OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X | EISSN 1526-5463 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000
© 0000 INFORMS

Distributionally Robust Optimization under Moment Uncertainty with Application to Data-Driven Problems

Erick Delage
Department of Management Sciences, HEC Montréal, Montreal, Quebec, Canada, H3T 2A7
erick.delage@hec.ca, http://neumann.hec.ca/pages/erick.delage

Yinyu Ye
Department of Management Science and Engineering, Stanford University, Stanford, California, USA
yinyu-ye@stanford.edu, http://www.stanford.edu/∼yyye

Stochastic programming can effectively describe many decision making problems in uncertain environments. Unfortunately, such programs are often computationally demanding to solve. In addition, their solution can be misleading when there is ambiguity in the choice of a distribution for the random parameters. In this paper, we propose a model that describes uncertainty in both the distribution form (discrete, Gaussian, exponential, etc.) and moments (mean and covariance matrix). We demonstrate that for a wide range of cost functions the associated distributionally robust (or min-max) stochastic program can be solved efficiently. Furthermore, by deriving a new confidence region for the mean and the covariance matrix of a random vector, we provide probabilistic arguments for using our model in problems that rely heavily on historical data. These arguments are confirmed in a practical example of portfolio selection, where our framework leads to better performing policies on the "true" distribution underlying the daily returns of financial assets.

Subject classifications: Programming: stochastic, Statistics: estimation, Finance: portfolio.
Area of review: Optimization.
History: Received February 2008; revisions received September 2008, December 2008, February 2009, accepted April 2009.

1. Introduction

Stochastic programming can effectively describe many decision making problems in uncertain environments. For instance, given that one is interested in solving a convex optimization problem of the type

$$\operatorname*{minimize}_{\boldsymbol{x}\in\mathcal{X}} \quad h(\boldsymbol{x},\boldsymbol{\xi}) \;\;,$$

where $\mathcal{X}$ is a convex set of feasible solutions and $h(\boldsymbol{x},\boldsymbol{\xi})$ is a convex cost function in $\boldsymbol{x}$ that depends on some vector of parameters $\boldsymbol{\xi}$, it is often the case that at the time of optimization, the parameters have not yet been fully resolved. For example, an investment manager cannot know the exact return for all available securities, or in a different context, a manufacturing manager might not know the exact amount of future demand.

If one chooses to represent his uncertainty about $\boldsymbol{\xi}$ through a distribution F, one can instead resort to minimizing the expected cost. This leads to solving a stochastic program:

$$\text{(SP)} \qquad\qquad \operatorname*{minimize}_{\boldsymbol{x}\in\mathcal{X}} \quad \mathbb{E}\left[h(\boldsymbol{x},\boldsymbol{\xi})\right] \;\;,$$

where the expectation is taken with respect to the random parameters $\boldsymbol{\xi}\in\mathbb{R}^m$. Thus, based on a well formulated stochastic model, our investment banker can now choose a portfolio of stocks which maximizes long-term expected return, or similarly our manufacturing company can take an early decision which leads to the highest expected profits. Unfortunately, although the SP is a convex optimization problem, in order to solve it one must often resort to Monte Carlo approximations (see Shapiro and Homem-de-Mello (2000)), which can be computationally challenging. A more challenging difficulty that arises in practice is the need to commit to a distribution F given only limited information about the stochastic parameters.

In an effort to address these issues, a robust formulation for stochastic programming was proposed by Scarf (1958). In this model, after defining a set $\mathcal{D}$ of probability distributions that is assumed to include the true distribution F, the objective function is reformulated with respect to the worst case expected cost over the choice of a distribution in this set. Hence, this leads to solving the Distributionally Robust Stochastic Program:

$$\text{(DRSP)} \qquad \underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} \left(\max_{F\in\mathcal{D}} \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})] \right) ,$$

where $\mathbb{E}_F[\cdot]$ is the expectation taken with respect to the random vector $\boldsymbol{\xi}$ given that it follows the probability distribution F.

Since its introduction, this model has gained a lot of interest in the context of computing upper bounds on the moment of a random vector (*i.e.*, the moment problem as reviewed in Landau (1987)), of computing upper bounds on the optimal value of a stochastic program (*e.g.*, in Birge and Wets (1987) and in Kall (1988)), and of providing robust decisions when distribution information is limited (e.g. in Dupacová (1987) and in Shapiro and Kleywegt (2002)).

Depending on the context, authors have considered a wide range of forms for the distributional set $\mathcal{D}$. Interestingly, if one uses the distributional set that contains distributions that put all of their weight at a single point in the parameter's support set $\mathcal{S}$, then the DRSP reduces to the so-called robust optimization problem (*e.g.*, in Ben-Tal and Nemirovski (1998) and in Bertsimas et al. (2008)). Otherwise, in Lagoa and Barmish (2002) and in Shapiro (2006), the authors consider a set containing unimodal distributions that satisfy some given support constraints. Under some conditions on $h(\boldsymbol{x},\boldsymbol{\xi})$, they characterize the worst distribution as being the uniform distribution. The most popular type of distributional set $\mathcal{D}$ imposes linear constraints on moments of the distribution as is discussed in Scarf (1958), in Dupacová (1987), in Prékopa (1995) and in Bertsimas and Popescu (2005). While many more forms of distributional set can be found in the literature (see Dupacová (2001) and reference therein), our work falls in the category of approaches that consider constraints on the first and second moments of the distribution.

In order to make the DRSP model tractable, approaches that consider moment constraints have typically assumed that these moments are known exactly. For instance, in his original model, Scarf considered a newsvendor problem with a one dimensional decision variable x and a single parameter ξ, which represented respectively how much inventory the vendor should hold and a random amount, with known mean and variance, of demand for the newspapers. The cost function had the form $h(x,\xi) = \max\{c(x-\xi)\ ,\ r(\xi - x)\}$. To solve this model, Scarf exploited the fact that the worst case distribution of demand could be chosen to be one with all of its weight on two points. This idea was reused in Yue et al. (2006), in Zhu et al. (2006), and in Popescu (2007) where, although the objective function takes more interesting forms, the authors assume known first and second moments of the stochastic demand. Moreover, in each case, the proposed solution method relies on characterizing the worst case distribution as a point distribution.

The computational difficulties related to dealing with $\boldsymbol{\xi}$ of larger dimension and with richer objective functions have limited the practical application of the DRSP model. More specifically, it is rare that the worst case moment expression can be simplified analytically, like in the linear chance constraint problem considered in Calafiore and El Ghaoui (2006). Instead, it is more common that the model is intractable and that only global optimization methods can be used to get an optimal solution (*e.g.*, in Ermoliev et al. (1985) and in Gaivoronski (1991)). Furthermore, the current approaches can lead to a false sense of security since they often falsely assume exact knowledge of mean and covariance statistics for the stochastic parameters. In many data-driven problems, one must estimate these moments based on limited historical data assumed to be drawn from F. As the experiment presented in Section 4 will demonstrate, disregarding the uncertainty (or noise) in these estimates can lead to taking poor decisions.

The main contribution of this paper is two-fold. First, we present a new set $\mathcal{D}$ of distributions that takes into account the knowledge of the distribution's support and of a *confidence region* for its mean and its second moment matrix. In Section 2, we show that under this distributional set the DRSP can be solved in polynomial time for a large range of objective functions. In fact, the structure of our distribution set allows us to solve instances of the DRSP that are known to be intractable under a moment matching approach (see Example 1 of Section 2.3 for more details). As a second contribution, in Section 3, after deriving a new form of confidence region for a covariance matrix, we show how our proposed distributional set can be well justified when addressing data-driven problems (*i.e.*, problems where the knowledge of $\boldsymbol{\xi}$ is solely derived from historical data). Finally, our model is applied to a portfolio selection problem in Section 4. In the context of this application, our experiments demonstrate that, besides computational advantages, we will provide empirical evidence that our DRSP approach addresses more effectively the true uncertainty that is present in the daily return of stocks.

2. Robust Stochastic Programming with Moment Uncertainty

As we mentioned earlier, it is often the case in practice that one has limited information about the distribution F driving the uncertain parameters which are involved in the decision making process. In such situations, it might instead be safer to rely on estimates of the mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$ of the random vector: *e.g.*, using empirical estimates. However we believe that in such problems, it is also rarely the case that one is entirely confident in these estimates. For this reason, we propose representing this uncertainty using two constraints parameterized by $\gamma_1 \geq 0$ and $\gamma_2 \geq 1$:

$$(\mathbb{E}[\boldsymbol{\xi}] - \boldsymbol{\mu}_0)^{\mathsf{T}} \boldsymbol{\Sigma}_0^{-1} (\mathbb{E}[\boldsymbol{\xi}] - \boldsymbol{\mu}_0) \leq \gamma_1 \tag{1a}$$

$$\mathbb{E}[(\boldsymbol{\xi} - \boldsymbol{\mu}_0)(\boldsymbol{\xi} - \boldsymbol{\mu}_0)^{\mathsf{T}}] \preceq \gamma_2 \boldsymbol{\Sigma}_0 \ . \tag{1b}$$

While Constraint (1a) assumes that the mean of $\boldsymbol{\xi}$ lies in an ellipsoid of size γ_1 centered at the estimate $\boldsymbol{\mu}_0$, Constraint (1b) forces $\mathbb{E}[(\boldsymbol{\xi} - \boldsymbol{\mu}_0)(\boldsymbol{\xi} - \boldsymbol{\mu}_0)^{\mathsf{T}}]$, which we will refer to as the centered second moment matrix of $\boldsymbol{\xi}$, to lie in a positive semi-definite cone defined with a matrix inequality. In other words, it describes how likely $\boldsymbol{\xi}$ is to be close to $\boldsymbol{\mu}_0$ in terms of the correlations expressed in $\boldsymbol{\Sigma}_0$. Finally, the parameters γ_1 and γ_2 provide natural means of quantifying one's confidence in $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ respectively.

In what follows, we will study the DRSP model under the distributional set

$$\mathcal{D}_1(\mathcal{S}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \gamma_1, \gamma_2) = \left\{ F \in \mathcal{M} \;\middle|\; \begin{array}{l} \mathbb{P}(\boldsymbol{\xi} \in \mathcal{S}) = 1 \\ (\mathbb{E}[\boldsymbol{\xi}] - \boldsymbol{\mu}_0)^{\mathsf{T}} \boldsymbol{\Sigma}_0^{-1} (\mathbb{E}[\boldsymbol{\xi}] - \boldsymbol{\mu}_0) \leq \gamma_1 \\ \mathbb{E}[(\boldsymbol{\xi} - \boldsymbol{\mu}_0)(\boldsymbol{\xi} - \boldsymbol{\mu}_0)^{\mathsf{T}}] \preceq \gamma_2 \boldsymbol{\Sigma}_0 \end{array} \right\} ,$$

where $\mathcal{M}$ is the set of all probability measures on the measurable space $(\mathbb{R}^m, \mathcal{B})$, with $\mathcal{B}$ the Borel σ-algebra on $\mathbb{R}^m$, and $\mathcal{S} \subseteq \mathbb{R}^m$ is any closed convex set known to contain the support of F. The set $\mathcal{D}_1(\mathcal{S}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \gamma_1, \gamma_2)$, which will also be referred to in short-hand notation as $\mathcal{D}_1$, can be seen as a generalization of many previously proposed sets. For example, $\mathcal{D}_1(\mathcal{S}, \boldsymbol{\mu}_0, \boldsymbol{I}, 0, \infty)$ imposes exact mean and support constraints as is studied in Dupacová (1987) and in Bertsimas and Popescu (2005). Similarly, $\mathcal{D}_1(\mathbb{R}^m, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, 0, 1)$ relates closely to the exact mean and covariance matrix constraints considered in Scarf (1958), in Yue et al. (2006), and in Popescu (2007). We will soon show that there is a lot to be gained, both on a theoretical and practical point of view, by formulating the DRSP model using the set $\mathcal{D}_1(\mathcal{S}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \gamma_1, \gamma_2)$ which constrains all three types of statistics: support, mean and second moment matrix.

REMARK 1. While our proposed uncertainty model cannot be used to express an arbitrarily large confidence in the second-order statistics of $\boldsymbol{\xi}$, in Section 3 we will show how in practice there are natural ways of assigning $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, γ_1 and γ_2 based on historical data. Of course, in some situations it might be interesting to add the following constraint:

$$\gamma_3 \boldsymbol{\Sigma}_0 \preceq \mathbb{E}[(\boldsymbol{\xi} - \boldsymbol{\mu}_0)(\boldsymbol{\xi} - \boldsymbol{\mu}_0)^{\mathsf{T}}] \ , \tag{2}$$

where $0 \leq \gamma_3 \leq \gamma_2$. Unfortunately, this leads to important computational difficulties for the DRSP model. Furthermore, in most applications of our model, we expect the worst case distribution to actually achieve maximum variance, thus making Constraint (2) irrelevant. For example, an instance of the portfolio optimization problem presented in Section 4 will have this characteristic.

2.1. Complexity of the Inner Moment Problem

We start by considering the question of solving the inner maximization problem of a DRSP that uses the set $\mathcal{D}_1$.

DEFINITION 1. Given any fixed $\boldsymbol{x} \in \mathcal{X}$, let $\Psi(\boldsymbol{x};\gamma_1,\gamma_2)$ be the optimal value of the moment problem:

$$\underset{F\,\in\,\mathcal{D}_1}{\text{maximize}} \;\; \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})] \;\; , \tag{3}$$

where $\mathbb{E}_F[\cdot]$ is the expectation taken with respect to the random vector $\boldsymbol{\xi}$ given that it follows the probability distribution $F \in \mathcal{D}_1$.

Since F is a probability measure on $(\mathbb{R}^m, \mathcal{B})$, Problem (3) can be described as the conic linear problem:

$$\begin{aligned}
&\underset{F}{\text{maximize}} && \int_{\mathcal{S}} h(\boldsymbol{x},\boldsymbol{\xi})dF(\boldsymbol{\xi}) && \text{(4a)}\\
&\text{subject to} && \int_{\mathcal{S}} dF(\boldsymbol{\xi}) = 1 && \text{(4b)}\\
& && \int_{\mathcal{S}} (\boldsymbol{\xi}-\boldsymbol{\mu}_0)(\boldsymbol{\xi}-\boldsymbol{\mu}_0)^{\mathsf{T}} dF(\boldsymbol{\xi}) \preceq \gamma_2 \boldsymbol{\Sigma}_0 && \text{(4c)}\\
& && \int_{\mathcal{S}} \begin{bmatrix} \boldsymbol{\Sigma}_0 & (\boldsymbol{\xi}-\boldsymbol{\mu}_0) \\ (\boldsymbol{\xi}-\boldsymbol{\mu}_0)^{\mathsf{T}} & \gamma_1 \end{bmatrix} dF(\boldsymbol{\xi}) \succeq 0 && \text{(4d)}\\
& && F \in \mathcal{M} \;\; . && \text{(4e)}
\end{aligned}$$

As it is often done with moment problems of this form, we can circumvent the difficulty of finding the optimal value of this problem by making use of duality theory (see Rockafeller (1970) and Rockafeller (1974) for a detailed theory of duality in infinite dimensional convex problems and both Isii (1963) and Shapiro (2001) for the case of conic linear moment problems).

