holds for problem (2).*

Although the SDP problem (2) is a convex optimization problem, its constraints are determined by a random vector ξ and prohibit the application of deterministic convex optimization algorithms or stochastic optimization algorithms, which are applicable to optimization problems that only contain randomness in the objective function. Existing algorithms for optimization problems with stochastic constraints find solutions that satisfy the constraints in expectation [7], [8]. However, the meaning of the expectation of constraints is undefined in many applications and a robust solution that satisfies each constraint with high probability is desired.

This work presents two new formulations of problem (2), which support high-probability bounds on the optimal solution and are developed using tools from distributionally robust optimization (DRO) [9]. Based on the empirical distribution of ξ, the new formulations admit an optimal solution X^* under the condition that the constraints are satisfied with high probability under the data-generating distribution of ξ. To be more specific, the formulations admit a solution X^* with the minimal objective value among all X for which it holds that

$$\mathbb{P}^0\left[\sum\nolimits_{i\in[m]} \omega_i \left(\langle M_i, X^*\rangle - \xi_i\right) \geq 0\right] \geq \beta, \tag{3}$$

for all weight vector[1] $\omega \in \mathbb{R}^m$, where $\mathbb{P}^0(\cdot)$ is the probability under the data-generating distribution of ξ and $\beta \in [0,1)$

*Equal Contribution.

[1]A vector $\omega \in \mathbb{R}^m$ is called a weight vector if $\omega_i \geq 0$ for all $i \in [m]$ and $\sum_{i\in[m]} \omega_i = 1$.

is any pre-specified probability. Note that the bound (3) is stronger than that in [10]. Most existing works on DRO focused on convex optimization problems; see [9], [11] for a review. However, in practice, a variety of applications include non-convex optimization problems. Our work is the first to provide a bound, as well as the first to apply DRO, to a nonconvex problem under strong duality.

The paper is organized as follows. In section II, we first introduce the DRO formulation of problem (2) and develop an optimization problem that is based on the expectation of ξ. Next, in Section III, we modify the expectation-based formulation and derive another optimization problem that is based on the quantiles of ξ, which is able to provide stronger high-probability bounds. We provide the theoretical guarantees of the solutions to both formulations in Sections II and III, respectively. Finally, in Section IV, we implement the proposed algorithms to verify the theory and illustrate the empirical performances. We conclude the paper in Section V.

II. Expectation-based Formulation

In this section, we first introduce the expectation-based DRO formulation of problem (2) and then establish the theoretical guarantees satisfied by the optimal solution of the optimization problem. To apply DRO techniques, we consider the dual problem of problem (2) with a fixed instance of ξ:

$$\max_{\nu\in\mathbb{R}^m} \xi^T\nu \quad \text{s.t. } M_0 - \sum\nolimits_{i\in[m]}\nu_i M_i \succeq 0, \quad \nu \geq 0, \quad (4)$$

where the vector inequality $\nu \geq 0$ means that $\nu_i \geq 0$ for all $i \in [m]$. Since problem (2) is a SDP problem with a finite optimal value, strong duality holds and solving problems (2) and (4) is equivalent. Compared with the primal problem (2), the randomness in the dual problem (4) only appears in the objective function $\xi^S\nu$. This property allows the application of various techniques in stochastic optimization. One common approach to solving problem (4) is minimizing the expectation of $\xi^S\nu$ under the empirical distribution. More specifically, we define the cost function and its expectation as

$$\gamma(\nu,\xi) := \xi^T\nu, \quad c(\nu,\mathbb{P}) := \mathbb{E}_{\xi\sim\mathbb{P}}\left[\gamma(\nu,\xi)\right], \quad \forall \nu,\xi\in\mathbb{R}^m, \quad \mathbb{P}\in\mathcal{P}.$$

Then, the empirical mean minimization of problem (4) can be written as

$$\max_{\nu\in\mathbb{R}^m} c(\nu,\hat{\mathbb{P}}) \quad \text{s.t. } M_0 - \sum\nolimits_{i\in[m]}\nu_i M_i \succeq 0, \quad \nu \geq 0,$$

where

$$\hat{\mathbb{P}}_S := \frac{1}{S}\sum\nolimits_{i\in[S]}\delta_{\xi^i}$$

is the empirical distribution of ξ and $\xi^1,\ldots,\xi^S$ are S independently and identically distributed samples from the distribution $\mathbb{P}^0$, where δ_ξ is the Dirac measure at ξ. To deal with the discrepancy between the true distribution and the empirical distribution of ξ, the DRO formulation in [12] serves as a useful tool. We first define the distributionally robust predictor.

Definition 1 (Distributionally Robust Predictor)**.** *Suppose that $\mathcal{P}$ is the set of Borel distributions and $r \geq 0$ is a constant. For all $\mathbb{P}'\in\mathcal{P}$ and input $\nu\in\mathcal{V}$, the distributionally robust predictor is defined as*

$$\hat{c}_r(\nu,\mathbb{P}') := \sup_{\mathbb{P}\in\mathcal{P}}\left\{c(\nu,\mathbb{P}) \mid I(\mathbb{P}',\mathbb{P}) \leq r\right\}, \quad (5)$$

where $I(\cdot,\cdot)$ is the relative entropy as defined in [13]. In the case when $\mathbb{P}' = \hat{\mathbb{P}}_S$, we denote the distributionally robust predictor as $\hat{c}_{r,\hat{\mathbb{P}}_S}$ for the notational simplicity.

We note that the above definition is the opposite to that in Definition 6 of [12], which considers the infimum of $c(\nu,\cdot)$ under the entropy constraint. Intuitively, this is because our ultimate goal is to derive bounds for the primal problem (2) through the dual problem. Now, we define the corresponding distributionally robust prescriptor.

Definition 2 (Distributionally Robust Prescriptor)**.** *For all $\mathbb{P}'\in\mathcal{P}$, the distributionally robust prescriptor $\hat{\nu}_r(\mathbb{P}')$ a quasi-continuous function that is a maximizer of the problem*

$$\max_{\nu\in\mathbb{R}^m} \hat{c}_r(\nu,\mathbb{P}') \quad \text{s.t. } M_0 - \sum\nolimits_{i\in[m]}\nu_i M_i \succeq 0, \quad \nu\geq 0.$$

In the case when $\mathbb{P}' = \hat{\mathbb{P}}_S$, we denote the distributionally robust prescriptor as $\hat{\nu}_{r,\hat{\mathbb{P}}_S}$ for the notational simplicity.

To ensure the existence and the regularity of $\hat{\nu}_r(\cdot)$, we make the following assumption on the feasible set.

