2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

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 a novel optimization formulation to solve the QCQP problems under strong duality. The proposed formulation does not contain stochastic constraints. The solutions to the optimization formulation 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 the new formulation. 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 $[m] := \{1, \ldots, m\}$ for positive integer m. The distribution of ξ is usually unknown and only a few samples $\xi^1, \ldots, \xi^S$, 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 $M_0, \ldots, M_m$, problems (1) and (2) are equivalent [3] (i.e., the relaxation is tight). Various optimization solvers are designed to efficiently solve SDP problems and the solution to the SDP problem (2) provides a good approximation to the solution to the original nonconvex problem, even when the relaxation is not tight; see Section 1 of [4] for examples of guarantees on the approximation. In this work, we make the following assumption, 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. The randomness in the constraints prohibits 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 a new formulation of problem (2), which support high-probability bounds on the optimal solution and are developed using tools from distributionally robust optimization (DRO) [9]. The statistical and optimization tools from DRO are able to provide efficient bounds on the worst-case behavior of the stochastic system. In a number of real-world applications, including but not limited to those in power systems, some constraints of problem (1) are safety-related and it is crucial to ensure the satisfaction of those constraints in the worst case. More specifically, based on the empirical distribution of ξ, the new formulation generates a

*Equal Contribution.

solution X^* such that

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

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 $\beta \in [0,1)$ is any pre-specified probability. We conjecture that the solution X^* attains the minimal objective value among all X for which the condition (3) holds. We note that the bound (3) is stronger than those 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 quantiles of ξ, which is able to provide stronger high-probability bounds than expectation-based formulations in [7], [8]. We provide the theoretical guarantees of the solutions to the new formulation in Section II. Finally, in Section III, we implement the proposed algorithms to verify the theory and illustrate the empirical performances. We conclude the paper in Section IV.

II. Quantile-based Formulation

In this section, we provide a new DRO formulation to problem (2), which is able to provide stronger theoretical guarantees than the expectation-based formulations [7], [8] and avoid their limitations. To apply DRO techniques, we consider the dual problem of problem (2) with a fixed instance of ξ:

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

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) are equivalent. Compared with the primal problem (2), the randomness in the dual problem (4) only appears in the objective function $\xi^T\nu$. This property allows the application of various techniques in stochastic optimization.

Suppose there are S independently and identically distributed samples, $\xi^1, \ldots, \xi^S$, from the distribution $\mathbb{P}^0$. We define the empirical distribution of ξ as

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

where δ_ξ is the Dirac measure at ξ. The goal of the DRO formulation is to find solutions that satisfy the high-probability constraint with the form of (3) under the true distribution $\mathbb{P}^0$ using the empirical distribution $\hat{\mathbb{P}}_S$. Since we want to provide high-probability guarantees, we can directly enforce the probability bound using the quantiles of $\gamma(\cdot,\xi)$ as the objective function, where the dual objective function is

$$\gamma(\nu,\xi) := \xi^T\nu, \quad \forall \nu,\xi \in \mathbb{R}^m.$$

For all $\alpha \in [0,1]$, we define the α-quantile of $\gamma(\nu,\xi)$ as

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

where we define the dual feasible set as

$$\mathcal{V} := \left\{\nu\in\mathbb{R}^m \mid \nu\geq 0, M_0 - \sum\nolimits_{i\in[m]}\nu_i M_i \succeq 0\right\}.$$

To deal with the discrepancy between the true distribution and the empirical distribution, the DRO formulation in [12] serves as a useful tool. We first define the distributionally robust predictor.

Definition 1 (Distributionally Robust Predictor)**.** *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*

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

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 $\hat{q}_{\alpha,r,\hat{\mathbb{P}}_S}(\cdot)$ for the notational simplicity.