LEMMA 1. *For a fixed $\boldsymbol{x} \in \mathbb{R}^n$, suppose that $\gamma_1 \geq 0$, $\gamma_2 \geq 1$, $\boldsymbol{\Sigma}_0 \succ 0$, and that $h(\boldsymbol{x},\boldsymbol{\xi})$ is F-integrable for all $F \in \mathcal{D}_1$. Then, $\Psi(\boldsymbol{x};\gamma_1,\gamma_2)$ must be equal to the optimal value of the problem:*

$$\begin{aligned}
&\underset{\boldsymbol{Q},\boldsymbol{q},r,t}{\text{minimize}} && r+t && \text{(5a)}\\
&\text{subject to} && r \geq h(\boldsymbol{x},\boldsymbol{\xi}) - \boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\xi} - \boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{q} \;\; \forall\, \boldsymbol{\xi} \in \mathcal{S} && \text{(5b)}\\
& && t \geq \left(\gamma_2\boldsymbol{\Sigma}_0 + \boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathsf{T}}\right) \bullet \boldsymbol{Q} + \boldsymbol{\mu}_0^{\mathsf{T}}\boldsymbol{q} + \sqrt{\gamma_1}\, \|\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{q}+2\boldsymbol{Q}\boldsymbol{\mu}_0)\| && \text{(5c)}\\
& && \boldsymbol{Q} \succeq 0 \;\; , && \text{(5d)}
\end{aligned}$$

where $(\boldsymbol{A} \bullet \boldsymbol{B})$ refers to the Frobenius inner product between matrices, $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ is a symmetric matrix, the vector $\boldsymbol{q} \in \mathbb{R}^m$, and $r,t \in \mathbb{R}$. In addition, if $\Psi(\boldsymbol{x};\gamma_1,\gamma_2)$ is finite then the set of optimal solutions to Problem (5) *must be non-empty.*

We defer the proof of this Lemma to the appendix since it is the result of applying some well established concepts in duality theory.

To show that there exists a tractable solution method for solving Problem (5), we employ a famous equivalence between convex optimization and separation of a convex set from a point.

LEMMA 2. *(Grötschel et al. (1981)) Consider a convex optimization problem of the form*

$$\underset{\boldsymbol{z}\in\mathcal{Z}}{\text{minimize}} \;\; \boldsymbol{c}^{\mathsf{T}}\boldsymbol{z}$$

with linear objective and convex feasible set $\mathcal{Z}$. Given that the set of optimal solutions is non-empty, the problem can be solved to any precision ϵ in time polynomial in $\log(1/\epsilon)$ and in the size of the problem by using the ellipsoid method if and only if $\mathcal{Z}$ satisfies the following two conditions :

1. *for any $\bar{\boldsymbol{z}}$, it can be verified whether $\bar{\boldsymbol{z}} \in \mathcal{Z}$ or not in time polynomial in the dimension of $\boldsymbol{z}$;*
2. *for any infeasible $\bar{\boldsymbol{z}}$, a hyperplane that separates $\bar{\boldsymbol{z}}$ from the feasible set $\mathcal{Z}$ can be generated in time polynomial in the dimension of $\boldsymbol{z}$.*

A first application of this lemma leads to quantifying the difficulty of solving the feasibility problem associated with Constraint (5b).

ASSUMPTION 1. *The support set $\mathcal{S} \subset \mathbb{R}^m$ is convex and compact (closed and bounded), and it is equipped with an oracle that can for any $\boldsymbol{\xi} \in \mathbb{R}^m$ either confirm that $\boldsymbol{\xi} \in \mathcal{S}$ or provide a hyperplane that separates $\boldsymbol{\xi}$ from $\mathcal{S}$ in time polynomial in m.*

LEMMA 3. *Let function $h(\boldsymbol{x}, \boldsymbol{\xi})$ be concave in $\boldsymbol{\xi}$ and be such that one can provide a super-gradient of $\boldsymbol{\xi}$ in time polynomial in m. Then, under Assumption 1, for any fixed assignment $\boldsymbol{x}$, $\boldsymbol{Q} \succeq 0$, and $\boldsymbol{q}$, one can find an assignment $\boldsymbol{\xi}_*$ that is ϵ-optimal with respect to the problem*

$$\begin{aligned} \underset{t,\boldsymbol{\xi}}{\text{maximize}} \quad & t && \text{(6a)} \\ \text{subject to} \quad & t \le h(\boldsymbol{x}, \boldsymbol{\xi}) - \boldsymbol{\xi}^\mathsf{T} \boldsymbol{Q} \boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T} \boldsymbol{q} && \text{(6b)} \\ & \boldsymbol{\xi} \in \mathcal{S} \ , && \text{(6c)} \end{aligned}$$

in time polynomial in $\log(1/\epsilon)$ and the size of the problem.

Proof: First, the feasible set of the problem is convex since $\boldsymbol{Q} \succeq 0$ so that $h(\boldsymbol{x}, \boldsymbol{\xi}) - \boldsymbol{\xi}^\mathsf{T} \boldsymbol{Q} \boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T} \boldsymbol{q}$ is a concave function in $\boldsymbol{\xi}$. Because $\mathcal{S}$ is compact, the set of optimal solutions for Problem (6) is therefore non-empty. By Assumption 1, conditions (1) and (2) of Lemma 2 are met for Constraint (6c). On the other hand, feasibility of Constraint (6b) can be verified directly after the evaluation of $h(\boldsymbol{x}, \boldsymbol{\xi})$; furthermore, for an infeasible assignment $(\bar{\boldsymbol{\xi}}, \bar{t})$, the following separating hyperplane can be generated in polynomial time:

$$t - (\nabla_{\boldsymbol{\xi}} h(\boldsymbol{x}, \bar{\boldsymbol{\xi}}) - 2\boldsymbol{Q}\bar{\boldsymbol{\xi}} - \boldsymbol{q})^\mathsf{T} \boldsymbol{\xi} \le h(\boldsymbol{x}, \bar{\boldsymbol{\xi}}) - \nabla_{\boldsymbol{\xi}} h(\boldsymbol{x}, \bar{\boldsymbol{\xi}})^\mathsf{T} \bar{\boldsymbol{\xi}} + \bar{\boldsymbol{\xi}}^\mathsf{T} \boldsymbol{Q} \bar{\boldsymbol{\xi}} \ ,$$

where $\nabla_{\boldsymbol{\xi}} h(\boldsymbol{x}, \boldsymbol{\xi})$ is a super-gradient of $h(\boldsymbol{x}, \cdot)$. It follows from Lemma 2 that the ellipsoid method will converge to an ϵ-optimal solution in polynomial time. □

We are now able to derive an important result about the complexity of solving Problem (5) under a general form for $h(\boldsymbol{x}, \boldsymbol{\xi})$.

ASSUMPTION 2. *The function $h(\boldsymbol{x}, \boldsymbol{\xi})$ has the form $h(\boldsymbol{x}, \boldsymbol{\xi}) = \max_{k \in \{1,\ldots,K\}} \ h_k(\boldsymbol{x}, \boldsymbol{\xi})$ such that for each k, $h_k(\boldsymbol{x}, \boldsymbol{\xi})$ is concave in $\boldsymbol{\xi}$. In addition, given a pair $(\boldsymbol{x}, \boldsymbol{\xi})$, it is assumed that one can in polynomial time:*
1. *evaluate the value of $h_k(\boldsymbol{x}, \boldsymbol{\xi})$*
2. *find a super-gradient of $h_k(\boldsymbol{x}, \boldsymbol{\xi})$ in $\boldsymbol{\xi}$.*

Furthermore, for any $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{q} \in \mathbb{R}^m$, and any positive semi-definite $\boldsymbol{Q} \in \mathbb{R}^{m \times m}$, the set $\{y \in \mathbb{R} \mid \exists \boldsymbol{\xi} \in \mathcal{S}, \ y \le h(\boldsymbol{x}, \boldsymbol{\xi}) - \boldsymbol{q}^\mathsf{T} \boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T} \boldsymbol{Q} \boldsymbol{\xi}\}$ is closed.

PROPOSITION 1. *Given that $\mathcal{S}$ satisfies Assumption 1 and that $h(\boldsymbol{x}, \boldsymbol{\xi})$ satisfies Assumption 2 and satisfies the condition of Lemma 1, then Problem* (5) *is a convex optimization problem whose optimal value is finite and equal to $\Psi(\boldsymbol{x}; \gamma_1, \gamma_2)$. Moreover, Problem* (5) *can be solved to any precision ϵ in time polynomial in $\log(1/\epsilon)$ and the size of the problem.*

Proof: First, the constraints of Problem (5) describe a convex set since for any $\boldsymbol{\xi} \in \mathcal{S}$ the function $h(\boldsymbol{x}, \boldsymbol{\xi}) - \boldsymbol{\xi}^\mathsf{T} \boldsymbol{Q} \boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T} \boldsymbol{q}$ is linear in $(\boldsymbol{Q}, \boldsymbol{q})$, and the function $(\gamma_2 \boldsymbol{\Sigma}_0 + \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^\mathsf{T}) \bullet \boldsymbol{Q} + \boldsymbol{\mu}_0^\mathsf{T} \boldsymbol{q} + \sqrt{\gamma_1} \ \|\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{q} + 2\boldsymbol{Q}\boldsymbol{\mu}_0)\|$ is convex in $(\boldsymbol{Q}, \boldsymbol{q})$. The feasible set is also non-empty since the assignment $\boldsymbol{Q} = \boldsymbol{I}$, $\boldsymbol{q} = 0$, $t = \textbf{trace}\,(\gamma_2 \boldsymbol{\Sigma}_0 + \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^\mathsf{T}) + 2\sqrt{\gamma_1} \ \|\boldsymbol{\Sigma}_0^{1/2} \boldsymbol{\mu}_0\|$, $r = \sup_{\boldsymbol{\xi} \in \mathcal{S}} \ \max_{k \in \{1,\ldots,K\}} \ h_k(\boldsymbol{x}, \boldsymbol{\xi}) - \|\boldsymbol{\xi}\|^2$ is necessarily feasible. Note that the assumption that each $h_k(\boldsymbol{x}, \boldsymbol{\xi})$ is concave ensures that such an assignment for r exists. Based on Lemma 1 and the fact that the optimal value of Problem (4) is bounded below by $h(\boldsymbol{x}, \boldsymbol{\mu}_0)$, since the Dirac measure[1] $\delta_{\boldsymbol{\mu}_0}$ is in $\mathcal{D}_1$, we can conclude that $\Psi(\boldsymbol{x};\gamma_1,\gamma_2)$ is finite and that the optimal solution set of Problem (5) is non-empty.

We can now use Lemma 2 to show that Problem (5) can be solved efficiently given that we verify the two conditions for each of its constraints. In the case of Constraint (5d), feasibility can be verified in $O(m^3)$ arithmetic operations. Moreover, a separating hyperplane can be generated, if necessary, based on the eigenvector corresponding to the lowest eigenvalue. The feasibility of Constraint (5c) is also easily verified. Based on an infeasible assignment $(\bar{\boldsymbol{Q}},\bar{\boldsymbol{q}},\bar{r},\bar{t})$, a separating hyperplane can be constructed in polynomial time:

$$\left(\gamma_2\boldsymbol{\Sigma}_0+\boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathsf{T}}+\nabla_{\boldsymbol{Q}}g(\bar{\boldsymbol{Q}},\bar{\boldsymbol{q}})\right)\bullet\boldsymbol{Q}+(\boldsymbol{\mu}_0+\nabla_{\boldsymbol{q}}g(\bar{\boldsymbol{Q}},\bar{\boldsymbol{q}}))^{\mathsf{T}}\boldsymbol{q}-t\le\nabla_{\boldsymbol{q}}g(\bar{\boldsymbol{Q}},\bar{\boldsymbol{q}})^{\mathsf{T}}\bar{\boldsymbol{q}}+\nabla_{\boldsymbol{Q}}g(\bar{\boldsymbol{Q}},\bar{\boldsymbol{q}})\bullet\bar{\boldsymbol{Q}}-g(\bar{\boldsymbol{Q}},\bar{\boldsymbol{q}})\,,$$

where $g(\boldsymbol{Q},\boldsymbol{q})=\sqrt{\gamma_1}\;\|\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{q}+2\boldsymbol{Q}\boldsymbol{\mu}_0)\|$ and where $\nabla_{\boldsymbol{Q}}g(\boldsymbol{Q},\boldsymbol{q})$ and $\nabla_{\boldsymbol{q}}g(\boldsymbol{Q},\boldsymbol{q})$ are the gradients of $g(\boldsymbol{Q},\boldsymbol{q})$ in $\boldsymbol{Q}$ and $\boldsymbol{q}$ respectively. Finally, given the assumed structure of $h(\boldsymbol{x},\boldsymbol{\xi})$, Constraint (5b) can be decomposed into K sub-constraints

$$r\ge h_k(\boldsymbol{x},\boldsymbol{\xi})-\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\xi}-\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{q}\;\;\forall\,\boldsymbol{\xi}\in\mathcal{S}\;\;\forall\,k\in\{1,2,...,K\}$$

Considering the k-th sub-constraint, Lemma 3 states that $\sup_{\boldsymbol{\xi}\in\mathcal{S}}h_k(\boldsymbol{x},\boldsymbol{\xi})-\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\xi}-\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{q}$ can be solved to any precision ϵ in polynomial time. Given that the optimal value is found to be above $r+\epsilon$, one can conclude infeasibility of the constraint and generate an associated separating hyperplane using any optimal solution $\boldsymbol{\xi}_*$ as follows:

$$(\boldsymbol{\xi}_*{\boldsymbol{\xi}_*}^{\mathsf{T}})\bullet\boldsymbol{Q}+{\boldsymbol{\xi}_*}^{\mathsf{T}}\boldsymbol{q}+r\ge h_{k^*}(\boldsymbol{x},\boldsymbol{\xi}_*)\;\;.$$

Since K is finite, the conditions derived from Grötschel et al. (1981) are necessarily met by Problem (5). We therefore conclude that $\Psi(\boldsymbol{x};\gamma_1,\gamma_2)$ can be computed up to any precision ϵ in polynomial time using the ellipsoid method. □

2.2. Complexity of the Distributionally Robust Stochastic Program

Based on our result with the inner moment problem, we can now address the existence of a tractable solution method for the DRSP model under the distributional set $\mathcal{D}_1$:

$$\underset{\boldsymbol{x}}{\text{minimize}}\;\;\left(\max_{F\,\in\,\mathcal{D}_1}\;\mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})]\right)\qquad(7a)$$

$$\text{subject to}\;\;\boldsymbol{x}\in\mathcal{X}\;\;.\qquad(7b)$$

ASSUMPTION 3. *The set $\mathcal{X}\subset\mathbb{R}^n$ is convex and compact (closed and bounded), and it is equipped with an oracle that can for any $\boldsymbol{x}\in\mathbb{R}^n$ either confirm that $\boldsymbol{x}\in\mathcal{X}$ or provide a hyperplane that separates $\boldsymbol{x}$ from $\mathcal{X}$ in time polynomial in n.*

ASSUMPTION 4. *The function $h(\boldsymbol{x},\boldsymbol{\xi})$ is convex in $\boldsymbol{x}$. In addition, it is assumed that one can find in polynomial time a sub-gradient of $h(\boldsymbol{x},\boldsymbol{\xi})$ in $\boldsymbol{x}$.*

Based on these new assumptions, the proposition that follows states that the distributionally robust optimization model is tractable.