Assumption 2. *The feasible set $\mathcal{V} := \{\nu\in\mathbb{R}^m | \nu\geq 0, M_0 - \sum_{i\in[m]}\nu_i M_i \succeq 0\}$ is compact.*

By Proposition 4 of [12], Assumption 2 guarantees that the function $\hat{\nu}_r(\cdot)$ exists and is quasi-continuous in $\mathbb{P}'$. Moreover, the pair $(\hat{c}_{r,\hat{\mathbb{P}}_S}, \hat{\nu}_{r,\hat{\mathbb{P}}_S})$ is the strong solution to the meta-optimization problem (6) in [12]. Namely, $\hat{c}_{r,\hat{\mathbb{P}}_S}(\hat{\nu}_{r,\hat{\mathbb{P}}_S})$ is the minimal value among all predictors that are larger than the population expectation with high probability in S.

Lemma 1. *Suppose that $\mathbb{P}^\infty$ is the sample path distribution under $\mathbb{P}^0$. For all $r>0$, the following two claims hold:*

1) *The predictor $\hat{c}_{r,\hat{\mathbb{P}}_S}$ satisfies*

$$\hat{c}_{r,\hat{\mathbb{P}}_S}(\bar{\nu}) \leq \hat{c}'_{r,\hat{\mathbb{P}}_S}(\bar{\nu}), \quad \forall\bar{\nu}\in\mathcal{V}$$

for all predictors $\hat{c}'_{r,\hat{\mathbb{P}}_S}$ such that for all $\bar{\nu}\in\mathcal{V}$,

$$\limsup_{S\to+\infty}\frac{1}{S}\log\mathbb{P}^\infty\left[c(\bar{\nu},\mathbb{P}^0) > \hat{c}'_{r,\hat{\mathbb{P}}_S}(\bar{\nu})\right] \leq -r.$$

2) *The pair $(\hat{c}_{r,\hat{\mathbb{P}}_S}, \hat{\nu}_{r,\hat{\mathbb{P}}_S})$ satisfies*

$$\hat{c}_{r,\hat{\mathbb{P}}_S}(\hat{\nu}_{r,\hat{\mathbb{P}}_S}) \leq \hat{c}'_{r,\hat{\mathbb{P}}_S}(\hat{\nu}'_{r,\hat{\mathbb{P}}_S})$$

for all pair $(\hat{c}'_{r,\hat{\mathbb{P}}_S}, \hat{\nu}'_{r,\hat{\mathbb{P}}_S})$ satisfying

$$\limsup_{S\to+\infty}\frac{1}{S}\log\mathbb{P}^\infty\left[c(\hat{\nu}'_{r,\hat{\mathbb{P}}_S},\mathbb{P}^0) > \hat{c}'_{r,\hat{\mathbb{P}}_S}(\hat{\nu}'_{r,\hat{\mathbb{P}}_S})\right] \leq -r.$$

The proof is the same as those of Theorems 4 and 7 in [12] and we omit it. Utilizing the distributionally robust

Algorithm 1 Algorithm for the expectation-based formulation.

1: **Input:** Matrices $M_0, \ldots, M_m$, samples $\xi^1, \ldots, \xi^S$, constant $r > 0$.
2: **Output:** Primal solution $\hat{X}_{r,\hat{\mathbb{P}}_S}$.
3: Find the distributionally robust prescriptor $\hat{\nu}_{r,\hat{\mathbb{P}}_S}$.
4: Find the distribution $\tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}$ such that

$$c(\hat{\nu}_{r,\hat{\mathbb{P}}_S}, \tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}) = \hat{c}_r(\hat{\nu}_{r,\hat{\mathbb{P}}_S}, \hat{\mathbb{P}}_S).$$

5: Solve the primal problem (2) with ξ fixed as $\mathbb{E}_{\xi\sim\tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}}(\xi)$ and return the solution $\hat{X}_{r,\hat{\mathbb{P}}_S}$.

prescriptor $\hat{\nu}_{r,\hat{\mathbb{P}}_S}$, we are able to generate a primal solution $\hat{X}_{r,\hat{\mathbb{P}}_S}$ with guarantees on the constraint satisfaction; see Algorithm 1. More concretely, we use the DRO approach to find the "worst-case" distribution $\hat{\mathbb{P}}_S$ and solve the primal problem (2) with ξ fixed to be the expectation under the worst-case distribution.

Now, we provide a practical algorithm to find the primal solution $\hat{X}_{r,\hat{\mathbb{P}}_S}$. First, we show that for all $\nu \in \mathcal{V}$, it is possible to evaluate the distributionally robust predictor $\hat{c}_{r,\hat{\mathbb{P}}_S}(\nu)$ and find the corresponding distribution that attains the supremum in $\hat{c}_{r,\hat{\mathbb{P}}_S}(\nu)$, namely, the distribution $\tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S} \in \mathcal{P}$ that satisfies $\hat{c}_{r,\hat{\mathbb{P}}_S}(\nu) = c(\nu, \tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S})$. Note that the supremum in (5) can be attained since the feasible set is compact and the objective function is linear. Using Lemma 2 of [12], the set of feasible distributions can be restricted to the set of distributions that are absolutely continuous with respect to $\hat{\mathbb{P}}_S$ except on the set

$$\Xi^*(\nu) := \{\xi \mid \gamma(\nu, \xi) = \bar{\gamma}(\nu)\},$$

where $\bar{\gamma}(\nu) := \max_{\xi\in\Xi} \gamma(\nu, \xi)$. For any feasible distribution $\mathbb{P}$, we denote

$$p_i := \tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}(\xi^i), \quad \forall i \in [S].$$

Then, we have $\tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}[\Xi^*(\nu)] = 1 - \sum_{i\in[S]} p_i$ and problem (5) is equivalent to

$$\begin{array}{ll} \max\limits_{p_1,\ldots,p_S} & \sum_{i\in[S]} p_i \cdot \gamma(\nu, \xi^i) + \left(1 - \sum_{i\in[S]} p_i\right) \bar{\gamma}(\nu) \\ \text{s.t.} & \sum_{i\in[S]} \log(p_i) \geq -S(r + \log S), \\ & \sum_{i\in[S]} p_i \leq 1, \quad p_i \geq 0, \quad \forall i \in [S]. \end{array} \tag{6}$$

Noticing that problem (6) is convex in p_i and satisfies Slater's condition when $r > 0$, the optimal solution can be derived via the Karush-Kuhn-Tucker (KKT) conditions. After a direct analysis of the KKT conditions, we get the algorithm for computing the distribution $\tilde{\mathbb{P}}$ as follows. Denote

$$\delta_i := \bar{\gamma}(\nu) - \gamma(\nu, \xi^i) + \eta, \quad \forall i \in [S],$$

where η is a positive number such that

$$\left(\prod\nolimits_{i\in[S]} \delta_i\right)^{1/S} \cdot \left(\frac{1}{S} \sum\nolimits_{i\in[S]} \frac{1}{\delta_i}\right) = e^r$$

or $\eta = 0$ if such positive number does not exist. We can find η by the bi-section method. The worst-case distribution is given by

$$\tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}(\xi^i) = e^{-r} \left(\prod\nolimits_{j\in[S]} \delta_j\right)^{1/S} / (S\delta_i), \quad \forall i \in [S].$$