In our DRO formulation, the ambiguity set is characterized by the relative entropy $I(\cdot,\cdot)$. The relative entropy and the Wasserstein distance are commonly used as a measure of distance between distributions. However, the large deviation theory guarantees that the relative entropy between the true data-generation distribution and the empirical distribution can be bounded by a value that depends on the sample size; see [13] for more details. Hence, we can show that the true distribution is contained in the ambiguity set with high probability and establish worst-case bounds on the ambiguity set. As a result, we consider the relative entropy in the ambiguity set in this work. 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)**.** *Suppose that $\alpha \in [0,1]$ and $r \geq 0$ are constants. For all $\mathbb{P}' \in \mathcal{P}$, the distributionally robust prescriptor $\hat{\nu}_{\alpha,r}(\mathbb{P}')$ is a quasi-continuous function that is a maximizer of*

$$\max_{\nu\in\mathbb{R}^m} \hat{q}_{\alpha,r}(\nu,\mathbb{P}') \quad \text{s.t. } \nu\in\mathcal{V}. \qquad (5)$$

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

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

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 $\mathcal{V}$ is compact.*

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

Now, we 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 1. *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

$$Q_{\mathbb{P}',\alpha}[\gamma(\nu,\xi)] \geq Q_{\mathbb{P}'',\alpha}[\gamma(\nu,\xi)], \tag{6}$$

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 (6), 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

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

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 (5) as

$$\max_{\nu\in\mathbb{R}^m} \ \gamma_{(k)}(\nu;\hat{\mathbb{P}}_S), \quad \text{s.t. } \nu \in \mathcal{V}. \tag{7}$$

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 (7). 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 (7) 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 $X \succeq 0$ such that $d(X) < +\infty$.*

The following lemma characterizes the dual function.

Lemma 2. *We have $d(X) = \langle M_0, X\rangle < +\infty$ if and only if*

$$\gamma_{(k)}(\nu;\hat{\mathbb{P}}_S) \leq \sum\nolimits_{i\in[m]} \nu_i \langle M_i, X\rangle, \ \forall \nu \in \mathbb{R}^m, \ \text{s.t. } \nu \geq 0. \tag{8}$$

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 $d(X) = \langle M_0, X\rangle < +\infty$. □

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

Lemma 3. *The dual problem* $\min_{X\succeq 0} d(X)$ *has a finite optimal value.*

Proof. Under Assumption 3, it suffices to prove that the following problem has a finite optimal value:

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times n}} \ & \langle M_0, X\rangle, \\ \text{s.t. } & d(X) < +\infty, \quad X \succeq 0. \end{aligned}$$

Denote the i-th unit basis of $\mathbb{R}^m$ as e_i for all $i \in [m]$. If we choose $\nu = e_i$ for $i \in [m]$ in condition (8), it follows that

$$\langle M_i, X\rangle \geq \xi_i^{(k)}, \quad \forall i \in [m].$$

By relaxing the condition $d(X) < +\infty$ with the above condition, we get the following relaxation of the dual problem:

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

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:

$$\hat{X}_{k,\hat{\mathbb{P}}_S} \in \operatorname*{arg\,min}_{X\in\mathbb{R}^{n\times n}} \ d(X), \quad \text{s.t. } X \succeq 0. \tag{9}$$

Intuitively, the condition (8) in Lemma 2 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 4. *Suppose that X satisfies the condition (8). For every weight vector $\omega \in \mathbb{R}^m$ and $k \in [S+1]$, it holds that*

$$\begin{aligned} &\mathbb{P}^0\left[\sum\nolimits_{i\in[m]} \omega_i \left(\langle M_i, X\rangle - \xi_i\right) \geq 0\right] \\ &\qquad \geq \alpha - \exp[-rS + o(S)]. \end{aligned} \tag{10}$$

Proof. Choosing $\nu = \omega$ in the condition (8), 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

$$\sum\nolimits_{j\in[m]} \omega_j \left[\langle M_j, X\rangle - \xi_j^i\right] \geq 0. \tag{11}$$

The condition (11) 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 the definition of α-quantile, 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 4 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 4 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 (8) and thus, it also satisfies the condition in Theorem 4. In practice, the user may first choose k and then choose a suitable α and r to maximize the right-hand side of (10). 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 (10).

Now, we provide an algorithm for the dual problem (9). The algorithm is based on the cutting-plane method [14] and

Algorithm 1 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:

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

    if condition (8) holds for X_t **then**
        break
    end if
    Find weight vector $\tilde{\nu} \in \mathbb{R}^m$ that violates (8), i.e.,

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

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

is described in Algorithm 1. 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 (8) 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 (8), which requires the inequality to hold for all weight vectors ν. If the solution of the relaxed problem X_t satisfies the condition (8), it must be an optimal solution to the dual problem (9).