PROPOSITION 2. *Given that assumptions 1, 2, 3, and 4 hold, then the DRSP model presented in Problem* (7) *can be solved to any precision ϵ in time polynomial in* $\log(1/\epsilon)$ *and the sizes of $\boldsymbol{x}$ and $\boldsymbol{\xi}$.*

Proof: The proof of this theorem follows similar lines as the proof for Proposition 1. We first reformulate the inner moment problem in its dual form, we then use the fact that min-min operations can be performed jointly and that the constraint involving $h(\boldsymbol{x},\boldsymbol{\xi})$ decomposes. This leads to an equivalent convex optimization form for Problem (7):

$$
\begin{aligned}
\underset{\boldsymbol{x},\boldsymbol{Q},\boldsymbol{q},r,t}{\text{minimize}}\quad & r+t && \text{(8a)}\\
\text{subject to}\quad & r \geq h_k(\boldsymbol{x},\boldsymbol{\xi}) - \boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\xi} - \boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{q}\ \ ,\ \ \forall\, \boldsymbol{\xi}\in\mathcal{S},\ \ k\in\{1,...,K\} && \text{(8b)}\\
& t \geq \left(\gamma_2\boldsymbol{\Sigma}_0 + \boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathsf{T}}\right)\bullet\boldsymbol{Q} + \boldsymbol{\mu}_0^{\mathsf{T}}\boldsymbol{q} + \sqrt{\gamma_1}\,\|\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{q}+2\boldsymbol{Q}\boldsymbol{\mu}_0)\| && \text{(8c)}\\
& \boldsymbol{Q}\succeq 0 && \text{(8d)}\\
& \boldsymbol{x}\in\mathcal{X}\ \ . && \text{(8e)}
\end{aligned}
$$

As in the proof of Proposition 1, we need to show that the ellipsoid method can be successfully applied. Because $\mathcal{X}$ is compact and non-empty, similar arguments to those presented in Proposition 1 ensure that the optimal solution set is once again non-empty. Regarding conditions (1) and (2) on the feasible set, the arguments that were presented in the proof of Proposition 1 still apply for Constraint (8c) and Constraint (8d). However, the argument for Constraint (8b) needs to be revisited since $\boldsymbol{x}$ is now considered an optimization variable. Feasibility of an assignment $(\bar{\boldsymbol{x}},\bar{\boldsymbol{Q}},\bar{\boldsymbol{q}},\bar{r})$ can still be verified in polynomial time because of Lemma 3 and of the fact that K is finite. However, in the case that one of the indexed constraints, say the k^*-th one, is found to be infeasible, one now needs to generate a separating hyperplane using the worst case $\boldsymbol{\xi}_*$ and $\nabla_{\boldsymbol{x}} h_k(\bar{\boldsymbol{x}},\boldsymbol{\xi}_*)$, a sub-gradient of $h_{k^*}(\cdot,\boldsymbol{\xi}_*)$ at $\bar{\boldsymbol{x}}$:

$$(\boldsymbol{\xi}_*\boldsymbol{\xi}_*{}^{\mathsf{T}})\bullet\boldsymbol{Q} + \boldsymbol{\xi}_*{}^{\mathsf{T}}\boldsymbol{q} + r - \nabla_{\boldsymbol{x}} h_{k^*}(\bar{\boldsymbol{x}},\boldsymbol{\xi}_*)^{\mathsf{T}}\boldsymbol{x} \geq h_{k^*}(\bar{\boldsymbol{x}},\boldsymbol{\xi}_*) - \nabla_{\boldsymbol{x}} h_{k^*}(\bar{\boldsymbol{x}},\boldsymbol{\xi}_*)^{\mathsf{T}}\bar{\boldsymbol{x}}\ \ .$$

Since by Assumption 4, a sub-gradient $\nabla_{\boldsymbol{x}} h_k(\bar{\boldsymbol{x}},\boldsymbol{\xi}_*)$ can be obtained in polynomial time and since, by Assumption 3, the conditions are met for Constraint (8e), we can conclude that Lemma 2 can be applied. Problem (8) can therefore be solved to any precision in polynomial time. □

We believe this result should be of high significance for both theoreticians and practitioners as it indicates that, if $\min_{\boldsymbol{x}}\max_{\boldsymbol{\xi}\in\mathcal{S}} h(\boldsymbol{x},\xi)$ is a tractable robust optimization problem (*cf*., Ben-Tal and Nemirovski (1998) and Bertsimas et al. (2008)), then the less-conservative DRSP $\min_{\boldsymbol{x}}\max_{F\in\mathcal{D}_1}\mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})]$ is also tractable. In some cases, the inner moment problem might even be reducible (see Section 4 for an example). Moreover, one also has access to the wide spectrum of methods for robust optimization problems: ranging from methods that use cutting planes more efficiently than the ellipsoid method such as in Goffin and Vial (1993), in Ye (1997), and in Bertsimas and Vempala (2004), to methods that approximate the feasible set with a finite number of sampled constraints such as in de Farias and Van Roy (2001) and in Calafiore and Campi (2005).

REMARK 2. The constraint $\boldsymbol{Q}\succeq 0$ plays an important role in making Problem (7) solvable in polynomial time. This constraint corresponds to the second moment matrix inequality in the construction of our distribution set $\mathcal{D}_1$. If the inequality is replaced by an equality, then $\boldsymbol{Q}$ becomes "free" and Problem (6) is no longer ensured to be a convex optimization problem. This explains why many DRSP problems under the "exact covariance constraint" actually become intractable.

REMARK 3. We also remark that the condition, in Assumption 1, that $\mathcal{S}$ be bounded is only imposed in order to simplify the exposition of our results. In the case that $\mathcal{S}$ is unbounded, Proposition 1 and Proposition 2 will hold as long as feasibility with respect to Constraint (5b) can be verified in polynomial time. In fact, given an infeasible assignment $(\bar{\boldsymbol{x}},\bar{\boldsymbol{Q}},\bar{\boldsymbol{q}},\bar{r})$, one can interrupt the solution process for Problem (6) when the achieved maximum is higher than $\bar{r}$. This is guaranteed to occur in polynomial time since either the problem is unbounded above or the set of optimal t^* for Problem (6) is non-empty due to the technical condition in Assumption 2.

2.3. Examples

Because our framework only imposes weak conditions on $h(\boldsymbol{x},\boldsymbol{\xi})$ through Assumption 2 and Assumption 4, it is possible to revisit many popular forms of stochastic programs and reformulate them taking into account distribution and moment uncertainty.

EXAMPLE 1. **Optimal Inequalities in Probability Theory**
Consider the problem of finding a tight upper bound on $\mathbb{P}(\boldsymbol{\xi}\in\mathcal{C})$ for a random vector $\boldsymbol{\xi}$ with known mean and covariance matrix, and some closed set $\mathcal{C}$. By formulating this problem as a semi-infinite linear program:

$$\underset{F\in\mathcal{D}}{\text{maximize}} \quad \int_{\mathcal{S}} \mathbb{1}\{\boldsymbol{\xi}\in\mathcal{C}\}dF(\boldsymbol{\xi}) \ , \tag{9}$$

many have proposed methods that provide extensions to the popular Chebyshev inequality (see Marshall and Olkin (1960) and Bertsimas and Popescu (2005)). However, these methods fail when dealing with support constraints. More specifically, if $\mathcal{C}$ is a finite union of disjoint convex sets, it is known that for Problem (9) with unconstrained support, $\mathcal{S}=\mathbb{R}^m$, the worst case value can be found in polynomial time. But, if the support is constrained, such as $\mathcal{S}=\mathbb{R}^+$, then the problem is known to potentially be NP-hard. In fact, the hardness of this last problem arises already in finding a distribution that is feasible.

Our framework recommends to relax the restrictions on the covariance of $\boldsymbol{\xi}$ and to instead consider the distributional set $\mathcal{D}_1(\mathcal{S},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0,\gamma_1,\gamma_2)$. Such a distributional set constrains all three types of statistics: support, mean and second moment matrix. If $\mathcal{C}$ is a finite union of disjoint convex sets $\mathcal{C}_k$ (equipped with their respective feasibility oracle), and if for each k, $\mathcal{C}_k\cap\mathcal{S}\neq\emptyset$, then our framework leads to a new Chebyshev inequality that can be evaluated in polynomial time. First, in our framework the problem of finding an $F\in\mathcal{D}_1$ is already resolved using the Dirac measure $\delta_{\boldsymbol{\mu}_0}$. We can also construct an $h(\boldsymbol{x},\boldsymbol{\xi})$ that satisfies Assumption 2 and Assumption 4 by choosing $h_0(\boldsymbol{x},\boldsymbol{\xi})=0$ and

$$h_k(\boldsymbol{x},\boldsymbol{\xi}) = \begin{cases} 1 & \text{, if } \boldsymbol{\xi}\in\mathcal{C}_k \\ -\infty & \text{, otherwise.} \end{cases}$$

Therefore, if the distribution of $\boldsymbol{\xi}$ is known to be a member of $\mathcal{D}_1$, then clearly by the fact that

$$\mathbb{P}(\boldsymbol{\xi}\in\mathcal{C}) \,=\, \mathbb{E}\left[\mathbb{1}\{\boldsymbol{\xi}\in\mathcal{C}\}\right] \,=\, \mathbb{E}\left[\max_k \ h_k(\boldsymbol{x},\boldsymbol{\xi})\right] \,=\, \mathbb{E}\left[h(\boldsymbol{x},\boldsymbol{\xi})\right] \,\le\, \max_{F\in\mathcal{D}_1}\ \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})] \ ,$$

a tight Chebyshev bound can be found in polynomial time. Note that by using the form $\mathcal{D}_1(\mathbb{R}^+,\boldsymbol{\mu},\boldsymbol{\Sigma},0,1)$ one can also provide useful upper bounds to the mentioned NP-hard versions of the problem which assumes the covariance matrix to be known.

EXAMPLE 2. **Distributionally Robust Optimization with Piecewise-Linear Convex Cost**
Assume that one is interested in solving the following DRSP model for a general piece-wise linear convex cost function of $\boldsymbol{x}$

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} \quad \left(\max_{F\in\mathcal{D}_1}\ \mathbb{E}_F[\max_k\ \boldsymbol{\xi}_k^{\mathsf{T}}\boldsymbol{x}]\right) \ ,$$

where each $\boldsymbol{\xi}_k\in\mathbb{R}^n$ is a random vector. This model is quite applicable since convex cost functions can be approximated by piecewise linear functions. By considering $\boldsymbol{\xi}$ to be a random matrix whose k-th column is the random vector $\boldsymbol{\xi}_k$ and taking $h_k(\boldsymbol{x},\boldsymbol{\xi})=\boldsymbol{\xi}_k^{\mathsf{T}}\boldsymbol{x}$, which is linear (hence concave) in $\boldsymbol{\xi}$, the results presented earlier allows one to conclude that the DRSP can be solved efficiently. In fact, due to $h_k(\boldsymbol{x},\boldsymbol{\xi})-\boldsymbol{\xi}^{\mathsf{T}}Q\boldsymbol{\xi}-\boldsymbol{\xi}^{\mathsf{T}}q$ being a concave quadratic function of $\boldsymbol{\xi}$, the DRSP can be solved more efficiently then suggested by Proposition 2. For instance, if $\mathcal{S}$ can be formulated as an ellipsoid then the DRSP reduces to a semi-definite program of finite size. Section 4 will exploit this property to solve a portfolio optimization problem efficiently.

EXAMPLE 3. **Distributionally Robust Conditional Value-at-Risk**
Conditional value-at-risk, also called mean excess loss, was recently introduced in the mathematical finance community as a risk measure for decision making. It is closely related to the more common value-at-risk measure, which for a risk tolerance level of $\vartheta \in (0,1)$ evaluates the lowest amount τ such that with probability $1-\vartheta$, the loss does not exceed τ. Instead, CVaR evaluates the conditional expectation of loss above the value-at-risk. In order to keep the focus of our discussion on the topic of DRSP models, we refer the reader to Rockafellar and Uryasev (2000) for technical details on this subject. CVaR has gained a lot of interest in the community because of its attractive computational properties. For instance, Rockafellar and Uryasev (2000) demonstrated that one can evaluate the ϑ-CVaR$[c(\boldsymbol{x},\boldsymbol{\xi})]$ of a cost function $c(\boldsymbol{x},\boldsymbol{\xi})$, where the random vector $\boldsymbol{\xi}$ is distributed according to F, by solving a convex minimization problem:

$$\vartheta\text{-CVaR}_F[c(\boldsymbol{x},\boldsymbol{\xi})] = \min_{\lambda\in\mathbb{R}} \ \lambda + \frac{1}{\vartheta}\mathbb{E}_F\left[(c(\boldsymbol{x},\boldsymbol{\xi})-\lambda)^+\right] \ \ ,$$

where $(y)^+ = \max\{y\ ,\ 0\}$ and where the notation ϑ-CVaR$_F[\cdot]$ emphasizes the fact that the measure depends on the distribution of $\boldsymbol{\xi}$.

While CVaR is an interesting risk measure, it still requires the decision maker to commit to a distribution F. This is a step that can be difficult to take in practice; thus, justifying the introduction of a distributionally robust version of the criterion such as in Čerbáková (2005) and in Zhu and Fukushima (2005). Using the results presented earlier, we can derive new conclusions for the general form of robust conditional value at risk. Given that the distribution is known to lie in a distributional set $\mathcal{D}_1$, let the Distributionally Robust ϑ-CVaR Problem be expressed as:

$$\text{(DR } \vartheta\text{-CVaR)} \qquad \underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} \ \left(\max_{F\in\mathcal{D}_1} \ \vartheta\text{-CVaR}_F[c(\boldsymbol{x},\boldsymbol{\xi})]\right) \ \ .$$

By the equivalence statement presented above, this problem is equivalent to the form

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} \ \left(\max_{F\in\mathcal{D}_1} \ \left(\min_{\lambda\in\mathbb{R}} \ \lambda + \frac{1}{\vartheta}\mathbb{E}_F\left[(c(\boldsymbol{x},\boldsymbol{\xi})-\lambda)^+\right]\right)\right) \ \ .$$

Given that $c(\boldsymbol{x},\boldsymbol{\xi})$ meets the conditions of Assumption 2 and Assumption 4, one can show that the minimax theorem holds for $\max_{F\in\mathcal{D}_1}\min_{\lambda\in\mathbb{R}}$ since the function $\lambda + \frac{1}{\vartheta}\mathbb{E}_F\left[(c(\boldsymbol{x},\boldsymbol{\xi})-\lambda)^+\right]$ is real valued, convex in λ and concave (actually linear) in F, and since $\mathcal{D}_1$ is weakly compact (see Shapiro (2001)). Thus, interchanging the $\max_F$ and $\min_\lambda$ operators leads to an equivalent formulation of the DR ϑ-CVaR Problem:

$$\underset{\boldsymbol{x}\in\mathcal{X},\lambda\in\mathbb{R}}{\text{minimize}} \ \left(\max_{F\in\mathcal{D}_1} \ \mathbb{E}_F[h(\boldsymbol{x},\lambda,\boldsymbol{\xi})]\right) \ \ ,$$

where $h(\boldsymbol{x},\lambda,\boldsymbol{\xi}) = \lambda + \frac{1}{\vartheta}(c(\boldsymbol{x},\boldsymbol{\xi})-\lambda)^+$. Since

$$\begin{aligned} h(\boldsymbol{x},\lambda,\boldsymbol{\xi}) &= \lambda + \frac{1}{\vartheta}\max\{\ 0\ \ ,\ \ c(\boldsymbol{x},\boldsymbol{\xi})-\lambda\ \} \\ &= \max\left\{\ \lambda\ \ ,\ \ \max_k \ \left(1-\frac{1}{\vartheta}\right)\lambda + \frac{1}{\vartheta}c_k(\boldsymbol{x},\boldsymbol{\xi})\ \right\} \ \ , \end{aligned}$$

it is clear that $h(\boldsymbol{x},\lambda,\boldsymbol{\xi})$ meets Assumption 2 and Assumption 4. Hence, Proposition 2 allows us to conclude that finding an optimal $\boldsymbol{x}$ (and its associated λ) with respect to the worst case conditional value-at-risk over the set of distributions $\mathcal{D}_1$ can be done in polynomial time.

3. Moment Uncertainty in Data-driven Problems

The computational results presented in the previous section rely heavily on the structure of the described distributional set $\mathcal{D}_1$. This set was built to take into account moment uncertainty in the stochastic parameters. We now turn ourselves to showing that such a structure can be naturally justified in the context of data-driven optimization problems. To be more specific, we now focus on problems where the knowledge of the stochastic parameters is restricted to a set of samples, $\{\boldsymbol{\xi}_i\}_{i=1}^M$, generated independently and randomly according to an unknown distribution F. Under such conditions, a common approach is to assume that the true moments lie in a neighborhood of their respective empirical estimates. In what follows, we will show how one can define a confidence region for the mean and the covariance matrix such that it is assured with high probability to contain the mean and covariance matrix of the distribution of $\boldsymbol{\xi}$. This result will in turn be used to derive a distributional set of the form $\mathcal{D}_1$ and will provide probabilistic guarantees that the solution found using our proposed DRSP model is robust with respect to the true distribution of the random vector.