Then, the distributionally robust predictor is given by $\hat{c}_{r,\hat{\mathbb{P}}_S}(\nu) = c(\nu, \tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S})$. Using the evaluation of $\hat{c}_{r,\hat{\mathbb{P}}_S}(\cdot)$, we can find an approximation to the distributionally robust prescriptor $\hat{\nu}_{r,\hat{\mathbb{P}}_S}$ by zeroth-order optimization methods [14], [15]. Furthermore, the distribution $\tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}$ for $\hat{\nu}_{r,\hat{\mathbb{P}}_S}$ can be computed using the above method. The final step is to solve problem (2) for $\hat{X}_{r,\hat{\mathbb{P}}_S}$, which is a SDP problem and can be solved by convex optimization methods.

The next theorem proves the theoretical properties satisfied by the solution $\hat{X}_{r,\hat{\mathbb{P}}_S}$.

Theorem 2. *For any $r > 0$, the solution $\hat{X}_{r,\hat{\mathbb{P}}_S}$ satisfies*

$$\langle M_0, \hat{X}_{r,\hat{\mathbb{P}}_S} \rangle \leq \langle M_0, \hat{X}'_{r,\hat{\mathbb{P}}_S} \rangle$$

for all $\hat{X}'_{r,\hat{\mathbb{P}}_S} \succeq 0$ such that

$$\begin{array}{l} \limsup\limits_{S\to+\infty} \frac{1}{S} \log \mathbb{P}^\infty \left(\begin{bmatrix} \langle M_1, \hat{X}'_{r,\hat{\mathbb{P}}_S} \rangle - \xi_1^0 \\ \vdots \\ \langle M_m, \hat{X}'_{r,\hat{\mathbb{P}}_S} \rangle - \xi_m^0 \end{bmatrix}^T \nu < 0 \right) \leq -r, \\ \forall \nu \in \mathcal{V}, \quad \text{s.t.} \left\langle M_0 - \sum\nolimits_{i\in[S]} \nu_i M_i, \hat{X}'_{r,\hat{\mathbb{P}}_S} \right\rangle = 0, \end{array} \tag{7}$$

where ξ^0 is the expectation of ξ under the distribution $\mathbb{P}^0$.

Proof. The proof is finished in two steps.

Step I. We first prove that the solution $\hat{X}_{r,\hat{\mathbb{P}}_S}$ satisfies condition (7). By definition, the pair $(\hat{\nu}_{r,\hat{\mathbb{P}}_S}, \tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S})$ is a solution to

$$\max_{\nu\in\mathbb{R}^m} \max_{\mathbb{P}\in\mathcal{P}} \; c(\nu, \mathbb{P}), \quad \text{s.t. } \nu \in \mathcal{V}, \quad I(\hat{\mathbb{P}}_S, \mathbb{P}) \leq r. \tag{8}$$

Switching the two maximization operations, it is equivalent to

$$\max_{\mathbb{P}\in\mathcal{P}} \max_{\nu\in\mathbb{R}^m} \; c(\nu, \mathbb{P}), \quad \text{s.t. } \nu \in \mathcal{V}, \quad I(\hat{\mathbb{P}}_S, \mathbb{P}) \leq r.$$

Using the strong duality of problem (2), the above problem has the same optimal value as

$$\begin{array}{ll} \max\limits_{\mathbb{P}\in\mathcal{P}} \min\limits_{X\in\mathbb{R}^{n\times n}} & \langle M_0, X \rangle, \\ & \text{s.t. } X \succeq 0, \quad \langle M_i, X \rangle \geq \mathbb{E}_{\xi\sim\mathbb{P}}(\xi_i), \; \forall i \in [S], \\ & \quad I(\hat{\mathbb{P}}_S, \mathbb{P}) \leq r. \end{array} \tag{9}$$

Since $\mathbb{P} = \tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}$ attains the optimal value of problem (8), the problem (9) has the same optimal value as

$$\begin{array}{ll} \min\limits_{X\in\mathbb{R}^{n\times n}} & \langle M_0, X \rangle, \\ & \text{s.t. } X \succeq 0, \quad \langle M_i, X \rangle \geq \mathbb{E}_{\xi\sim\tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}}(\xi_i), \; \forall i \in [m]. \end{array} \tag{10}$$

The solution to the above problem is exactly $\hat{X}_{r,\hat{\mathbb{P}}_S}$. Therefore, $\langle M_0, \hat{X}_{r,\hat{\mathbb{P}}_S}\rangle$ is the optimal value of the problem (10) and is equal to that of problem (8), which is $\hat{c}_{r,\hat{\mathbb{P}}_S}(\hat{\nu}_{r,\hat{\mathbb{P}}_S})$. Using the constraint of problem (10) and the condition on ν, we get

$$\begin{bmatrix} \langle M_1, \hat{X}_{r,\hat{\mathbb{P}}_S}\rangle \\ \vdots \\ \langle M_m, \hat{X}_{r,\hat{\mathbb{P}}_S}\rangle \end{bmatrix}^T \nu - \hat{c}_{r,\hat{\mathbb{P}}_S}(\nu) = \langle M_0, \hat{X}_{r,\hat{\mathbb{P}}_S}\rangle - \hat{c}_{r,\hat{\mathbb{P}}_S}(\nu)$$

$$= \hat{c}_{r,\hat{\mathbb{P}}_S}\left(\hat{\nu}_{r,\hat{\mathbb{P}}_S}\right) - \hat{c}_{r,\hat{\mathbb{P}}_S}(\nu) \geq 0,$$

$$\forall \nu \in \mathcal{V}, \quad \text{s.t. } \left\langle M_0 - \sum\nolimits_{i\in[S]} \nu_i M_i, \hat{X}_{r,\hat{\mathbb{P}}_S}\right\rangle = 0.$$

From the first claim of Lemma 1, it holds that

$$\limsup_{S\to+\infty} \frac{1}{S}\log \mathbb{P}^\infty\left[c(\nu,\mathbb{P}^0) > \hat{c}_{r,\hat{\mathbb{P}}_S}(\nu)\right] \leq -r, \quad \forall \nu \in \mathcal{V}.$$

Combining $c(\nu,\mathbb{P}^0) = (\xi^0)^T\nu$ with the last two inequalities, we get the condition (7) for $\hat{X}_{r,\hat{\mathbb{P}}_S}$.