Now, we describe an algorithm to check whether the condition (8) is satisfied for a given matrix X. In addition, if condition (8) 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:

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

where $C \gg 1$ is a large enough constant. Although we also solve an optimization problem with probabilistic constraints, i.e., problem (2), via the MIP approach, we cannot directly use the results in [15]. First, their formulation only allows a joint probabilistic constraint, whereas we require the condition (8) to hold for all $\nu \geq 0$. Additionally, they assume that the random vector $\xi \in \mathbb{R}^m$ has a finite support but we only need ξ to have a compact support.

The MIP problem is based on the big-M method [16]. 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

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

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 (8) 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 (8) is violated by X. Although Algorithm 1 requires solving an MIP problem, the algorithm runs efficiently in practice and exhibits good empirical performances in our examples; see more details in Section III. Since we need to deal with the SDP constraint $X \succeq 0$, Algorithm 1 is different with classical cutting-plane methods, e.g., [17]. Therefore, the convergence of Algorithm 1 cannot be directly derived from those of existing cutting-plane methods. We leave the theoretical analysis of Algorithm 1 to future works.

III. Numerical Experiments

In this section, we test Algorithm 1 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. To be more specific, we want to solve the original QCQP problem (1) in the complex case, namely,

$$\min_{x\in\mathbb{C}^n} x^H M_0 x, \quad \text{s.t. } x^H M_i x \geq \xi_i, \quad \forall i \in [m], \tag{12}$$

where v^H is the conjugate transpose of v for all $v \in \mathbb{C}^n$. Since the matrices $M_0, \ldots, M_m$ are real symmetric matrices and the graph $\mathcal{G}$ is acyclic, Theorem 6 of [3] guarantees that

the semi-definite relaxation of problem (12) is tight. Hence, problem (12) has the same optimal objective value as

$$\min_{X\in\mathbb{C}^{n\times n}} \langle M_0, X\rangle, \quad \text{s.t. } X \succeq 0,\ \langle M_i, X\rangle \geq \xi_i,\ \forall i \in [m], \tag{13}$$

where we define $\langle Y, Z\rangle := \mathrm{Re}[\mathrm{Tr}(Y^T Z)]$ for all matrices $Y, Z \in \mathbb{C}^{n\times n}$ and Re is the real part of a complex number. We note that we implicitly enforce the constraint that X is a Hermitian matrix in the condition $X \succeq 0$. As a result, problem (12) reduces to finding a rank-1 solution

$$X^* = x^*(x^*)^H$$

for problem (13), where $x^* \in \mathbb{C}^n$. Suppose that $x^* = y^* + i \cdot z^*$ for vectors $y^*, z^* \in \mathbb{R}^n$ and define

$$U^* := [y^* \quad z^*] \in \mathbb{R}^{n\times 2}.$$

Since matrices $M_0, \ldots, M_m$ are all real matrices, we have

$$(x^*)^H M_i x^* = \left\langle M_i, U^*(U^*)^T \right\rangle, \quad \forall i \in \{0\} \cup [m].$$

Therefore, finding a rank-1 solution for problem (13) is equivalent to finding a rank-2 solution for problem (2). We can apply our cutting-plane algorithm (Algorithm 1) to find a solution that satisfies the condition (8). Finally, one can generate the solution for problem (12) by the algorithms in [18].

To verify the results of Theorem 4, 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 1. In all tested examples, the optimal solution is found and the algorithm terminates within $t_{max} = 100$ iterations. The problem size and the sample size is $n = 20$ and $S = 20$, respectively. We use $S' = 10^4$ samples to verify the results of Theorem 4. We implement the Algorithm 1 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 1 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 is not guaranteed to be robust to the true distribution.