In order to simplify the derivations, we start by reformulating the random vector $\boldsymbol{\xi}$ in terms of a linear combination of uncorrelated random variables. More specifically, given the random vector $\boldsymbol{\xi} \in \mathbb{R}^m$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} \succ 0$, let us define $\boldsymbol{\zeta} \in \mathbb{R}^m$ to be the normalized random vector $\boldsymbol{\zeta} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\xi} - \boldsymbol{\mu})$ such that $\mathbb{E}[\boldsymbol{\zeta}] = 0$ and $\mathbb{E}[\boldsymbol{\zeta}\boldsymbol{\zeta}^{\mathsf{T}}] = \boldsymbol{I}$. Also, let us make the following assumption about $\boldsymbol{\zeta}$:

ASSUMPTION 5. *There exists a ball of radius R that contains the entire support of the unknown distribution of $\boldsymbol{\zeta}$. More specifically, there exist an $R \geq 0$ such that*

$$\mathbb{P}\left((\boldsymbol{\xi} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\xi} - \boldsymbol{\mu}) \leq R^2\right) = 1 \ .$$

In practice, even when one does not have information about $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we believe that one can often still make an educated and conservative guess about the magnitude of R. We will also revisit this issue in Section 3.3 where R will be derived based on the bounded support of $\boldsymbol{\xi}$ and a set of samples $\{\boldsymbol{\xi}_i\}_{i=1}^M$. In what follows, a confidence region for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ will be resolved based on Assumption 5 and on an inequality known as the “independent bounded differences inequality”, which was popularized by McDiarmid.[2] In fact, this inequality can be seen as a generalized version of Hoeffding’s inequality.

THEOREM 1. *(McDiarmid (1998)) Let $\{\boldsymbol{\xi}_i\}_{i=1}^M$ be a set of independent random vectors $\boldsymbol{\xi}_i$ taking values in a set $\mathcal{S}_i$ for each i. Suppose that the real-valued function $g(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, ..., \boldsymbol{\xi}_M)$ defined on $\mathcal{S}_1 \times \mathcal{S}_2 \times ... \times \mathcal{S}_M$ satisfies*

$$|g(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, ..., \boldsymbol{\xi}_M) - g(\boldsymbol{\xi}'_1, \boldsymbol{\xi}'_2, ..., \boldsymbol{\xi}'_M)| \leq c_j \tag{10}$$

whenever the vector sets $\{\boldsymbol{\xi}_i\}_{i=1}^M$ and $\{\boldsymbol{\xi}'_i\}_{i=1}^M$ differ only in the j-th vector. Then for any $t \geq 0$,

$$\mathbb{P}\left(g(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, ..., \boldsymbol{\xi}_M) - \mathbb{E}[g(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, ..., \boldsymbol{\xi}_M)] \leq -t\right) \leq \exp\left(\frac{-2t^2}{\sum_{j=1}^M c_j^2}\right) \ .$$

3.1. A Confidence Region for the Mean

A first use of the McDiarmid’s theorem leads to defining an ellipsoidal constraint relating the empirical estimate $\hat{\boldsymbol{\mu}} = M^{-1}\sum_{i=1}^M \boldsymbol{\xi}_i$ to the true mean and true covariance matrix of the random vector $\boldsymbol{\xi}$.

The following result is an interesting consequence of McDiarmid’s theorem.

LEMMA 4. *(Shawe-Taylor and Cristianini (2003)) Let $\{\boldsymbol{\zeta}_i\}_{i=1}^M$ be a set of M samples generated independently at random according to the distribution of $\boldsymbol{\zeta}$. If $\boldsymbol{\zeta}$ satisfies Assumption 5 then with probability at least $(1-\delta)$ over the choice of samples $\{\boldsymbol{\zeta}_i\}_{i=1}^M$, we have that*

$$\left\| \frac{1}{M}\sum_{i=1}^M \boldsymbol{\zeta}_i \right\|^2 \leq \frac{R^2}{M}\left(2 + \sqrt{2\ln(1/\delta)}\right)^2 \ .$$

This result can in turn be used to derive a similar statement about the random vector $\boldsymbol{\xi}$.

COROLLARY 1. *Let $\{\boldsymbol{\xi}_i\}_{i=1}^M$ be a set of M samples generated independently at random according to the distribution of $\boldsymbol{\xi}$. If $\boldsymbol{\xi}$ satisfies Assumption 5, then with probability greater than $1-\delta$, we have that*

$$(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \le \beta(\delta) \ \ , \tag{11}$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{M}\sum_{i=1}^M \boldsymbol{\xi}_i$ and $\beta(\delta) = (R^2/M)\left(2+\sqrt{2\ln(1/\delta)}\right)^2$.

Proof: This generalization for a $\boldsymbol{\xi}$ with arbitrary mean and covariance matrix is quite straightforward:

$$\begin{aligned}\mathbb{P}\left((\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \le \beta(\delta)\right) &= \mathbb{P}\left(\left\|\boldsymbol{\Sigma}^{-1/2}\left(\frac{1}{M}\sum_{i=1}^M \boldsymbol{\xi}_i - \boldsymbol{\mu}\right)\right\|^2 \le \beta(\delta)\right) \\ &= \mathbb{P}\left(\left\|\sum_{i=1}^M \boldsymbol{\zeta}_i\right\|^2 \le \beta(\delta)\right) \ge 1-\delta \ . \quad \square\end{aligned}$$

Since $\boldsymbol{\Sigma}$ is non-singular, the inequality of Equation (11) constrains the vector $\boldsymbol{\mu}$ and matrix $\boldsymbol{\Sigma}$ to a convex set. This set can be represented by the following linear matrix inequality after applying the principles of Schur's complement:

$$\begin{bmatrix} \boldsymbol{\Sigma} & (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \\ (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\mathsf{T} & \beta(\delta) \end{bmatrix} \succeq 0 \ \ .$$

3.2. New Confidence Region for the Covariance Matrix

In order for Constraint (11) to describe a bounded set, one must be able to bound the uncertainty in $\boldsymbol{\Sigma}$. While confidence regions for the covariance matrix are typically defined in terms of bounding the sum of square differences between each term of the matrix and its estimate (see for example Shawe-Taylor and Cristianini (2003)), we favor the structure imposed by two linear matrix inequalities bounding $\boldsymbol{\Sigma}$ around its empirical estimate $\hat{\boldsymbol{\Sigma}} = M^{-1}\sum_{i=1}^M (\boldsymbol{\xi}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{\xi}_i - \hat{\boldsymbol{\mu}})^\mathsf{T}$:

$$\mathbb{P}\left(c_{\min}\hat{\boldsymbol{\Sigma}} \preceq \boldsymbol{\Sigma} \preceq c_{\max}\hat{\boldsymbol{\Sigma}}\right) \ge 1-\delta \ \ . \tag{12}$$

Note that the difficulty of this task is mostly due to the fact that one needs to derive a confidence interval for the eigenvalues of the stochastic matrix $\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$. For the case that interests us, where $M \gg m$ with M finite and m fixed, prior work on this topic usually assumes $\boldsymbol{\xi}$ is a normally distributed random vector (see Anderson (1984) and Edelman (1989)). Under the Gaussian assumption, the empirical covariance matrix follows the Wishart distribution, thus one can formulate the distribution of eigenvalues in a closed form expression and derive such percentile bounds. In the case where $\boldsymbol{\xi}$ takes a non-normal form, the asymptotic distribution of eigenvalues was studied in Waternaux (1976) and Fujikoshi (1980) among others. However, to the best of our knowledge, our work is the first to formulate a confidence region with the characteristics presented in Equation (12) for a sample set of finite size. In what follows, we start by demonstrating how a confidence region of the form presented in Equation (12) can be defined around $\hat{\boldsymbol{I}} = M^{-1}\sum_i \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\mathsf{T}$ for the covariance matrix of $\boldsymbol{\zeta}$. Next, we will assume that the mean of $\boldsymbol{\xi}$ is exactly known and we will formulate the confidence region for $\boldsymbol{\Sigma}$ in terms of $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\mu}) = M^{-1}\sum_{i=1}^M(\boldsymbol{\xi}_i - \boldsymbol{\mu})(\boldsymbol{\xi}_i - \boldsymbol{\mu})^\mathsf{T}$. We conclude this section with our main result about a confidence region for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which relies solely on the M samples and on support information about the random vector $\boldsymbol{\xi}$.

LEMMA 5. *Let $\{\boldsymbol{\zeta}_i\}_{i=1}^M$ be a set of M samples generated independently at random according to the distribution of $\boldsymbol{\zeta}$. If $\boldsymbol{\zeta}$ satisfies Assumption 5, then with probability greater than $1-\delta$, we have that*

$$\frac{1}{1+\alpha(\delta/2)}\hat{\boldsymbol{I}} \preceq \boldsymbol{I} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\boldsymbol{I}} \ \ , \tag{13}$$

where $\alpha(\delta/2) = (R^2/\sqrt{M})\left(\sqrt{1-m/R^4} + \sqrt{\ln(2/\delta)}\right)$*, provided that*

$$M > R^4\left(\sqrt{1-m/R^4} + \sqrt{\ln(2/\delta)}\right)^2 \quad . \tag{14}$$

Proof: The proof of this theorem relies on applying Theorem 1 twice to show that both $\frac{1}{1+\alpha(\delta/2)}\hat{\boldsymbol{I}} \preceq \boldsymbol{I}$ and $\boldsymbol{I} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\boldsymbol{I}}$ occur with probability greater than $1-\delta/2$. Our statement then simply follows by the union bound. However, for the sake of conciseness, this proof will focus on deriving the upper bound since the steps that we follow can easily be modified for the derivation of the lower bound.

When applying Theorem 1 to show that $\boldsymbol{I} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\boldsymbol{I}}$ occurs with probability greater than $1-\delta/2$, the main step consists of defining $g(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, ..., \boldsymbol{\zeta}_M) = \min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}\boldsymbol{z}$ and finding a lower bound for $\mathbb{E}\left[g(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, ..., \boldsymbol{\zeta}_M)\right]$. One can start by showing that Constraint (10) is met when $c_j = R^2/M$ for all j.

$$|g(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, ..., \boldsymbol{\zeta}_M) - g(\boldsymbol{\zeta}'_1, \boldsymbol{\zeta}'_2, ..., \boldsymbol{\zeta}'_M)| = \left|\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}\boldsymbol{z} - \min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}'\boldsymbol{z}\right| \quad ,$$

where $\hat{\boldsymbol{I}}' = \frac{1}{M}\sum_{i=1}^M \boldsymbol{\zeta}'_i{\boldsymbol{\zeta}'_i}^{\mathsf{T}} = \hat{\boldsymbol{I}} + \frac{1}{M}(\boldsymbol{\zeta}'_j{\boldsymbol{\zeta}'_j}^{\mathsf{T}} - \boldsymbol{\zeta}_j\boldsymbol{\zeta}_j^{\mathsf{T}})$ since $\{\boldsymbol{\zeta}_i\}_{i=1}^M$ and $\{\boldsymbol{\zeta}'_i\}_{i=1}^M$ only differ in the j-th vector.

Now assume that $\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}\boldsymbol{z} \geq \min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}'\boldsymbol{z}$. Then, for any $\boldsymbol{z}^* \in \arg\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}'\boldsymbol{z}$

$$\begin{aligned}
|g(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, ..., \boldsymbol{\zeta}_M) - g(\boldsymbol{\zeta}'_1, \boldsymbol{\zeta}'_2, ..., \boldsymbol{\zeta}'_M))| &= \min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}\boldsymbol{z} - {\boldsymbol{z}^*}^{\mathsf{T}}\hat{\boldsymbol{I}}'\boldsymbol{z}^* \\
&\leq {\boldsymbol{z}^*}^{\mathsf{T}}(\hat{\boldsymbol{I}} - \hat{\boldsymbol{I}}')\boldsymbol{z}^* \\
&= {\boldsymbol{z}^*}^{\mathsf{T}}\frac{1}{M}(\boldsymbol{\zeta}_j\boldsymbol{\zeta}_j^{\mathsf{T}} - \boldsymbol{\zeta}'_j\boldsymbol{\zeta}_j'^{\mathsf{T}})\boldsymbol{z}^* \\
&= \frac{1}{M}\left((\boldsymbol{\zeta}_j^{\mathsf{T}}\boldsymbol{z}^*)^2 - (\boldsymbol{\zeta}_j'^{\mathsf{T}}\boldsymbol{z}^*)^2\right) \\
&\leq \frac{\|\boldsymbol{z}^*\|^2\|\boldsymbol{\zeta}_j\|^2}{M} \leq \frac{R^2}{M} \quad .
\end{aligned}$$

Otherwise, in the case that $\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}\boldsymbol{z} \leq \min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}'\boldsymbol{z}$, the same argument applies using $\boldsymbol{z}^* \in \arg\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}\boldsymbol{z}$.

The task of bounding $\mathbb{E}\left[g(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, ..., \boldsymbol{\zeta}_M)\right]$ is a bit harder. We can instead start by finding an upper bound on the expected maximum eigenvalue of $(\boldsymbol{I} - \hat{\boldsymbol{I}})$ since

$$\mathbb{E}\left[\max_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}(\boldsymbol{I} - \hat{\boldsymbol{I}})\boldsymbol{z}\right] = 1 - \mathbb{E}\left[\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}\hat{\boldsymbol{I}}\boldsymbol{z}\right] \quad , \tag{15}$$

where the expectation is taken with respect to the random matrix $\hat{\boldsymbol{I}}$. Using Jensen's inequality and basic linear algebra, one can show that

$$\begin{aligned}
\left(\mathbb{E}\left[\max_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}(\boldsymbol{I} - \hat{\boldsymbol{I}})\boldsymbol{z}\right]\right)^2 &\leq \mathbb{E}\left[\left(\max_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathsf{T}}(\boldsymbol{I} - \hat{\boldsymbol{I}})\boldsymbol{z}\right)^2\right] \leq \mathbb{E}\left[\sum_{i=1}^m \sigma_i^2(\boldsymbol{I} - \hat{\boldsymbol{I}})\right] = \mathbb{E}\left[\mathbf{trace}\left(\left(\boldsymbol{I} - \hat{\boldsymbol{I}}\right)^2\right)\right] \\
&= \mathbb{E}\left[\mathbf{trace}\left(\left(\frac{1}{M}\sum_{i=1}^M\left(\boldsymbol{I} - \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^{\mathsf{T}}\right)\right)^2\right)\right] \\
&= \mathbf{trace}\left(\frac{1}{M^2}\sum_{i=1}^M \mathbb{E}\left[\boldsymbol{I} - 2\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^{\mathsf{T}} + \left(\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^{\mathsf{T}}\right)^2\right]\right) \\
&= \frac{1}{M}\left(\mathbf{trace}\left(\mathbb{E}\left[\left(\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^{\mathsf{T}}\right)^2\right]\right) - \mathbf{trace}\left(\boldsymbol{I}\right)\right) = \frac{\mathbb{E}\left[\|\boldsymbol{\zeta}_i\|^4\right] - m}{M} \leq \frac{R^4 - m}{M} \quad ,
\end{aligned}$$

where $\sigma_i(\boldsymbol{I}-\hat{\boldsymbol{I}})$ refers to the i-th singular value of $\boldsymbol{I}-\hat{\boldsymbol{I}}$. The derivation above uses the fact that $\boldsymbol{\zeta}_i$ are sampled independently thus making $\mathbb{E}\left[(\boldsymbol{I}-\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\mathsf{T})(\boldsymbol{I}-\boldsymbol{\zeta}_j\boldsymbol{\zeta}_j^\mathsf{T})\right]=\mathbb{E}\left[\boldsymbol{I}-\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\mathsf{T}\right]\mathbb{E}\left[\boldsymbol{I}-\boldsymbol{\zeta}_j\boldsymbol{\zeta}_j^\mathsf{T}\right]=0$.