Step II. Now, we prove that $\hat{X}_{r,\hat{\mathbb{P}}_S}$ attains the minimal objective value among all predictors that satisfy condition (7). Suppose that $\hat{X}'_{r,\hat{\mathbb{P}}_S}$ also satisfies the condition (7). Without loss of generality, we can assume that $\hat{X}'_{r,\hat{\mathbb{P}}_S}$ is a minimizer to

$$\min_{X\in\mathbb{R}^{n\times n}} \ \langle M_0, X\rangle, \tag{11}$$
$$\text{s.t. } X \succeq 0, \quad \langle M_i, X\rangle \geq \langle M_i, \hat{X}'_{r,\hat{\mathbb{P}}_S}\rangle, \quad \forall i \in [S].$$

Otherwise, we can replace $\hat{X}'_{r,S}$ with a solution to the problem (11) and the value of $\langle M_0, \hat{X}'_{r,S}\rangle$ will decrease without violating the inequality (7). Define the function

$$\hat{c}'_{r,\hat{\mathbb{P}}_S}(\nu) := \begin{bmatrix} \langle M_1, \hat{X}'_{r,\hat{\mathbb{P}}_S}\rangle \\ \vdots \\ \langle M_m, \hat{X}'_{r,\hat{\mathbb{P}}_S}\rangle \end{bmatrix}^T \nu, \quad \forall \nu \in \mathcal{V}.$$

The dual problem of problem (11) can be written as

$$\max_{\nu\in\mathbb{R}^m} \ \hat{c}'_{r,\hat{\mathbb{P}}_S}(\nu), \quad \text{s.t. } \nu \in \mathcal{V}. \tag{12}$$

Let $\hat{\nu}'_{r,S}$ be a solution to the problem (12). Then, the complementary slackness condition of problem (11) implies that the pair $(\hat{c}'_{r,S}, \hat{\nu}'_{r,S})$ is a data-driven predictor-prescriptor pair that satisfies the condition (7). The second claim of Lemma 1 ensures that

$$\hat{c}'_{r,\hat{\mathbb{P}}_S}(\hat{\nu}'_{r,S}) \geq \hat{c}_{r,\hat{\mathbb{P}}_S}(\hat{\nu}_{r,S}).$$

In summary, we get

$$\langle M_0, \hat{X}'_{r,S}\rangle = \hat{c}'_{r,\hat{\mathbb{P}}_S}(\hat{\nu}'_{r,\hat{\mathbb{P}}_S}) \geq \hat{c}_{r,\hat{\mathbb{P}}_S}(\hat{\nu}_{r,\hat{\mathbb{P}}_S}) = \langle M_0, \hat{X}_{r,\hat{\mathbb{P}}_S}\rangle,$$

which implies that $\hat{X}_{r,\hat{\mathbb{P}}_S}$ has the minimal value $\langle M_0, \hat{X}_{r,\hat{\mathbb{P}}_S}\rangle$ among all $X \succeq 0$ that satisfies the condition (7). □

From the theorem, we can see that $\hat{X}_{r,\hat{\mathbb{P}}_S}$ attains the minimal objective value under the condition (7). Intuitively, this condition claims that a weighted combination of the (expected) constraints of problem (2) is satisfied with high probability for some weights $\nu \in \mathcal{V}$. However, the meaning of the comparison to the expected value ξ^0 in the constraints is not clear, especially when the sample size S is small. In the next section, we introduce the quantile-based formulation, which avoids this limitation.

III. Quantile-based Formulation

In this section, we provide another DRO formulation to problem (2), which is able to provide stronger theoretical guarantees than the expectation-based formulation and avoid the limitations. Since we want to find solutions that satisfy probability constraint with the form of (3), we can directly enforce the probability bound using the quantiles of $\gamma(\cdot,\xi)$ as the objective function. For all $\alpha \in [0,1]$, we define the α-quantile of $\gamma(\nu,\xi)$ as

$$q_\alpha(\nu,\mathbb{P}) := \sup\left\{\gamma \mid \mathbb{P}\left[\gamma(\nu,\xi) \leq \gamma\right] \leq \alpha\right\}, \quad \forall \nu \in \mathcal{V},\ \mathbb{P} \in \mathcal{P}.$$

Then, we define the distributionally robust predictor and the distributionally robust prescriptor in the same way as the expectation-based formulation.

Definition 3 (Distributionally Robust Predictor-Prescriptor)**.** *Suppose that $\alpha \in [0,1]$ and $r \geq 0$ are constants. For all $\mathbb{P}' \in \mathcal{P}$ and input $\nu \in \mathcal{V}$, the distributionally robust predictor is defined as*

$$\hat{q}_{\alpha,r}(\nu,\mathbb{P}') := \sup_{\mathbb{P}\in\mathcal{P}}\left\{q_\alpha(\nu,\mathbb{P}) \mid I(\mathbb{P}',\mathbb{P}) \leq r\right\}.$$

The corresponding distributionally robust prescriptor $\hat{\nu}_{\alpha,r}(\mathbb{P}')$ is a quasi-continuous function that is a maximizer of

$$\max_{\nu\in\mathbb{R}^m} \hat{q}_{\alpha,r}(\nu,\mathbb{P}') \quad \text{s.t. } \nu \in \mathcal{V}. \tag{13}$$

In the case when $\mathbb{P}' = \hat{\mathbb{P}}_S$, we denote the distributionally robust predictor-prescriptor pair as $(\hat{q}_{\alpha,r,\hat{\mathbb{P}}_S}(\cdot), \hat{\nu}_{\alpha,r,\hat{\mathbb{P}}_S})$ for the notational simplicity.

The existence of $\hat{\nu}_{\alpha,r}(\cdot)$ is also guaranteed by Assumption 2 and we omit the proof. Note that we still use the supremum in the definition of $\hat{q}_{\alpha,r}$. We first show that the distributionally robust predictor $\hat{q}_{\alpha,r,\hat{\mathbb{P}}_S}(\cdot)$ is also a quantile of $\gamma(\cdot,\xi)$ under the empirical distribution $\hat{\mathbb{P}}_S$.

Lemma 3. *For all $\alpha \in [0,1]$ and $r, S > 0$, there exists an integer $k(\alpha,r,S) \in [S+1]$ such that*

$$\hat{q}_{\alpha,r,\hat{\mathbb{P}}_S}(\nu) = \gamma_{(k(\alpha,r,S))}(\nu;\hat{\mathbb{P}}_S), \quad \forall \nu \in \mathcal{V},$$

where $\gamma_{(k)}(\nu;\hat{\mathbb{P}}_S)$ is the k-th smallest value of $\{\gamma(\nu,\xi^i), i \in [S]\} \cup \{\bar{\gamma}(\nu)\}$.