The results are summarized in Figure 1. For all $k \in [S]$, the output of Algorithm 1 (i.e., $\hat{X}_{k,\hat{\mathbb{P}}_S}$) satisfies the condition (8). 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 1 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 1 are larger than those of the naive algorithm. This is because the naive algorithm enforces a relaxed condition of (8). 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 the satisfaction rate of Algorithm 1 remains well above the theoretical bound, but the naive method may drop below as it does at $k = 16$. To further verify that condition (8) is satisfied with high probability by the solution of Algorithm 1, we generate $1,000$ random weight vectors by the uniform distribution on the set of weight vectors. For all random weight vectors, the solutions of both the naive algorithm and Algorithm 1 satisfy condition (8). This observation indicates that both solutions satisfy the condition with high probability over the uniform distribution of weight vectors. In addition, we conjecture that if condition (8) is satisfied by unit vectors, the condition will hold with high probability over random weight vectors. We leave the proof of this conjecture to future works.

In addition, Algorithm 1 finds more robust solutions than the naive algorithm, an advantage most prominent in the minimum satisfaction rate. The naive algorithm is not theoretically guaranteed to generate distributionally robust solutions. As k approaches T, the performances of the two algorithms become similar, though this behavior is not necessarily expected to hold for all problem instances or choices of weights ω. The gap between Algorithm 1 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 variance. In other words, as the empirical distribution approaches the true distribution, the methods become equivalent. As a summary, the naive algorithm can efficiently generate robust solutions in some cases, but Algorithm 1 is theoretically guaranteed and works better especially when the sample size S is small.

IV. 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 a new DRO formulation, and we prove that the solution to the DRO formulation 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 formulation and implement the algorithms on a few examples to illustrate the empirical performance. The new formulation is the first result on the application

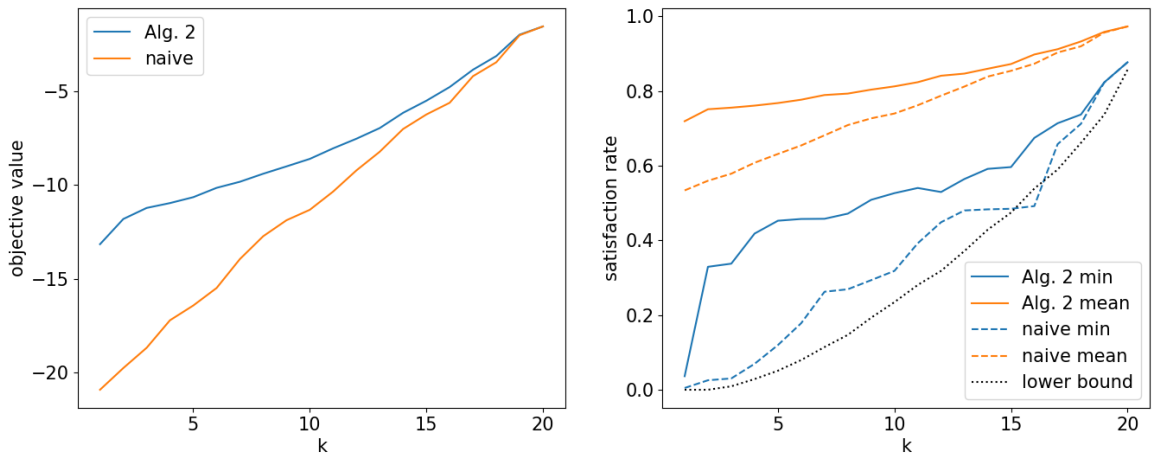

Fig. 1. Results of Algorithm 1 and the naive algorithm. The left plot compares the objective values of the two algorithms. The right plot compares the constraint satisfaction rate of the two algorithms.

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.

ACKNOWLEDGMENT

The authors were partially supported by grants from ARO, ONR, AFOSR and NSF. Eli Brock was partially supported by NSF GRFP. Haixiang Zhang was partially supported by the Two Sigma Ph.D. Fellowship.

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

[15] J. Luedtke, S. Ahmed, and G. L. Nemhauser, "An integer programming approach for linear programs with probabilistic constraints," *Mathematical programming*, vol. 122, no. 2, pp. 247–272, 2010.

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

[17] P. M. Vaidya, "A new algorithm for minimizing convex functions over convex sets," *Mathematical programming*, vol. 73, no. 3, pp. 291–341, 1996.

[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