By replacing this lower bound in Equation (15), we have that $\mathbb{E}\left[g(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2,...,\boldsymbol{\zeta}_M)\right]\geq 1-(R^2/\sqrt{M})\sqrt{1-m/R^4}$. More importantly, Theorem 1 allows us to confirm the proposed upper bound using the following argument. Since the statement

$$\mathbb{P}\left(\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{I}}\boldsymbol{z}-\mathbb{E}\left[\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{I}}\boldsymbol{z}\right]\leq-\epsilon\right)\leq\exp\left(\frac{-2\epsilon^2}{\sum_{j=1}^{M}(R^4/M^2)}\right)\ ,$$

implies that

$$\mathbb{P}\left(\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{I}}\boldsymbol{z}-\mathbb{E}\left[\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{I}}\boldsymbol{z}\right]\geq-\frac{R^2\sqrt{\ln(2/\delta)}}{\sqrt{M}}\right)\geq 1-\delta/2\ ,$$

and since replacing $\mathbb{E}\left[\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{I}}\boldsymbol{z}\right]$ to its lower bound can only include more random events, we necessarily have that

$$\mathbb{P}\left(\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{I}}\boldsymbol{z}\geq 1-\frac{R^2}{\sqrt{M}}\left(\sqrt{1-m/R^4}+\sqrt{\ln(2/\delta)}\right)\right)\geq 1-\delta/2\ .$$

Thus, given that M is large enough such that $1-\alpha(\delta/2)>0$, we can conclude that

$$\mathbb{P}\left(\boldsymbol{I}\preceq\frac{1}{1-\alpha(\delta/2)}\hat{\boldsymbol{I}}\right)\geq 1-\delta/2\ .$$

The task of showing that $1/(1+\alpha(\delta/2))\hat{\boldsymbol{I}}\preceq\boldsymbol{I}$ also occurs with probability $1-\delta/2$ is very similar. One needs to apply Theorem 1, now defining $g(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2,...,\boldsymbol{\zeta}_M)=-\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{I}}\boldsymbol{z}$, and to demonstrate that $\mathbb{E}\left[g(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2,...,\boldsymbol{\zeta}_M)\right]\geq-1-\alpha(\delta/2)$. The rest follows easily. □

REMARK 4. Considering the case where ζ is actually a random variable and where $\hat{\sigma}^2=\sum_{i=1}^{M}\zeta_i^2$, one can easily verify that the parameter $\alpha(\delta)$ in Lemma 5 is asymptotically of the right order in terms of M and R. Since $\mathbb{E}\left[\zeta^4\right]$ is bounded by R^4, the central limit theorem guarantees that $\sqrt{M}(\hat{\sigma}^2-\mathbb{E}\left[\zeta^2\right])$ converges in distribution to $\mathcal{N}(0,\mathbb{E}\left[\zeta^4\right]-1)$. Thus, it follows that the distribution of $\left(M/(\mathbb{E}\left[\zeta^4\right]-1)\right)\left(\hat{\sigma}^2-\mathbb{E}\left[\zeta^2\right]\right)^2$ converges to a χ^2-distribution with degree 1. For any $\delta>0$, one can find $c(\delta)$ such that $|\hat{\sigma}^2-\mathbb{E}\left[\zeta^2\right]|\leq\frac{c(\delta)\sqrt{\mathbb{E}\left[\zeta^4\right]-1}}{\sqrt{M}}$ is satisfied with probability greater than $1-\delta$. Hence, asymptotically the confidence region $-\left(1+\frac{c(\delta)R^2}{\sqrt{M}}\right)^{-1}\hat{\sigma}^2\ \leq\ 1\ \leq\ \left(1-\frac{c(\delta)R^2}{\sqrt{M}}\right)^{-1}\hat{\sigma}^2$ is tight.

We are now interested in extending Lemma 5 to a random vector with general mean and covariance matrix. Given the random event that Constraint (13) is satisfied, then:

$$\begin{aligned}\boldsymbol{I}\preceq\frac{1}{1-\alpha(\delta/2)}\hat{\boldsymbol{I}}&\Rightarrow\boldsymbol{\Sigma}^{1/2}\boldsymbol{I}\boldsymbol{\Sigma}^{1/2}\preceq\frac{1}{1-\alpha(\delta/2)}\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{I}}\boldsymbol{\Sigma}^{1/2}\\&\Rightarrow\boldsymbol{\Sigma}\preceq\frac{1}{1-\alpha(\delta/2)}\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\mathsf{T}\boldsymbol{\Sigma}^{1/2}\\&\Rightarrow\boldsymbol{\Sigma}\preceq\frac{1}{1-\alpha(\delta/2)}\frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{\xi}_i-\boldsymbol{\mu})(\boldsymbol{\xi}_i-\boldsymbol{\mu})^\mathsf{T}\\&\Rightarrow\boldsymbol{\Sigma}\preceq\frac{1}{1-\alpha(\delta/2)}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\mu})\ ,\end{aligned}$$

and similarly,

$$\frac{1}{1+\alpha(\delta/2)}\hat{\boldsymbol{I}}\preceq\boldsymbol{I}\Rightarrow\frac{1}{1+\alpha(\delta/2)}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\mu})\preceq\boldsymbol{\Sigma}\ .$$

Since Constraint (13) is satisfied with probability greater than $1-\delta$, the following corollary follows easily.

COROLLARY 2. *Let $\{\boldsymbol{\xi}_i\}_{i=1}^M$ be a set of M samples generated independently at random according to the distribution of $\boldsymbol{\xi}$. If $\boldsymbol{\xi}$ satisfies Assumption 5 and M satisfies Constraint* (14)*, then with probability greater then $1-\delta$, we have that*

$$\frac{1}{1+\alpha(\delta/2)}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\mu}) \preceq \boldsymbol{\Sigma} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\mu}) \ ,$$

where $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\mu}) = \frac{1}{M}\sum_{i=1}^M(\boldsymbol{\xi}_i-\boldsymbol{\mu})(\boldsymbol{\xi}_i-\boldsymbol{\mu})^\mathsf{T}$ and $\alpha(\delta/2)$ is defined as in Lemma 5.

Combined with Corollary 1, this statement leads to the description of a convex set which is constructed using empirical estimates of the mean and covariance matrix, and yet is guaranteed to contain the true mean and covariance matrix of $\boldsymbol{\xi}$ with high probability.

THEOREM 2. *Let $\{\boldsymbol{\xi}_i\}_{i=1}^M$ be a set of M samples generated independently at random according to the distribution of $\boldsymbol{\xi}$. If $\boldsymbol{\xi}$ satisfies Assumption 5 and M satisfies Equation* (14)*, then with probability greater than $1-\delta$ over the choice of $\{\boldsymbol{\xi}_i\}_{i=1}^M$, the following set of constraints are met:*

$$(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \le \beta(\delta/2) \tag{16a}$$

$$\boldsymbol{\Sigma} \preceq \frac{1}{1-\alpha(\delta/4)-\beta(\delta/2)}\hat{\boldsymbol{\Sigma}} \tag{16b}$$

$$\boldsymbol{\Sigma} \succeq \frac{1}{1+\alpha(\delta/4)}\hat{\boldsymbol{\Sigma}} \ , \tag{16c}$$

where $\hat{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{i=1}^M(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}})^\mathsf{T}$, $\alpha(\delta/4) = (R^2/\sqrt{M})\left(\sqrt{1-m/R^4}+\sqrt{\ln(4/\delta)}\right)$, $\beta(\delta/2) = (R^2/M)\left(2+\sqrt{2\ln(2/\delta)}\right)^2$.

Proof: By applying Corollary 1, 2, and Lemma 5, the union bound guarantees us with probability greater than $1-\delta$ that the following constraints are met:

$$\begin{aligned}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) &\le \beta(\delta/2)\\ \boldsymbol{\Sigma} &\preceq \frac{1}{1-\alpha(\delta/4)}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\mu})\\ \boldsymbol{\Sigma} &\succeq \frac{1}{1+\alpha(\delta/4)}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\mu}) \ .\end{aligned}$$

Note that our result is not proven yet since, although the first constraint is exactly Constraint (16a), the second and third constraints actually refer to a covariance matrix estimate that uses the true mean of the distribution instead of its empirical estimate. The following steps will convince us that these conditions are sufficient for Constraint (16b) and Constraint (16c) to hold:

$$\begin{aligned}(1-\alpha(\delta/4))\boldsymbol{\Sigma} \preceq \hat{\boldsymbol{\Sigma}}(\boldsymbol{\mu}) &= \frac{1}{M}\sum_{i=1}^M(\boldsymbol{\xi}_i-\boldsymbol{\mu})(\boldsymbol{\xi}_i-\boldsymbol{\mu})^\mathsf{T} = \frac{1}{M}\sum_{i=1}^M(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}}+\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}}+\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\mathsf{T}\\ &= \frac{1}{M}\sum_{i=1}^M(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}})^\mathsf{T}+(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\mathsf{T}+(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}})^\mathsf{T}+(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\mathsf{T}\\ &= \hat{\boldsymbol{\Sigma}}+(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\mathsf{T} \preceq \hat{\boldsymbol{\Sigma}}+\beta(\delta/2)\boldsymbol{\Sigma} \ ,\end{aligned}$$

where the last semi-definite inequality of the derivation can be explained using the fact that for any $\boldsymbol{x}\in\mathbb{R}^m$,

$$\begin{aligned}\boldsymbol{x}^\mathsf{T}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{x} &= (\boldsymbol{x}^\mathsf{T}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}))^2 = \left(\boldsymbol{x}^\mathsf{T}\boldsymbol{\Sigma}^{1/2}\,\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\right)^2\\ &\le \|\boldsymbol{x}^\mathsf{T}\boldsymbol{\Sigma}^{1/2}\|^2\|\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\|^2 \le \beta(\delta/2)\boldsymbol{x}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{x} \ .\end{aligned}$$

Thus we can conclude that Constraint (16b) is met. Similar steps can be used to show that Constraint (16c) also holds:

$$(1+\alpha(\delta/4))\boldsymbol{\Sigma} \succeq \hat{\boldsymbol{\Sigma}}(\boldsymbol{\mu}) = \frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{\xi}_i-\boldsymbol{\mu})(\boldsymbol{\xi}_i-\boldsymbol{\mu})^{\mathsf{T}} = \hat{\boldsymbol{\Sigma}} + (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^{\mathsf{T}} \succeq \hat{\boldsymbol{\Sigma}}\ . \quad \square$$

3.3. Bounding the Support of ζ using Empirical Data

The above derivations assumed that one can describe a ball containing the support of the “fictitious” random vector $\boldsymbol{\zeta}$. In fact, this assumption can be replaced by an assumption on the support of the more tangible random vector $\boldsymbol{\xi}$ as is presented in the following corollary.

COROLLARY 3. *Let $\{\boldsymbol{\xi}_i\}_{i=1}^{M}$ be a set of M samples generated independently at random according to the distribution of $\boldsymbol{\xi}$. Given that the support of the distribution of $\boldsymbol{\xi}$ is known to be contained in $\mathcal{S}_{\boldsymbol{\xi}}$, let*

$$\hat{R} = \sup_{\boldsymbol{\xi}\in\mathcal{S}_{\boldsymbol{\xi}}} \|\hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{\xi}-\hat{\boldsymbol{\mu}})\|_2$$

be a sample-based approximation of R and for any $\delta>0$, let

$$\bar{R} = \left(1-(\hat{R}^2+2)\frac{2+\sqrt{2\ln(4/\bar{\delta})}}{\sqrt{M}}\right)^{-1/2}\hat{R}\ ,$$

where $\bar{\delta}=1-\sqrt{1-\delta}$. If

$$M > \max\left\{(\hat{R}^2+2)^2\left(2+\sqrt{2\ln(4/\bar{\delta})}\right)^2\ ,\ \frac{\left(8+\sqrt{32\ln(4/\bar{\delta})}\right)^2}{\left(\sqrt{\hat{R}+4}-\hat{R}\right)^4}\right\}\ , \tag{17}$$

then with probability greater than $1-\delta$, constraints (16a), (16b), *and* (16c) *are satisfied with $\alpha(\delta/4)$ and $\beta(\delta/2)$ replaced with $\bar{\alpha}(\bar{\delta}/4)=(\bar{R}^2/\sqrt{M})\left(\sqrt{1-m/\bar{R}^4}+\sqrt{\ln(4/\bar{\delta})}\right)$ and $\bar{\beta}(\bar{\delta}/2)=(\bar{R}^2/M)\left(2+\sqrt{2\ln(2/\bar{\delta})}\right)^2$ respectively.*

Proof: Since we assumed that $\boldsymbol{\Sigma}$ was non-singular, the fact that the support of $\boldsymbol{\xi}$ is bounded by a ball of radius $R_{\boldsymbol{\xi}}$ implies that $\boldsymbol{\zeta}$ is also bounded. Thus, there exists an R such that $\mathbb{P}(\|\boldsymbol{\zeta}\|\leq R)=1$. Given that $\boldsymbol{\zeta}$ has a bounded support and given Condition (17), Theorem 4 guarantees us that with probability greater than $1-\bar{\delta}$, constraints (16a), (16b), and (16c) are met. Thus

$$\begin{aligned} R &= \sup_{\boldsymbol{\zeta}\in\mathcal{S}_{\boldsymbol{\zeta}}}\|\boldsymbol{\zeta}\| = \sup_{\boldsymbol{\xi}\in\mathcal{S}_{\boldsymbol{\xi}}}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\xi}-\boldsymbol{\mu})\| = \sup_{\boldsymbol{\xi}\in\mathcal{S}_{\boldsymbol{\xi}}}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\xi}-\boldsymbol{\mu}+\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\mu}})\| \\ &\leq \sup_{\boldsymbol{\xi}\in\mathcal{S}_{\boldsymbol{\xi}}}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\xi}-\hat{\boldsymbol{\mu}})\| + \|\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\| \\ &\leq \sup_{\boldsymbol{\xi}\in\mathcal{S}_{\boldsymbol{\xi}}}\sqrt{1+\alpha(\bar{\delta}/4)}\|\hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{\xi}-\hat{\boldsymbol{\mu}})\| + \sqrt{\beta(\bar{\delta}/2)} \\ &\leq \sqrt{1+\alpha(\bar{\delta}/4)}\hat{R} + \sqrt{\beta(\bar{\delta}/2)} \\ &\leq \hat{R}\sqrt{1+cR^2}+cR\ , \end{aligned}$$

where $c=\left(2+\sqrt{2\ln(4/\bar{\delta})}\right)/\sqrt{M}$.

A careful analysis of the function $\psi(R,\hat{R})=\hat{R}\sqrt{1+cR^2}+cR$ leads to the observation that if M satisfies Condition (17) then the fact that $R\leq\psi(R,\hat{R})$ necessarily implies that $R\leq\bar{R}$. We can therefore conclude that $\mathbb{P}(R\leq\bar{R})\geq 1-\bar{\delta}$.

Given the event that $R \leq \bar{R}$ occurs, since

$$\begin{aligned}\alpha(\bar{\delta}/4) &= (R^2/\sqrt{M})\left(\sqrt{1-m/R^4}+\sqrt{2\ln(4/\bar{\delta})}\right)\\ &\leq (\bar{R}^2/\sqrt{M})\left(\sqrt{1-m/\bar{R}^4}+\sqrt{2\ln(4/\bar{\delta})}\right) = \bar{\alpha}(\bar{\delta}/4)\end{aligned}$$

and since

$$\beta(\bar{\delta}/2) = (R^2/M)\left(2+\sqrt{2\ln(2/\bar{\delta})}\right)^2 \leq (\bar{R}^2/M)\left(2+\sqrt{2\ln(2/\bar{\delta})}\right)^2 = \bar{\beta}(\bar{\delta}/2) \ \ ,$$

we can conclude with a second application of Theorem 2 that with probability greater than $1-\bar{\delta}$ the following statements are satisfied:

$$\begin{aligned}&(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \leq \beta(\bar{\delta}/2) \leq \bar{\beta}(\bar{\delta}/2) \ \ ,\\ &\boldsymbol{\Sigma} \preceq \frac{1}{1-\alpha(\delta/4)-\beta(\delta/2)}\hat{\boldsymbol{\Sigma}} \preceq \frac{1}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)}\hat{\boldsymbol{\Sigma}} \ \ ,\\ &\boldsymbol{\Sigma} \succeq \frac{1}{1-\alpha(\delta/4)}\hat{\boldsymbol{\Sigma}} \succeq \frac{1}{1-\bar{\alpha}(\delta/4)}\hat{\boldsymbol{\Sigma}} \ \ .\end{aligned}$$

It follows that Theorem 2 applies with $\bar{\alpha}(\bar{\delta}/4)$ and $\bar{\beta}(\bar{\delta}/4)$ because the probability that the event $\mathcal{E}$ that constraints (16a), (16b), and (16c) equipped with $\bar{\alpha}(\bar{\delta}/4)$ and $\bar{\beta}(\bar{\delta}/4)$ are met is necessarily greater than $1-\delta$:

$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}(\mathcal{E}|R \leq \bar{R})\mathbb{P}(R \leq \bar{R}) \geq (1-\bar{\delta})(1-\bar{\delta}) = 1-\delta \ \ . \quad \square$$

3.4. Data-driven DRSP Optimization

In some practical situations where one wishes to formulate a DRSP model, it might not be clear how to define an uncertainty set for the mean and second moment matrix of the random vector of parameters $\boldsymbol{\xi}$. It is more likely the case that one only has in hand a set of independent samples, $\{\boldsymbol{\xi}_i\}_{i=1}^M$, drawn according to the distribution of $\boldsymbol{\xi}$ and wishes to guarantee that the solution of the DRSP model is robust with respect to what the unknown distribution of the random vector $\boldsymbol{\xi}$ might be.