Proof. We first show that for the predictor $\hat{q}_{\alpha,r,\hat{\mathbb{P}}_S}(\cdot)$, the set of feasible distributions can also be restricted to the set of distributions that are absolutely continuous with respect to $\hat{\mathbb{P}}$ except on the set

$$\Xi^*(\nu) := \{\xi \mid \gamma(\nu,\xi) = \bar{\gamma}(\nu)\}.$$

The proof is the same as that of Lemma 2 of [12] except the bound on the expectation, i.e., the second last inequality in the proof. To deal with this issue, we only need to prove that for all $\nu \in \mathcal{V}$, $p \in [0,1]$, $\xi^* \in \Xi^*$, $\mathbb{P}_c \ll \hat{\mathbb{P}}_S$ and $\mathbb{P}_\perp \in \mathcal{P}$ such that $\mathbb{P}_\perp \perp \mathbb{P}_c$, it holds that

$$Q_{\mathbb{P}',\alpha}[\gamma(\nu,\xi)] \geq Q_{\mathbb{P}'',\alpha}[\gamma(\nu,\xi)], \tag{14}$$

where

$$\mathbb{P}' := p \cdot \mathbb{P}_c + (1-p) \cdot \delta_{\xi^*}, \quad \mathbb{P}'' := p \cdot \mathbb{P}_c + (1-p) \cdot \mathbb{P}_\perp.$$

Let $F'(\gamma)$ and $F''(\gamma)$ be the cumulative distribution function of $\gamma(\nu,\xi)$ under the distribution $\mathbb{P}'$ and $\mathbb{P}''$, respectively. By the definition of the quantile, to prove inequality (14), it is sufficient to show that

$$F'(\gamma) \geq F''(\gamma), \quad \forall \gamma,$$

which is equivalent to

$$\mathbb{E}_{\xi\sim\mathbb{P}'}\left[\mathbf{1}(\gamma(\nu,\xi) \leq \gamma)\right] \geq \mathbb{E}_{\xi\sim\mathbb{P}''}\left[\mathbf{1}(\gamma(\nu,\xi) \leq \gamma)\right], \quad \forall \gamma,$$

where $\mathbf{1}(\gamma(\nu,\xi) \leq \gamma)$ is an indicator function. This can be proved in the same way as the proof in [12]. As a result, there exists an integer $k \in [S+1]$ such that $\hat{q}_{\alpha,r,\hat{\mathbb{P}}_S}(\nu) = \gamma_{(k)}(\nu;\hat{\mathbb{P}}_S)$.

Next, we prove that the integer k does not depend on ν and $\hat{\mathbb{P}}_S$. Let $\tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}$ be the worst-case distribution that attains $\hat{q}_{r,\hat{\mathbb{P}}_S}(\nu)$. Assume without loss of generality that

$$\gamma(\nu,\xi^1) \leq \cdots \leq \gamma(\nu,\xi^S).$$

Denote

$$p_i := \tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}(\xi^i), \quad \forall i \in [S], \quad p_{S+1} := \tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}(\Xi^*).$$

Then, the integer k is the solution to

$$\begin{aligned}\max_{k\in[S],p\in\mathbb{R}^{S+1}} \quad & k, \\ \text{s.t.} \quad & \sum\nolimits_{i\in[k]} p_i \leq \alpha, \quad -\frac{1}{S}\sum\nolimits_{i\in[S]} \log(Sp_i) \leq r, \\ & \sum\nolimits_{i\in[S+1]} p_i = 1, \quad p_i \geq 0, \quad \forall i \in [S+1],\end{aligned}$$

which is independent of ν and $\hat{\mathbb{P}}_S$. Intuitively, k is the smallest integer such that the probability $\tilde{\mathbb{P}}_{r,\hat{\mathbb{P}}_S}$ on the smallest k elements is at least α and the relative entropy constraint is not violated. □

When there is no confusion about α, r and S, we denote $k := k(\alpha,r,S)$ for simplicity and re-write problem (13) as

$$\max_{\nu\in\mathbb{R}^m} \ \gamma_{(k)}(\nu;\hat{\mathbb{P}}_S), \quad \text{s.t. } \nu \in \mathcal{V}. \tag{15}$$

In the case when $k = S+1$, the evaluation of $\gamma_{(S+1)}(\nu;\hat{\mathbb{P}}_S)$ requires the knowledge of Ξ, which may be unknown in practice. Hence, we focus on the case when $k \in [S]$ in the remainder of the paper. The distributionally robust prescriptor $\hat{\nu}_{k,\hat{\mathbb{P}}_S}$ is a solution to problem (15). To get a solution for problem (2), we define the Lagrangian function

$$L(\nu,X;\hat{\mathbb{P}}_S) := \gamma_{(k)}(\nu;\hat{\mathbb{P}}_S) + \left\langle X, M_0 - \sum\nolimits_{i\in[S]} \nu_i M_i \right\rangle.$$

Then, we consider the mini-max problem

$$\min_{X\in\mathbb{R}^{n\times n}} \max_{\nu\in\mathbb{R}^m} \ L(\nu,X;\hat{\mathbb{P}}_S), \quad \text{s.t. } \nu \geq 0, \quad X \succeq 0.$$

Then, the dual function to problem (15) is defined as

$$d(X) := \max_{\nu\in\mathbb{R}^m} \ L(\nu,X;\hat{\mathbb{P}}_S), \quad \text{s.t. } \nu \geq 0.$$

We make the following assumption on the dual problem.

Assumption 3. *The dual problem $\min_{X\succeq 0} d(X)$ is feasible, i.e., there exists $X \succeq$ such that $d(X) < +\infty$.*

Under Assumption 3, we show that the dual problem has a finite optimal value.

Lemma 4. *The dual problem $\min_{X\succeq 0} d(X)$ has a finite optimal value.*

Proof. We consider the following relaxation of the dual problem:

$$\begin{aligned}\min_{X\in\mathbb{R}^{n\times n}} \quad & \langle M_0, X\rangle, \\ \text{s.t.} \quad & \langle M_i, X\rangle \geq \xi_i^{(k)}, \quad \forall i \in [m], \quad X \succeq 0,\end{aligned}$$

where $\xi_i^{(k)}$ is the k-th smallest value in $\{\xi_i^1,\ldots,\xi_i^S\}$. The dual problem has a finite optimal value if the relaxed problem has a finite optimal value. Since the relaxed problem is a SDP problem, it has the dual problem

$$\max_{\nu\in\mathbb{R}^m} \ \gamma(\nu,\xi^{(k)}), \quad \text{s.t. } \nu \in \mathcal{V}.$$

Since the dual problem is a special case of problem (2), it is feasible with a bounded optimal value by Assumption 2. Hence, the duality theory implies that the relaxed problem is also feasible and has a bounded optimal value. This finishes the proof. □

As a result, we can choose the primal solution $\hat{X}_{k,\hat{\mathbb{P}}_S}$ to be an optimum of the dual problem:

$$\hat{X}_{k,\hat{\mathbb{P}}_S} \in \operatorname*{arg\,min}_{X\in\mathbb{R}^{n\times n}} \ d(X), \quad \text{s.t. } X \succeq 0. \tag{16}$$

The following lemma characterizes the dual function.