We will first use our recent result to define, based on the samples $\{\boldsymbol{\xi}_i\}_{i=1}^M$, a set of distributions which is known to contain the distribution of $\boldsymbol{\xi}$ with high probability, given that M is sufficiently large.

DEFINITION 2. Given a set $\{\boldsymbol{\xi}_i\}_{i=1}^M$ of M samples, for any $\delta > 0$ let $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, $\bar{\gamma}_1$ and $\bar{\gamma}_2$ be defined as

$$\hat{\boldsymbol{\mu}} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\xi}_i \ \ , \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}})^{\mathsf{T}}$$

$$\bar{\gamma}_1 = \frac{\bar{\beta}(\bar{\delta}/2)}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)} \ \ , \qquad \bar{\gamma}_2 = \frac{1+\bar{\beta}(\bar{\delta}/2)}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)} \ \ .$$

where $\bar{\alpha}(\bar{\delta}/4) = O(1/\sqrt{M})$ and $\bar{\beta}(\bar{\delta}/2) = O(1/M)$ are constants defined in Corollary 3.

Note that it follows from Definition 2 that $\bar{\gamma}_1 \rightarrow 0$ and $\bar{\gamma}_2 \rightarrow 1$ as M goes to infinity.

COROLLARY 4. *Let* $\{\boldsymbol{\xi}_i\}_{i=1}^M$ *be a set of* M *samples generated independently at random according to the distribution of* $\boldsymbol{\xi}$*. If* M *satisfies Constraint* (17) *and* $\boldsymbol{\xi}$ *has a support contained in a bounded set* $\mathcal{S}$*, then with probability greater than* $1-\delta$ *over the choice of* $\{\boldsymbol{\xi}_i\}_{i=1}^M$*, the distribution of* $\boldsymbol{\xi}$ *lies in the set* $\mathcal{D}_1(\mathcal{S},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}},\bar{\gamma}_1,\bar{\gamma}_2)$.

Proof: This result can be derived from Corollary 3. One can show that given any estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ that satisfy both constraints (16a) and (16b) equipped with $\bar{\alpha}(\bar{\delta}/4)$ and $\bar{\beta}(\bar{\delta}/2)$, these estimates should also satisfy constraints (1a) and (1b). First, Constraint (1a) is necessarily met since for such $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$,

$$(1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2))(\hat{\boldsymbol{\mu}}-\mathbb{E}[\boldsymbol{\xi}])^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}-\mathbb{E}[\boldsymbol{\xi}]) \leq (\hat{\boldsymbol{\mu}}-\mathbb{E}[\boldsymbol{\xi}])^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}-\mathbb{E}[\boldsymbol{\xi}]) \leq \bar{\beta}(\bar{\delta}/2)\ ,$$

where we used the fact that Constraint (16a) implies that $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \geq (1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2))\boldsymbol{x}^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x}$ for any $\boldsymbol{x}\in\mathbb{R}^m$. Similarly, the same $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ can be shown to satisfy Constraint (1b):

$$\begin{aligned}\frac{1}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)}\hat{\boldsymbol{\Sigma}} \succeq \boldsymbol{\Sigma} &= \mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}}]-\mathbb{E}[\boldsymbol{\xi}]\mathbb{E}[\boldsymbol{\xi}]^{\mathsf{T}} \\ &\succeq \mathbb{E}[(\boldsymbol{\xi}-\hat{\boldsymbol{\mu}})(\boldsymbol{\xi}-\hat{\boldsymbol{\mu}})^{\mathsf{T}}]-\frac{\bar{\beta}(\bar{\delta}/2)}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)}\hat{\boldsymbol{\Sigma}}\ ,\end{aligned}$$

since for all $\boldsymbol{x}\in\mathbb{R}^m$,

$$\begin{aligned}\boldsymbol{x}^{\mathsf{T}}\mathbb{E}[\boldsymbol{\xi}]\mathbb{E}[\boldsymbol{\xi}]^{\mathsf{T}}\boldsymbol{x} &= \left(\boldsymbol{x}^{\mathsf{T}}(\mathbb{E}[\boldsymbol{\xi}]-\hat{\boldsymbol{\mu}}+\hat{\boldsymbol{\mu}})\right)^2 = \left(\boldsymbol{x}^{\mathsf{T}}(\mathbb{E}[\boldsymbol{\xi}]-\hat{\boldsymbol{\mu}})\right)^2+2\boldsymbol{x}^{\mathsf{T}}(\mathbb{E}[\boldsymbol{\xi}]-\hat{\boldsymbol{\mu}})\hat{\boldsymbol{\mu}}^{\mathsf{T}}\boldsymbol{x}+\left(\boldsymbol{x}^{\mathsf{T}}\hat{\boldsymbol{\mu}}\right)^2 \\ &= \mathbf{trace}\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}(\mathbb{E}[\boldsymbol{\xi}]-\hat{\boldsymbol{\mu}})(\mathbb{E}[\boldsymbol{\xi}]-\hat{\boldsymbol{\mu}})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\right)+2\boldsymbol{x}^{\mathsf{T}}\mathbb{E}[\boldsymbol{\xi}]\hat{\boldsymbol{\mu}}^{\mathsf{T}}\boldsymbol{x}-(\boldsymbol{x}^{\mathsf{T}}\hat{\boldsymbol{\mu}})^2 \\ &\leq (\mathbb{E}[\boldsymbol{\xi}]-\hat{\boldsymbol{\mu}})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbb{E}[\boldsymbol{\xi}]-\hat{\boldsymbol{\mu}})\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{x}+2\boldsymbol{x}^{\mathsf{T}}\mathbb{E}[\boldsymbol{\xi}]\hat{\boldsymbol{\mu}}^{\mathsf{T}}\boldsymbol{x}-(\boldsymbol{x}^{\mathsf{T}}\hat{\boldsymbol{\mu}})^2 \\ &\leq \boldsymbol{x}^{\mathsf{T}}\left(\frac{\bar{\beta}(\bar{\delta}/2)}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)}\hat{\boldsymbol{\Sigma}}+\mathbb{E}[\boldsymbol{\xi}]\hat{\boldsymbol{\mu}}^{\mathsf{T}}+\hat{\boldsymbol{\mu}}\mathbb{E}[\boldsymbol{\xi}]^{\mathsf{T}}-\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathsf{T}}\right)\boldsymbol{x} \\ &= \boldsymbol{x}^{\mathsf{T}}\left(\frac{\bar{\beta}(\bar{\delta}/2)}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)}\hat{\boldsymbol{\Sigma}}+\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}}]-\mathbb{E}[(\boldsymbol{\xi}-\hat{\boldsymbol{\mu}})(\boldsymbol{\xi}-\hat{\boldsymbol{\mu}})^{\mathsf{T}}]\right)\boldsymbol{x}\ .\end{aligned}$$

By Corollary 3, the random variables $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are known to satisfy constraints (16a) and (16b) with probability greater than $1-\delta$, therefore they also satisfy Constraint (1a) and Constraint (1b) with probability greater than $1-\delta$. □

We can now extend the results presented in Section 2 to a data-driven framework where moments of the distribution are estimated using independent samples. Based on the computational argument of Proposition 2 and the probabilistic guarantees provided by Corollary 4, we present an important result for data-driven problems.

THEOREM 3. *Let $\{\boldsymbol{\xi}_i\}_{i=1}^M$ be a set of M samples generated independently at random according to the distribution F whose support is contained in the set $\mathcal{S}$. For any $\delta>0$, if assumptions 1, 2, 3, and 4 are satisfied then, given the set $\{\boldsymbol{\xi}_i\}_{i=1}^M$, one can solve Problem* (7) *in polynomial time under the set $\mathcal{D}_1(\mathcal{S},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}},\bar{\gamma}_1,\bar{\gamma}_2)$ where $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, $\bar{\gamma}_1$, and $\bar{\gamma}_2$ are assigned as in Definition 2. Furthermore, if M satisfies Constraint* (17), *then with probability greater than $1-\delta$ over the choice of $\{\boldsymbol{\xi}_i\}_{i=1}^M$, we have that any optimal solution $\boldsymbol{x}^*$ of the DRSP formed using these samples will satisfy the constraint*

$$\mathbb{E}[h(\boldsymbol{x}^*,\boldsymbol{\xi})] \leq \Psi(\boldsymbol{x}^*;\bar{\gamma}_1,\bar{\gamma}_2)\ ,$$

where $\mathbb{E}[\cdot]$ is the expectation with respect to the true distribution of $\boldsymbol{\xi}$.

Since we believe the moment problem to be interesting in its own right, we wish to mention a simple consequence of the above result for moment problems in a data-driven framework.

COROLLARY 5. *Let $\delta>0$ and let $\{\boldsymbol{\xi}_i\}_{i=1}^M$ be a set of M samples generated independently at random according to the distribution F which support is contained in the set $\mathcal{S}$. For any $\delta>0$ and function $g(\boldsymbol{\xi})$, if $\mathcal{S}$ satisfies Assumption 1 and the function $h(\boldsymbol{x},\boldsymbol{\xi})=g(\boldsymbol{\xi})$ satisfies Assumption 2 then, given the set $\{\boldsymbol{\xi}_i\}_{i=1}^M$, one can solve in polynomial time the moment problem*

$$\underset{F\ \in\ \mathcal{D}_1(\mathcal{S},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}},\bar{\gamma}_1,\bar{\gamma}_2)}{\text{maximize}}\quad \mathbb{E}_F[g(\boldsymbol{\xi})]\ ,$$

where $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, $\bar{\gamma}_1$ *and* $\bar{\gamma}_2$ *are assigned as in Definition 2. Furthermore, if* M *satisfies Constraint* (17)*, then with probability greater than* $1-\delta$ *over the choice of* $\{\boldsymbol{\xi}_i\}_{i=1}^M$*, we have that*

$$\mathbb{E}\left[g(\boldsymbol{\xi})\right] \;\leq\; \Psi(0;\bar{\gamma}_1,\bar{\gamma}_2)\ ,$$

where $\mathbb{E}\left[\cdot\right]$ *is the expectation with respect to the true distribution of* $\boldsymbol{\xi}$*.*

4. Application to Portfolio Optimization

We now turn ourselves to applying the DRSP framework to an instance of portfolio optimization. In such a problem, one is interested in maximizing his expected utility obtained from the single step return of his investment portfolio. Given that n investment options are available, the expected utility can be defined as $\mathbb{E}\left[u(\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x})\right]$, where $u(\cdot)$ is a non-decreasing function and $\boldsymbol{\xi}\in\mathbb{R}^n$ is a random vector of returns for the different options. In the robust approach to this problem, one defines a distributional set $\mathcal{D}$ that is known to contain the distribution F and chooses the portfolio which is optimal according to the following Distributionally Robust Portfolio Optimization model:

$$\text{(DRPO)}\qquad \underset{\boldsymbol{x}}{\text{maximize}}\quad \min_{F\,\in\,\mathcal{D}}\ \mathbb{E}_F[u(\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x})] \tag{18a}$$

$$\text{subject to}\quad \sum_{i=1}^n x_i = 1\ ,\ \ \boldsymbol{x}\geq 0\ . \tag{18b}$$

In Popescu (2007), the author addressed the case of Problem (18) where $\mathbb{E}\left[\boldsymbol{\xi}\right]$ and $\mathbb{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}^\mathsf{T}\right]$ are known exactly and one considers $\mathcal{D}$ to be the set of all distributions with such first and second moments. Based on these assumptions, the author presents a parametric quadratic programming algorithm that is efficient for a large family of utility functions $u(\cdot)$. This approach is interesting since it provides the means to take into account uncertainty in the form of the distribution of returns. However, our experiments show that in practice it is highly sensitive to the noise in the empirical estimation of these moments (see Section 4.3). The proposed algorithm also relies on solving a one-dimensional non-convex mathematical program; thus, there are no guarantees of finding a near optimal solution in polynomial time. Although the approach that we are about to propose addresses a smaller family of utility functions, it will take into account moment uncertainty and will lead to the formulation of a semi-definite program, which can be solved efficiently using interior point methods.

In Goldfarb and Iyengar (2003), the authors attempt to account for moment uncertainty in Markowitz models. Their motivation is closely aligned with ours and many of the techniques that they propose can be applied in our context: *e.g.*, the use of factor models to reduce the dimensionality of $\boldsymbol{\xi}$. Similarly, the results presented in Section 3 for a data-driven framework should extend easily to the context of Markowitz models. Because Problem (18) reduces to a Markowitz model when the utility function is quadratic and concave, we consider our model to be richer than the one considered in Goldfarb and Iyengar (2003). On the other hand, a robust Markowitz model typically gives rise to a problem that is simpler to solve.

4.1. Portfolio Optimization with Moment Uncertainty

In order to apply our framework, we make the assumption that the utility function is piecewise linear and concave, such that $u(y)=\min_{k\in\{1,2,\ldots,K\}}\ a_k y + b_k$ with $a_k\geq 0$. This assumption is not very limiting since most interesting utility functions are concave and can usually be approximated accurately using simple piecewise linear functions. We use historical knowledge of investment returns $\{\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\ldots,\boldsymbol{\xi}_M\}$ to define a distributional uncertainty set for F. This is done using the set $\mathcal{D}_1(\mathcal{S},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}},\gamma_1,\gamma_2)$ where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are assigned as the empirical estimates of the mean $\hat{\boldsymbol{\mu}}=M^{-1}\sum_{i=1}^M\boldsymbol{\xi}_i$ and covariance matrix $\hat{\boldsymbol{\Sigma}}=M^{-1}\sum_{i=1}^M(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}})^\mathsf{T}$ of $\boldsymbol{\xi}$ respectively.[3] We consider two options for the choice of $\mathcal{S}$: either $\mathcal{S}=\mathbb{R}^n$, or is an "ellipsoidal" set $\mathcal{S}=\{\boldsymbol{\xi}|(\boldsymbol{\xi}-\boldsymbol{\xi}_0)^\mathsf{T}\boldsymbol{\Theta}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)\leq 1\}$, where $\boldsymbol{\Theta}$ has at least one strictly positive eigenvalue.

Building on the results presented in Section 2, one can make the following statement about the tractability of the DRPO model.

THEOREM 4. *Given that $u(\cdot)$ is piecewise linear concave and that $\mathcal{S}=\mathbb{R}^n$ or ellipsoidal, finding an optimal solution $\boldsymbol{x}\in\mathbb{R}^n$ to the DRPO model, Problem* (18)*, equipped with the set of distributions $\mathcal{D}_1(\mathcal{S},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}},\gamma_1,\gamma_2)$ can be done in $O(n^{6.5})$.*

Proof: We first reformulate Problem (18) as a minimization problem :

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} \left(\max_{F\,\in\,\mathcal{D}_1(\mathcal{S},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}},\gamma_1,\gamma_2)} \mathbb{E}_F[\max_k\ -a_k\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x}-b_k]\right) \ ,$$

where $\mathcal{X}=\{x\in\mathbb{R}^n|\boldsymbol{x}\geq 0,\,\sum_{i=1}^n x_i=1\}$. After confirming that $\mathcal{S}$, with $\boldsymbol{\Theta}\succeq 0$, satisfies the weaker version of Assumption 1 (see Remark 3) and that $h(\boldsymbol{x},\boldsymbol{\xi})=\max_k\ -a_k\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x}-b_k$ satisfies Assumption 2 and Assumption 4, a straightforward application of Proposition 2 shows that Problem (18) can be solved in polynomial time. In order to get a more precise computational bound, one needs to take a closer look at the dual formulation presented in Lemma 1 (*cf*., the details in the appendix), and exploit the special structure of the objective function $h(\boldsymbol{x},\boldsymbol{\xi})$ of Problem (18):

$$\begin{aligned}
&\underset{\boldsymbol{x},\boldsymbol{Q},\boldsymbol{q},r,\boldsymbol{P},\boldsymbol{p},s}{\text{minimize}} && \gamma_2(\hat{\boldsymbol{\Sigma}}\bullet\boldsymbol{Q})-\hat{\boldsymbol{\mu}}^\mathsf{T}\boldsymbol{Q}\hat{\boldsymbol{\mu}}+r+(\hat{\boldsymbol{\Sigma}}\bullet\boldsymbol{P})-2\hat{\boldsymbol{\mu}}^\mathsf{T}\boldsymbol{p}+\gamma_1 s && \text{(19a)}\\
&\text{subject to} && \begin{bmatrix}\boldsymbol{P} & \boldsymbol{p}\\ \boldsymbol{p}^\mathsf{T} & s\end{bmatrix}\succeq 0\ ,\ \boldsymbol{p}=-\boldsymbol{q}/2-\boldsymbol{Q}\hat{\boldsymbol{\mu}}\ ,\ \boldsymbol{Q}\succeq 0 && \text{(19b)}\\
& && \boldsymbol{\xi}^\mathsf{T}\boldsymbol{Q}\boldsymbol{\xi}+\boldsymbol{\xi}^\mathsf{T}\boldsymbol{q}+r\geq -a_k\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x}-b_k\ ,\ \forall\boldsymbol{\xi}\in\mathcal{S},k\in\{1,2,...,K\} && \text{(19c)}\\
& && \sum_{i=1}^n x_i=1\ ,\ \boldsymbol{x}\geq 0\ . && \text{(19d)}
\end{aligned}$$

Given that $\mathcal{S}=\mathbb{R}^n$, one can use Schur's complement to replace Constraint (19c) by an equivalent linear matrix inequality:

$$\begin{aligned}
&\underset{\boldsymbol{x},\boldsymbol{Q},\boldsymbol{q},r,\boldsymbol{P},\boldsymbol{p},s}{\text{minimize}} && \gamma_2(\hat{\boldsymbol{\Sigma}}\bullet\boldsymbol{Q})-\hat{\boldsymbol{\mu}}^\mathsf{T}\boldsymbol{Q}\hat{\boldsymbol{\mu}}+r+(\hat{\boldsymbol{\Sigma}}\bullet\boldsymbol{P})-2\hat{\boldsymbol{\mu}}^\mathsf{T}\boldsymbol{p}+\gamma_1 s\\
&\text{subject to} && \begin{bmatrix}\boldsymbol{P} & \boldsymbol{p}\\ \boldsymbol{p}^\mathsf{T} & s\end{bmatrix}\succeq 0\ ,\ \boldsymbol{p}=-\boldsymbol{q}/2-\boldsymbol{Q}\hat{\boldsymbol{\mu}}\\
& && \begin{bmatrix}\boldsymbol{Q} & \boldsymbol{q}/2+a_k\boldsymbol{x}/2\\ \boldsymbol{q}^\mathsf{T}/2+a_k\boldsymbol{x}^\mathsf{T}/2 & r+b_k\end{bmatrix}\succeq 0\ ,\ \forall k\\
& && \sum_{i=1}^n x_i=1\ ,\ \boldsymbol{x}\geq 0\ .
\end{aligned}$$

On the other hand, if $\mathcal{S}$ is ellipsoidal and Θ has at least one positive eigenvalue, then the S-Lemma (*cf*., Theorem 2.2 in Pólik and Terlaky (2007)) can be used for any given $k\in\{1,...,K\}$ to replace Constraint (19c), which enforces that

$$\nexists\,\boldsymbol{\xi}\in\mathbb{R}^n \text{ such that } \boldsymbol{\xi}^\mathsf{T}\boldsymbol{Q}\boldsymbol{\xi}+\boldsymbol{\xi}^\mathsf{T}\boldsymbol{q}+r<-a_k\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x}-b_k\ \bigwedge\ (\boldsymbol{\xi}-\boldsymbol{\xi}_0)^\mathsf{T}\boldsymbol{\Theta}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)\leq 1\ ,$$

with the equivalent constraint that

$$\exists\,\tau_k\geq 0 \text{ such that } \forall\,\boldsymbol{\xi}\in\mathbb{R}^n\ ,\ \boldsymbol{\xi}^\mathsf{T}\boldsymbol{Q}\boldsymbol{\xi}+\boldsymbol{\xi}^\mathsf{T}\boldsymbol{q}+r+a_k\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x}+b_k\geq -\tau_k\left((\boldsymbol{\xi}-\boldsymbol{\xi}_0)^\mathsf{T}\boldsymbol{\Theta}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)-1\right)\ .$$

The problem can therefore be reformulated as a semi-definite program:

$$\begin{aligned}\underset{\boldsymbol{x},\boldsymbol{Q},\boldsymbol{q},r,\boldsymbol{P},\boldsymbol{p},s,\tau}{\text{minimize}} \quad & \gamma_2(\hat{\boldsymbol{\Sigma}} \bullet \boldsymbol{Q}) - \hat{\boldsymbol{\mu}}^{\mathsf{T}}\boldsymbol{Q}\hat{\boldsymbol{\mu}} + r + (\hat{\boldsymbol{\Sigma}} \bullet \boldsymbol{P}) - 2\hat{\boldsymbol{\mu}}^{\mathsf{T}}\boldsymbol{p} + \gamma_1 s \\ \text{subject to} \quad & \begin{bmatrix} \boldsymbol{P} & \boldsymbol{p} \\ \boldsymbol{p}^{\mathsf{T}} & s \end{bmatrix} \succeq 0 \ , \ \ \boldsymbol{p} = -\boldsymbol{q}/2 - \boldsymbol{Q}\hat{\boldsymbol{\mu}} \ , \ \ \boldsymbol{Q} \succeq 0 \\ & \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{q}/2 + a_k\boldsymbol{x}/2 \\ \boldsymbol{q}^{\mathsf{T}}/2 + a_k\boldsymbol{x}^{\mathsf{T}}/2 & r + b_k \end{bmatrix} \succeq -\tau_k \begin{bmatrix} \boldsymbol{\Theta} & -\boldsymbol{\Theta}\boldsymbol{\xi}_0 \\ -\boldsymbol{\xi}_0^{\mathsf{T}}\boldsymbol{\Theta} & \boldsymbol{\xi}_0^{\mathsf{T}}\boldsymbol{\Theta}\boldsymbol{\xi}_0 - 1 \end{bmatrix} \ , \ \forall k \\ & \tau_k \geq 0 \ \ \forall k \\ & \sum_{i=1}^{n} x_i = 1 \ , \ \ \boldsymbol{x} \geq 0 \, ,\end{aligned}$$

where $\tau \in \mathbb{R}^K$ is a new vector of optimization variables.

In both cases, the optimization problem that needs to be solved is a semi-definite program. It is well known that an interior point algorithm can be used to solve an SDP of the form

$$\begin{aligned}\underset{\boldsymbol{x}\in\mathbb{R}^{\tilde{n}}}{\text{minimize}} \quad & \boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}_i(\boldsymbol{x}) \succeq 0 \ \ \forall i = 1,2,...,\tilde{K}\end{aligned}$$

in $O\left(\left(\sum_{i=1}^{\tilde{K}} \tilde{m}_i\right)^{0.5}\left(\tilde{n}^2 \sum_{i=1}^{\tilde{K}} \tilde{m}_i^2 + \tilde{n}\sum_{i=1}^{\tilde{K}} \tilde{m}_i^3\right)\right)$, where $\tilde{m}_i$ stands for the dimension of the positive semi-definite cone (*i.e.,* $\boldsymbol{A}_i(\boldsymbol{x}) \in \mathbb{R}^{\tilde{m}_i \times \tilde{m}_i}$) (see Nesterov and Nemirovski (1994)). In both SDPs that interest us, one can show that $\tilde{n} \leq n^2 + 4n + 2 + K$ and that both problems can be solved in $O(K^{3.5}n^{6.5})$ operations, with K being the number of pieces in the utility function $u(\cdot)$. We conclude that the portfolio optimization problem can be solved in $O(n^{6.5})$. □

The results presented in Theorem 4 are related to Bertsimas et al. (2000) where the authors proposed semi-definite programming models for solving moment problems that are similar to the one present in the objective of the DRPO. However, notice how the two SDP models involved in the proof of Theorem 4 actually address the harder task of finding an optimal robust decision and yet do not lead to in a heavier computational load. It is also the case that our proposed SDP models consider a more practical set of distributions which accounts for mean and covariance matrix uncertainty (in the form of a linear matrix inequality) and support information.

REMARK 5. The computational complexity presented here is based on general theory for solving semi-definite programs. Based on an implementation that uses SeDuMi (Sturm (1999)), we actually observed empirically that complexity grows in the order of $O(n^5)$ for dense problems. In practice, one may also be able to exploit the structure of problems where subsets (or linear combinations) of assets are known to behave independently from each other.

4.2. A Case of Worst Distribution with Largest Second Moment Matrix

When presenting our distributionally robust framework, we argued in Remark 1 that a positive semi-definite lower bound on the centered second moment matrix was uninteresting. Actually, in the case of a portfolio optimization problem with piecewise linear concave utility function, the argument can be made more formally. The proof of the following proposition also provides valuable insights on the structure of a worst case distribution for the distributionally robust portfolio optimization problem.

PROPOSITION 3. *The distributionally robust portfolio optimization problem with piecewise linear concave utility and no support constraint on the distribution is an instance of Problem* (7) *where Constraint* (4c) *of its inner moment problem is tight for a worst case distribution.*

Proof: Consider the inner problem in the robust portfolio optimization with no support constraint on the distribution:

$$\max_{F \,\in\, \mathcal{D}_1(\mathbb{R}^n,\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}},0,\gamma_2)} \quad \mathbb{E}_F[\max_k \; -a_k\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x} - b_k] \;\; . \tag{20}$$

For simplicity, we consider that there is no uncertainty in the mean of the distribution (*i.e.*, $\gamma_1 = 0$); thus, Constraint (4c) reduces to an upper bound on the covariance matrix of $\boldsymbol{\xi}$. The dual of this problem can be shown to reduce to:

$$\begin{array}{ll} \underset{\boldsymbol{Q},\boldsymbol{q},r}{\text{minimize}} & (\hat{\boldsymbol{\Sigma}}\bullet\boldsymbol{Q}) + \hat{\boldsymbol{\mu}}^\mathsf{T}\boldsymbol{Q}\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^\mathsf{T}\boldsymbol{q} + r \\ \text{subject to} & \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{q}/2 + a_k\boldsymbol{x}/2 \\ \boldsymbol{q}^\mathsf{T}/2 + a_k\boldsymbol{x}^\mathsf{T}/2 & r + b_k \end{bmatrix} \succeq 0 \;,\; \forall\, k \in \{1,2,...,K\} \;\; . \end{array}$$

Applying duality theory a second time leads to formulating a new equivalent version of the primal problem, which by strong duality achieves the same optimum:

$$\underset{\{(\boldsymbol{\Lambda}_k,\boldsymbol{\lambda}_k,\nu_k)\}_{k=1}^K}{\text{maximize}} \quad \sum_{k=1}^K a_k\boldsymbol{x}^\mathsf{T}\boldsymbol{\lambda}_k + \nu_k b_k \tag{21a}$$

$$\text{subject to} \quad \sum_{k=1}^K \boldsymbol{\Lambda}_k \preceq \gamma_2\hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\mathsf{T} \tag{21b}$$

$$\sum_{k=1}^K \boldsymbol{\lambda}_k = \hat{\boldsymbol{\mu}} \;\;,\;\; \sum_{k=1}^K \nu_k = 1 \tag{21c}$$

$$\begin{bmatrix} \boldsymbol{\Lambda}_k & \boldsymbol{\lambda}_k \\ \boldsymbol{\lambda}_k^\mathsf{T} & \nu_k \end{bmatrix} \succeq 0 \quad \forall\, k \in \{1,2,...,K\} \;\; . \tag{21d}$$

We can show that there always exists an optimal solution such that Constraint (21b) is satisfied with equality. Given an optimal assignment $X^* = \{(\boldsymbol{\Lambda}_k^*,\boldsymbol{\lambda}_k^*,\nu_k^*)\}_{k=1}^K$, consider an alternate solution $X' = \{(\boldsymbol{\Lambda}_k',\boldsymbol{\lambda}_k',\nu_k')\}_{k=1}^K$ which is exactly the same as the original solution X^* except for $\boldsymbol{\Lambda}_1' = \boldsymbol{\Lambda}_1^* + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = \gamma_2\hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\mathsf{T} - \sum_{k=1}^K \boldsymbol{\Lambda}_k^* \succeq 0$. Obviously the two solutions achieve the same objective values since $\{(\boldsymbol{\lambda}_k^*,\nu_k^*)\}_{k=1}^K$ and $\{(\boldsymbol{\lambda}_k',\nu_k)\}_{k=1}^K$ are the same. If we can show that X' is also feasible then it is necessarily optimal. The only feasibility constraint that needs to be verified is the following:

$$\begin{bmatrix} \boldsymbol{\Lambda}_1' & \boldsymbol{\lambda}_1' \\ \boldsymbol{\lambda}_1'^\mathsf{T} & \nu_1' \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_1^* & \boldsymbol{\lambda}_1^* \\ \boldsymbol{\lambda}_1^{*\mathsf{T}} & \nu_1^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Delta} & 0 \\ 0 & 0 \end{bmatrix} \succeq 0 \;\; ,$$

and is necessarily satisfied since by definition X^* is feasible and by construction $\boldsymbol{\Delta}$ is positive semi-definite. It is therefore the case that there exists a solution X^* that is optimal with respect to Problem (21) and satisfies Constraint (21b) with equality. Furthermore, one is assured that $\sum_{k=1}^K a_k\boldsymbol{x}^\mathsf{T}\boldsymbol{\lambda}_k^* + \nu_k^* b_k$ is equal to the optimal value of Problem (20).

After assuming without loss of generality that all $\nu_k^* > 0$, let us now construct K random vectors $\{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2,...,\boldsymbol{\zeta}_K\}$ that satisfy the following conditions:

$$\mathbb{E}[\boldsymbol{\zeta}_k] = \frac{1}{\nu_k^*}\boldsymbol{\lambda}_k^* \;\;,\;\; \mathbb{E}[\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^\mathsf{T}] = \frac{1}{\nu_k^*}\boldsymbol{\Lambda}_k^* \;\; .$$

Note that since X^* satisfies Constraint (21d), we are assured that

$$\mathbb{E}[\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^\mathsf{T}] - \mathbb{E}[\boldsymbol{\zeta}_k]\mathbb{E}[\boldsymbol{\zeta}_k]^\mathsf{T} = \begin{bmatrix} \boldsymbol{I} \\ -\mathbb{E}[\boldsymbol{\zeta}_k]^\mathsf{T} \end{bmatrix}^\mathsf{T} \begin{bmatrix} \mathbb{E}[\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^\mathsf{T}] & \mathbb{E}[\boldsymbol{\zeta}_k] \\ \mathbb{E}[\boldsymbol{\zeta}_k]^\mathsf{T} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{I} \\ -\mathbb{E}[\boldsymbol{\zeta}_k]^\mathsf{T} \end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} \boldsymbol{I} \\ -\mathbb{E}\,[\boldsymbol{\zeta}_k]^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \frac{1}{\nu_k}\boldsymbol{\Lambda}_k^* & \frac{1}{\nu_k}\boldsymbol{\lambda}_k^* \\ \frac{1}{\nu_k}{\boldsymbol{\lambda}_k^*}^{\mathsf{T}} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{I} \\ -\mathbb{E}\,[\boldsymbol{\zeta}_k]^{\mathsf{T}} \end{bmatrix} \\
&= \frac{1}{\nu_k^*} \begin{bmatrix} \boldsymbol{I} \\ -\mathbb{E}\,[\boldsymbol{\zeta}_k]^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \boldsymbol{\Lambda}_k^* & \boldsymbol{\lambda}_k^* \\ {\boldsymbol{\lambda}_k^*}^{\mathsf{T}} & \nu_k^* \end{bmatrix} \begin{bmatrix} \boldsymbol{I} \\ -\mathbb{E}\,[\boldsymbol{\zeta}_k]^{\mathsf{T}} \end{bmatrix} \succeq 0 \ .
\end{aligned}
$$

Hence, such a set of random vectors $\{\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, ..., \boldsymbol{\zeta}_K\}$ exists. For instance, if $\mathbb{E}\,[(\boldsymbol{\zeta}_k - \mathbb{E}\,[\boldsymbol{\zeta}_k])(\boldsymbol{\zeta}_k - \mathbb{E}\,[\boldsymbol{\zeta}_k])^{\mathsf{T}}] \succ 0$, then $\boldsymbol{\zeta}_k$ can take the form of a multivariate Gaussian distribution with such mean and covariance matrix. Otherwise, one could construct a lower dimensional random vector; in particular, if $\mathbb{E}\,[(\boldsymbol{\zeta}_k - \mathbb{E}\,[\boldsymbol{\zeta}_k])(\boldsymbol{\zeta}_k - \mathbb{E}\,[\boldsymbol{\zeta}_k])^{\mathsf{T}}] = 0$ then the random vector could have the Dirac measure $\delta_{\mathbb{E}\,[\boldsymbol{\zeta}_k]}$ as a distribution.