Lemma 5. *We have $d(X) = \langle M_0, X\rangle < +\infty$ if and only if*

$$\gamma_{(k)}(\nu;\hat{\mathbb{P}}_S) \leq \sum\nolimits_{i\in[m]} \nu_i \langle M_i, X\rangle, \ \forall \nu \in \mathbb{R}^m, \ \text{s.t. } \nu \geq 0. \tag{17}$$

Proof. We first prove the necessity part. Suppose that there exists $\nu \in \mathbb{R}^m$ such that

$$\nu \geq 0, \quad \gamma_{(k)}(\nu) > \sum\nolimits_{i\in[m]} \nu_i \langle M_i, X\rangle.$$

Then, we choose a constant $C > 0$ and consider

$$\begin{aligned}& L_{(k)}(C\nu, X;\hat{\mathbb{P}}_S) \\ =& C \cdot \gamma_{(k)}(\nu;\hat{\mathbb{P}}_S) + \left\langle X, M_0 - C \cdot \sum\nolimits_{i\in[m]} \nu_i M_i \right\rangle \\ =& C\left(\gamma_{(k)}(\nu;\hat{\mathbb{P}}_S) - \sum\nolimits_{i\in[m]} \nu_i \langle M_i, X\rangle\right) + \langle M_0, X\rangle.\end{aligned}$$

Letting $C \to +\infty$, we have

$$d(X) \geq L_{(k)}(C\nu, X; \hat{\mathbb{P}}_S) \to +\infty.$$

This is a contradiction to the condition that $d(X) < +\infty$.

Then, we prove the sufficiency part. By the condition,

$$\begin{aligned} L_{(k)}(\nu, X; \hat{\mathbb{P}}_S) &= \gamma_{(k)}(\nu; \hat{\mathbb{P}}_S) - \sum_{i\in[m]} \nu_i \langle M_i, X\rangle + \langle M_0, X\rangle \\ &\leq \langle M_0, X\rangle. \end{aligned}$$

Therefore, $\nu = 0$ is a maximizer of the Lagrangian function over ν and $d(X) = \langle M_0, X\rangle < +\infty$. □

Intuitively, the condition (17) implies that the constraints of problem (2) are satisfied with probability at least $k/S - \exp[-rS + o(S)]$ under the true data-generation distribution $\mathbb{P}^0$. To be more concrete, we have the following theorem.

Theorem 6. *Suppose that X satisfies the condition (17). For all weight vector $\omega \in \mathbb{R}^m$ and $k \in [S+1]$, it holds that*

$$\begin{aligned} &\mathbb{P}^0\left[\sum\nolimits_{i\in[m]} \omega_i \left(\langle M_i, X\rangle - \xi_i\right) \geq 0\right] \\ &\qquad \geq \alpha - \exp[-rS + o(S)]. \end{aligned} \tag{18}$$

Proof. Choosing $\nu = \omega$ in the condition (17), it follows that for at least k samples in $\{\xi^i, i \in [S]\}$, it holds that

$$\gamma(\omega, \xi^i) \leq \sum\nolimits_{j\in[m]} \omega_j \langle M_j, X\rangle.$$

By the definition $\gamma(\nu, \xi) = \nu^T \xi$, it follows that

$$\sum\nolimits_{j\in[m]} \omega_j \left[\langle M_j, X\rangle - \xi^i_j\right] \geq 0. \tag{19}$$

The condition (19) says that a weighted average of the constraints is satisfied with weight ω_j. Therefore, under the empirical distribution $\hat{\mathbb{P}}_S$, we have

$$\hat{\mathbb{P}}_S\left[\sum\nolimits_{j\in[m]} \omega_j \left[\langle M_j, X\rangle - \xi_j\right] \geq 0\right] \geq \frac{k}{S}.$$

Now, Theorem 10 of [12] implies that

$$\limsup_{S\to+\infty} \frac{1}{S} \log\left\{\mathbb{P}^\infty\left[q_\alpha(\omega, \mathbb{P}^0) > \gamma_{(k)}(\omega; \hat{\mathbb{P}}_S)\right]\right\} \leq -r.$$

Combining the last inequality with, we get

$$\mathbb{P}^0\left[\sum\nolimits_{j\in[m]} \omega_j \left(\langle X, M_j\rangle - \xi_j\right) > 0\right] \geq \alpha - \exp[-rS + o(S)].$$

This finishes the proof. □

In practice, natural choices of ω might include the unit vectors $e_1, \ldots, e_m$. In this case, Theorem 6 guarantees that each of the constraints individually is satisfied with the stated probability. However, ω can also be chosen to encode any constraint "budget" by setting the weights according to the relative value of the satisfaction (or violation) margin among the m constraints. The strength of Theorem 6 is that it holds for any such budget under the unknown true distribution.

By definition, the primal solution $\hat{X}_{k,\hat{\mathbb{P}}_S}$ satisfies the condition (17) and thus, it also satisfies the condition in Theorem 6. In practive, the user may first choose k and then choose a suitable α and r to maximize the right-hand side of (18). Given $k \in [S]$ and $\alpha \in [0, k/S]$, the maximal radius r such that $k(\alpha, r, S) = k$ is given by

$$r = -\frac{k}{S}\log\left(\frac{S\alpha}{k}\right) - \frac{S-k}{S}\log\left(\frac{S(1-\alpha)}{S-k}\right),$$

where we define $0\log(0) = 0$. Therefore, given the sample size $S \gg 1$ and the parameter $k \in [S]$, one wants to solve the maximization problem

$$p^*_{k,S} := \max_{\alpha\in[0,k/S]} \ \alpha - \frac{S^S}{k^k(S-k)^{S-k}} \cdot \alpha^k (1-\alpha)^{S-k}.$$

The solution of the above problem will maximize the right-hand side of (18).

Algorithm 2 Algorithm for the quantile-based formulation.

Input: Matrices $M_0, \ldots, M_n$, empirical distribution $\hat{\mathbb{P}}_S$, number of iterations t_{max}, parameter $k \in [S]$.
Output: Primal solution $\hat{X}_{k,\hat{\mathbb{P}}_S}$.
Initialize $\mathcal{S}_1 \leftarrow \{e_i \mid i \in [m]\}$.
for $t = 1, 2, \ldots, t_{max}$ **do**
  Update X_t to be a maximizer to the SDP problem:

$$\begin{aligned} \max_{X\in\mathbb{R}^{n\times n}} \ & \langle M_0, X\rangle, \\ \text{s.t.} \ & \textstyle\sum_{i\in[m]} \nu_i \langle M_i, X\rangle \geq \gamma_{(k)}(\nu), \quad \forall \nu \in \mathcal{S}_t, \\ & X \succeq 0. \end{aligned}$$

  if condition (17) holds for X_t **then**
    break
  end if
  Find weight vector $\tilde{\nu} \in \mathbb{R}^m$ that violates (17), i.e.,

$$\textstyle\sum_{i\in[m]} \tilde{\nu}_i \langle M_i, X\rangle < \gamma_{(k)}(\tilde{\nu}).$$

  Update $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{\tilde{\nu}\}$.
end for
Return the last iterate of X_t as $\hat{X}_{k,\hat{\mathbb{P}}_S}$.

Now, we provide an algorithm for the dual problem (16). The algorithm is based on the cutting-plane method [16] and is described in Algorithm 2. Here, we denote the i-th unit basis of $\mathbb{R}^m$ as e_i for all $i \in [m]$. Basically, we approximate the condition (17) by a finite number of linear constraints

$$\sum\nolimits_{i\in[m]} \nu_i \langle M_i, X\rangle \geq \gamma_{(k)}(\nu), \quad \forall \nu \in \mathcal{S}_t.$$

These constraints provide a relaxed condition of (17), which requires the inequality to hold for all weight vectors ν. If the solution of the relaxed problem X_t satisfies the condition (17), it must be an optimal solution to the dual problem (16).

Now, we describe an algorithm to check whether the condition (17) is satisfied for a given matrix X. In addition, if condition (17) fails, the algorithm finds a weight vector $\tilde{\nu}$ that violates the condition. The algorithm is based on the

following mixed-integer programming (MIP) problem:

$$\begin{aligned}\min_{z\in\mathbb{R}^S, t\in\mathbb{R}, \nu\in\mathbb{R}^m} \quad & t, \\ \text{s.t. } & t + C\cdot z_i \geq \sum\nolimits_{j\in[m]} \nu_j \left(\langle M_0, X\rangle - \xi_j^i\right), \\ & z_i \in \{0,1\},\ \forall i \in [S], \quad \sum\nolimits_{i\in[S]} z_i = k-1,\end{aligned}$$

where $C \gg 1$ is a large enough constant. The MIP problem is based on the big-M method [17]. If the variable $z_i = 1$, since the constant C is sufficiently large, there is no constraint on t. Otherwise if the variable $z_i = 0$, the constraint requires that

$$t \geq \sum\nolimits_{j\in[m]} \nu_j \left(\langle M_0, X\rangle - \xi_j^i\right).$$

This means that t should be the maximal value of the right-hand side over all indices i such that $z_i = 0$. With a given ν, to minimize the value of t, z_i is equal to one for indices with the $k-1$ smallest values of the right-hand side. Then, the optimal value of t should be the k-th smallest value of the right-hand side over all samples. If we further minimize over the weight vector ν, the condition (17) holds if and only if the optimal value t^* is non-negative. In addition, if $t < 0$, the corresponding vector ν^* provides a weight vector such that condition (17) is violated by X. Although Algorithm 2 requires solving an MIP problem, the algorithm runs efficiently in practice and exhibits good empirical performances in our examples; see more details in Section IV.

IV. Numerical Experiments

In this section, we test the Algorithm 2 for the quantile-based formulation on a synthetic example. For a given dimension n, we choose $m = 2(n-1)$ and generate matrices $M_0, \ldots, M_m$ as follows. Let $\mathcal{G}$ be a connected, undirected, acyclic graph with n nodes. In our experiments, we choose $\mathcal{G}$ to be a tree with n nodes. For each $i \in \{0, \ldots, n-1\}$, we define

$$(M_i)_{j,k} := \begin{cases} 0, & \text{if } (j,k) \notin \mathcal{G} \\ \psi_{i,j,k}, & \text{if } (j,k) \in \mathcal{G}, \end{cases}$$

where $\{\psi_{i,j,k} \mid i \in [n-1], (j,k) \in \mathcal{G}\}$ are independent uniform random variables on $[0,1]$. Then, we define

$$M_{i+n-1} := -M_i, \quad \forall i \in [n-1].$$

For the random vector ξ, its first $n-1$ entries are independent uniform random variables on $[-1,0]$. The last $n-1$ entries of ξ are equal to the first $n-1$ entries. This definition of $M_1, \ldots, M_m$ and ξ leads to the constraints

$$\xi_i \leq \langle M_i, X\rangle \leq -\xi_i, \quad \forall i \in [n-1].$$

We choose graph-structured matrices as they mirror the objective and constraint functions of OPF and similar network flow problems. It is shown in [3], [4] that the graph structure and the non-negativeness of the entries of M_i ensure that the problem (2) admits an exact SDP relaxation. As a result, one can generate the decision variable for problem (1) from a solution to problem (2) by the algorithm in [18].

To verify the results of Theorem 6, we generate $S' \gg S$ independent samples of ξ, which are denoted as $\tilde{\xi}^1, \ldots, \tilde{\xi}^{S'}$. For each $i \in [m]$, we count the number of samples that satisfy the constraint

$$\left\langle M_i, \hat{X}_{k,\hat{\mathbb{P}}_S} \right\rangle \geq \xi_i^j.$$

By theory, we expect at least $p^*_{k,S} S'$ samples to satisfy the above condition. We choose the maximal number of iterations $t_{max} = 100$ for Algorithm 2. In all tested examples, the optimal solution is found and the algorithm terminates in less than $t_{max} = 100$ iterations. The problem size and the sample size is $n = 10$ and $S = 20$, respectively. We use $S' = 10^4$ samples to verify the results of Theorem 6. We implement the Algorithm 2 for all quantiles $k \in [S]$ and compare the performances. The algorithms are implemented in Python 3.10 and MATLAB 2023a environment equipped with solvers MOSEK 10.0 [19] and Gurobi 10.0 [20].

As a baseline for comparison, we test Algorithm 2 against the naive approach of requiring that each constraint be satisfied for at least k samples from the empirical distribution. Specifically, the naive approach chooses the solution of (2) with $\xi = \xi^{(k)}$, where the i-th element of $\xi^{(k)}$ is the k-th smallest of $\{\xi_i^1, \ldots, \xi_i^S\}$. As S grows, this naive algorithm becomes more reliable, but for problems with a small number of samples and many constraints, it will not be robust to the true distribution.

The results are summarized in Figure 1. For all $k \in [S]$, the output of Algorithm 2 (i.e., $\hat{X}_{k,\hat{\mathbb{P}}_S}$) satisfies the condition (17). From the figure, we can see the trade-offs between the the optimal objective value and the constraint satisfaction rate, which can be adjusted by choosing the parameter k. As the high-probability bound becomes stricter with a larger k, the objective value becomes larger but the constraints are satisfied by more samples. Hence, both Algorithm 2 and the naive algorithm exhibit the expected behavior with respect to k. From the left plot, we can see that the objective values of Algorithm 2 are larger than those of the naive algorithm. This is because the naive algorithm enforces a relaxed condition of (17). In the right plot, we compute the probability that the solution satisfies a given constraint for the S' extra samples. We compare the pointwise (over k) minimum and the mean satisfaction rate among all m constraints, and we also compare the rates with the theoretical lower bound $p^*_{k,S}$. We can see that both solutions satisfy the constraint with a probability that is larger than the lower bound, except for the naive solution at $k = 14$. In addition, the Algorithm 2 finds more robust solutions than the naive algorithm. The naive algorithm is not theoretically guaranteed to generate distributionally robust solutions. As k approaches T, the performance of the two algorithms become similar, though this behavior is not necessarily expected to hold for all problem instances or choices of weights w. The gap between Algorithm 2 and the naive algorithm is expected to grow when the sample size S is small compared to the number of constraints m, or when the true distribution has a high covariance. In other words, as the empirical distribution approaches

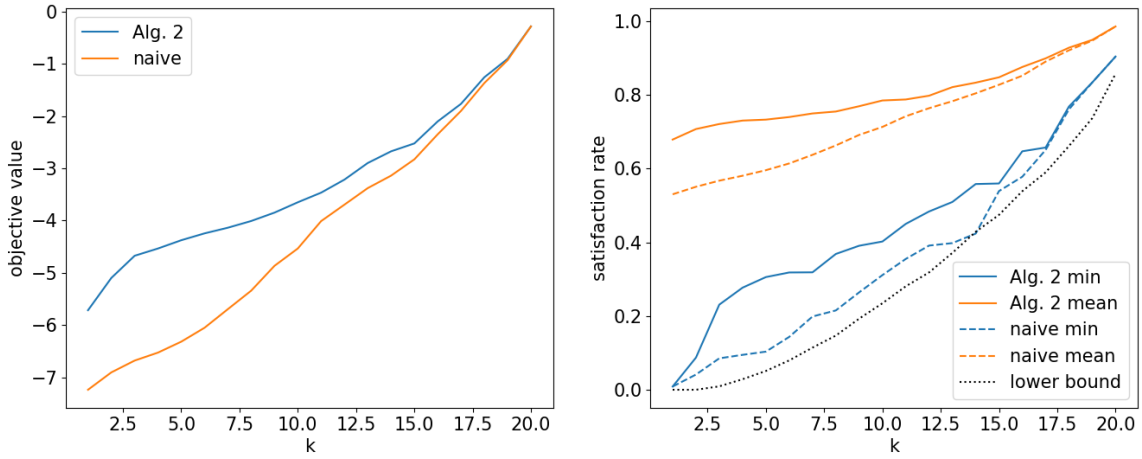

Fig. 1. Results of Algorithm 2 and the naive algorithm. The left plot compares the objective values of the two algorithms. The right plot compares the constraint satisfaction rate of the two algorithms.

the true distribution, the methods become equivalent. As a summary, the naive algorithm can efficiently generate robust solutions in some cases, but Algorithm 2 is theoretically guaranteed and works better especially when the sample size S is small.

V. Conclusion

In this work, we consider the nonconvex QCQPs with stochastic constraints under strong duality. Existing stochastic optimization algorithms only allow randomness in the objective function and thus, they are not applicable. We propose two new DRO formulations, and we prove that the solution to the DRO formulations attains the optimal objective value among all solutions that satisfy the constraints with high probability under the data-generating distribution, even when we only have access to a few samples from the distribution. In addition, we develop corresponding algorithms that solve the proposed DRO formulations and implement the algorithms on a few examples to illustrate the empirical performance. The new formulations are the first results on the application of DRO techniques to a nonconvex optimization problem with stochastic constraints. The approach can be extended to a broad class of nonconvex optimization problems with stochastic constraints and generate robust solutions that satisfy the constraints with high probability.

References

[1] E. De Klerk, "The complexity of optimizing over a simplex, hypercube or sphere: a short survey," *Central European Journal of Operations Research*, vol. 16, pp. 111–125, 2008.

[2] J. Lavaei and S. H. Low, "Zero duality gap in optimal power flow problem," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 92–107, 2011.

[3] S. Sojoudi and J. Lavaei, "Exactness of semidefinite relaxations for nonlinear optimization problems with underlying graph structure," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 1746–1778, 2014.

[4] S. Kim and M. Kojima, "Exact solutions of some nonconvex quadratic optimization problems via SDP and SOCP relaxations," *Computational optimization and applications*, vol. 26, pp. 143–154, 2003.

[5] S. Bose, D. F. Gayme, K. M. Chandy, and S. H. Low, "Quadratically constrained quadratic programs on acyclic graphs with application to power flow," *IEEE Transactions on Control of Network Systems*, vol. 2, no. 3, pp. 278–287, 2015.

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

[7] M. Mahdavi, T. Yang, and R. Jin, "Stochastic convex optimization with multiple objectives," *Advances in neural information processing systems*, vol. 26, 2013.

[8] H. Yu, M. Neely, and X. Wei, "Online convex optimization with stochastic constraints," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[9] J. Goh and M. Sim, "Distributionally robust optimization and its tractable approximations," *Operations research*, vol. 58, no. 4-part-1, pp. 902–917, 2010.

[10] D. Bertsimas, V. Gupta, and N. Kallus, "Data-driven robust optimization," *Mathematical Programming*, vol. 167, pp. 235–292, 2018.

[11] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," *arXiv preprint arXiv:1908.05659*, 2019.

[12] B. P. Van Parys, P. M. Esfahani, and D. Kuhn, "From data to decisions: Distributionally robust optimization is optimal," *Management Science*, vol. 67, no. 6, pp. 3387–3402, 2021.

[13] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[14] A. R. Conn, K. Scheinberg, and L. N. Vicente, "Global convergence of general derivative-free trust-region algorithms to first-and second-order critical points," *SIAM Journal on Optimization*, vol. 20, no. 1, pp. 387–415, 2009.

[15] ——, *Introduction to derivative-free optimization*. SIAM, 2009.

[16] Y. Nesterov *et al.*, *Lectures on convex optimization*. Springer, 2018, vol. 137.

[17] L. A. Wolsey and G. L. Nemhauser, *Integer and combinatorial optimization*. John Wiley & Sons, 1999, vol. 55.

[18] R. Madani, G. Fazelnia, S. Sojoudi, and J. Lavaei, "Low-rank solutions of matrix inequalities with applications to polynomial optimization and matrix completion problems," in *53rd IEEE Conference on Decision and Control*. IEEE, 2014, pp. 4328–4335.

[19] M. ApS, "Mosek optimization toolbox for matlab," *User's Guide and Reference Manual, Version*, vol. 4, p. 1, 2019.

[20] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2023. [Online]. Available: https://www.gurobi.com