Let $\tilde{k}$ be a discrete random variable that follows a distribution with parameters $(\nu_1^*, \nu_2^*, ..., \nu_K^*)$, such that $\mathbb{P}(\tilde{k} = i) = \nu_i^*$, and use it to construct the random vector $\boldsymbol{\xi} = \boldsymbol{\zeta}_{\tilde{k}}$. Since X^* satisfies Constraint (21b) and Constraint (21c) tightly, one can show that the distribution function of $\boldsymbol{\xi}^*$ lies in $\mathcal{D}(\mathbb{R}^n, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, 0, \gamma_2)$ and has largest covariance.

$$
\begin{aligned}
\mathbb{E}\,[\boldsymbol{\xi}^*] &= \sum_{k=1}^{K} \mathbb{E}\,[\boldsymbol{\zeta}_k | \tilde{k} = k]\mathbb{P}(\tilde{k} = l) = \sum_{k=1}^{K} \frac{1}{\nu_k^*}\boldsymbol{\lambda}_k^* \nu_k^* = \hat{\boldsymbol{\mu}} \\
\mathbb{E}\,[\boldsymbol{\xi}^* {\boldsymbol{\xi}^*}^{\mathsf{T}}] &= \sum_{k=1}^{K} \mathbb{E}\,[\boldsymbol{\zeta}_k \boldsymbol{\zeta}_k^{\mathsf{T}} | \tilde{k} = k]\mathbb{P}(\tilde{k} = l) = \sum_{k=1}^{K} \frac{1}{\nu_k^*}\boldsymbol{\Lambda}_k^* \nu_k^* = \gamma_2 \hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathsf{T}}
\end{aligned}
$$

Moreover, when used as a candidate distribution in Problem (20) it actually achieves the maximum since we can show that it must be greater or equal to it.

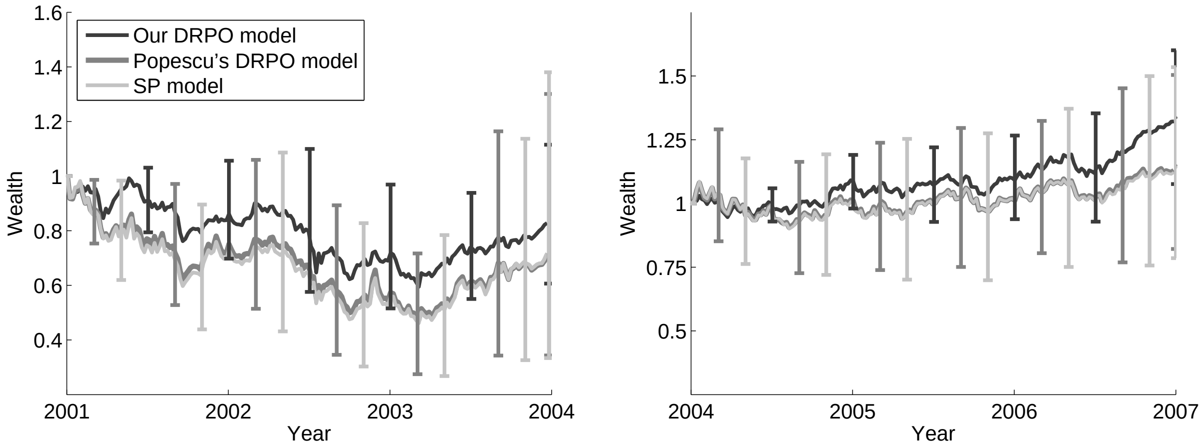
$$
\begin{aligned}
\mathbb{E}\left[\max_l \ -a_l \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\xi}^* - b_l\right] &= \sum_{k=1}^{K} \mathbb{E}\left[\max_l \ -a_l \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\zeta}_{\tilde{k}} - b_l \middle| \tilde{k} = k\right] \mathbb{P}(\tilde{k} = k) \\
&\geq \sum_{k=1}^{K} \mathbb{E}\,[-a_k \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\zeta}_k - b_k]\mathbb{P}(\tilde{k} = k) \\
&= \sum_{k=1}^{K} -a_k \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\lambda}_k^* - b_k \nu_k^* \\
&= \max_{F \,\in\, \mathcal{D}_1(\mathbb{R}^m, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, 0, \gamma_2)} \mathbb{E}_F[\max_k \ -a_k \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\xi} - b_k]
\end{aligned}
$$

We conclude that we just constructed a worst case distribution that does have the largest covariance. □

An interesting consequence of Proposition 3 is that in the framework considered in Popescu (2007), if the utility function is piecewise linear concave, one can find the optimal portfolio in polynomial time using our semi-definite programming formulation with the distributional set $\mathcal{D}_1(\mathbb{R}^n, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, 0, 1)$. Theoretically, our semi-definite program formulation is more tractable than the method proposed in Popescu (2007). However, it is true that our framework does not provide a polynomial time algorithm for the larger range of utility functions considered in Popescu (2007).

REMARK 6. Since the submission of this article, we became aware of independent work presented in Natarajan et al. (2008), which also addresses the computational difficulties related to the method proposed by Popescu. Their work is closely related to our result. Actually, for the case of unbounded support, their derivations lead to a further reduction of the DRPO model with known moments to the form of a second-order cone program. On the other hand, they do not consider support constraints and do not study the effect of moment uncertainty on the performance of a portfolio. Their approach is therefore susceptible, in practice, to the same deficiencies as Popescu's method when the moments are estimated using historical data.

Figure 1 Comparison of wealth evolution in 300 experiments conducted over the years 2001-2007.



Note. For each approach, the figures indicate periodically the 10th and 90th percentile of the distribution of accumulated wealth.

4.3. Experiments with Historical Stock Data

We evaluate our portfolio optimization framework using an historical data set of 30 assets over a horizon of 15 years (1992-2007), obtained from the Yahoo! Finance web site.[4] Each experiment consists of randomly choosing 4 assets, and building a dynamic portfolio with these assets through the years 2001-2007. At any given day of the experiment, the algorithms are allowed to use a period of 30 days from the most recent history to assign the portfolio. All methods assume that in this period the samples are independent and identically distributed. Note that 30 samples of data might be insufficient to generate good empirical estimates of the mean and covariance matrix of returns; however, using a larger history would make the assumption of independent and identical samples somewhat unrealistic.

In implementing our method, referred as the DRPO model, the distributional set is formulated as $\mathcal{D}_1(\mathbb{R}^4, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, 1.35, 8.32)$, where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the empirical estimates of the mean and covariance matrix of $\boldsymbol{\xi}$ respectively. Due to the sample size being too small to set the parameters $\bar{\gamma}_1$ and $\bar{\gamma}_2$ as shown in Definition 2, instead these parameters are chosen based on a simple statistical analysis of the amount of noise present in the estimation of mean and covariance matrix during the years 1992-2001.[5] We compare our approach to the one proposed by Popescu (2007), where the mean and covariance matrix of the distribution F is assumed to be equal to the empirical estimates measured on the last 30 days. The method is also compared to using a naive approximation of the stochastic program, referred as the SP model, which maximizes the average utility over the last 30 days. We believe that the statistics obtained over the set of 300 experiments and presented in Table 1 demonstrate how much there is to gain in terms of average performance and risk reduction by considering an optimization model that accounts for both distribution and moment uncertainty.

Table 1 Comparison of short and long term performance over six years of trading.

Method	Single Day Utility (2001-2007)		Yearly Return (2001-2004)		Yearly Return (2004-2007)	
	Avg.	1st perc.	Avg.	10th perc.	Avg.	10th perc.
Our DRPO model	1.000	0.983	0.944	0.846	1.102	1.025
Popescu's DRPO model	1.000	0.975	0.700	0.334	1.047	0.936
SP model	1.000	0.973	0.908	0.694	1.045	0.923

First, from the analysis of the daily returns generated by each method, one observes that they achieve comparable average daily utility. However, our DRPO model stands out as being more reliable. For instance, the lower 1st percentile of the utility distribution is 0.8% higher then the two competing methods. Also, this difference in reliability becomes more obvious when considering the respective long term performances. Figure 1 presents the average evolution of wealth on a six-year period when managing a portfolio of 4 assets on a daily basis with any of the three methods. In Table 1, the performances over the years 2001-2004 are presented separately from the performances over the years 2004-2007 in order to measure how they are affected by different levels of economic growth. The figures also periodically indicate the 10th and 90th percentile of the wealth distribution over the set of 300 experiments. The statistics of the long term experiments demonstrate empirically that our method significantly outperforms the two other ones in terms of average return and risk during both the years of economic growth and the years of decline. More specifically, our DRPO model outperformed Popescu's DRPO model in terms of total return accumulated over the period 2001-2007 in 79.2% of our experiments. Also, it performed on average at least 1.67 times better than any competing model. Note that these experiments are purely illustrative of the strengths and weaknesses of the different models. For instance, the returns measured in each experiment do not take into account transaction fees. The realized returns are also biased due to the fact that the assets involved in our experiments were known to be major assets in their category in January 2007. On the other hand, the realized returns were also negatively biased due to the fact that in each experiment the models were managing a portfolio of only four assets. Overall, we believe that these biases affected all methods equally.

Appendix. Proof of Lemma 1

We first establish the primal-dual relationship between Problem (4) and Problem (5). In a second step, we demonstrate that the conditions for strong duality to hold are met.

STEP 1. One can first show, by formulating the Lagrangian of Problem (3), that the dual can take the following form

$$\begin{aligned} \underset{r,\boldsymbol{Q},\boldsymbol{P},\boldsymbol{p},s}{\text{minimize}} \quad & (\gamma_2\boldsymbol{\Sigma}_0 - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\mathsf{T})\bullet\boldsymbol{Q} + r + (\boldsymbol{\Sigma}_0\bullet\boldsymbol{P}) - 2\boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{p} + \gamma_1 s && \text{(22a)}\\ \text{subject to} \quad & \boldsymbol{\xi}^\mathsf{T}\boldsymbol{Q}\boldsymbol{\xi} - 2\boldsymbol{\xi}^\mathsf{T}(\boldsymbol{p}+\boldsymbol{Q}\boldsymbol{\mu}_0) + r - h(\boldsymbol{x},\boldsymbol{\xi}) \ge 0 \ , \ \forall\, \boldsymbol{\xi}\in\mathcal{S} && \text{(22b)}\\ & \boldsymbol{Q}\succeq 0 && \text{(22c)}\\ & \begin{bmatrix} \boldsymbol{P} & \boldsymbol{p} \\ \boldsymbol{p}^\mathsf{T} & s \end{bmatrix} \succeq 0 \ , && \text{(22d)} \end{aligned}$$

where $r\in\mathbb{R}$ and $\boldsymbol{Q}\in\mathbb{R}^{m\times m}$ are the dual variables for Constraint (4b) and Constraint (4c) respectively, while $\boldsymbol{P}\in\mathbb{R}^{m\times m}$, $\boldsymbol{p}\in\mathbb{R}^m$ and $s\in\mathbb{R}$ form together a matrix which is the dual variable associated with Constraint (4d).

We can further simplify this dual problem by solving analytically for the variables $(\boldsymbol{P},\boldsymbol{p},s)$, while keeping $(\boldsymbol{Q},r)$ fixed. Because of Constraint (22d), we can consider two cases for the variable s^*: either $s^*=0$ or $s^*>0$. Assuming that $s^*=0$, then it must be that $\boldsymbol{p}^*=0$ otherwise ${\boldsymbol{p}^*}^\mathsf{T}\boldsymbol{p}^*>0$ and

$$\begin{bmatrix} \boldsymbol{p}^* \\ y \end{bmatrix}^\mathsf{T} \begin{bmatrix} \boldsymbol{P}^* & \boldsymbol{p}^* \\ {\boldsymbol{p}^*}^\mathsf{T} & s^* \end{bmatrix} \begin{bmatrix} \boldsymbol{p}^* \\ y \end{bmatrix} = {\boldsymbol{p}^*}^\mathsf{T}\boldsymbol{P}^*\boldsymbol{p}^* - 2{\boldsymbol{p}^*}^\mathsf{T}\boldsymbol{p}^* y < 0 \ , \text{ for } y > \frac{{\boldsymbol{p}^*}^\mathsf{T}\boldsymbol{P}^*\boldsymbol{p}^*}{2{\boldsymbol{p}^*}^\mathsf{T}\boldsymbol{p}^*} \ ,$$

which contradicts Constraint (22d). Similarly, $\boldsymbol{P}^*=0$ is an optimal solution since it minimizes the objective. We conclude that if $s^*=0$ then, after replacing $\boldsymbol{q}=-2(\boldsymbol{p}+\boldsymbol{Q}\boldsymbol{\mu}_0)$, Problem (22)'s objective does indeed reduce to

$$\gamma_2(\boldsymbol{\Sigma}_0\bullet\boldsymbol{Q}) - \boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{Q}\boldsymbol{\mu}_0 + r \ = \ r + \gamma_2(\boldsymbol{\Sigma}_0\bullet\boldsymbol{Q}) + \boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{Q}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{q} + \sqrt{\gamma_1}\|\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{q}+2\boldsymbol{Q}\boldsymbol{\mu}_0)\| \ .$$

If instead one assumes that $s^*>0$, then using Schur's complement, Constraint (22d) can be shown equivalent to $\boldsymbol{P}\succeq\frac{1}{s}\boldsymbol{p}\boldsymbol{p}^\mathsf{T}$. Since $\boldsymbol{\Sigma}_0\succeq 0$, $\boldsymbol{P}^*=\frac{1}{s}\boldsymbol{p}\boldsymbol{p}^\mathsf{T}$ is a valid optimal solution and can be replaced in the objective. It remains to solve for $s^*>0$, which reduces to solving the one dimensional convex optimization problem $\text{minimize}_{s>0}\,\frac{1}{s}\boldsymbol{p}^\mathsf{T}\boldsymbol{\Sigma}_0\boldsymbol{p}+\gamma_1 s$. By setting the derivative of the objective function to zero, we obtain that $s^* = \sqrt{\frac{1}{\gamma_1}\boldsymbol{p}^\mathsf{T}\boldsymbol{\Sigma}_0\boldsymbol{p}}$. Thus, once again, after replacing $\boldsymbol{q} = -2(\boldsymbol{p} + \boldsymbol{Q}\boldsymbol{\mu}_0)$, the optimal value of Problem (22) reduces to the form of Problem (5):

$$r + \gamma_2(\boldsymbol{\Sigma}_0 \bullet \boldsymbol{Q}) + \boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{Q}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{q} + \sqrt{\gamma_1}\|\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{q} + 2\boldsymbol{Q}\boldsymbol{\mu}_0)\| \ .$$

STEP 2. One can easily show that the conditions on γ_1, γ_2 and $\boldsymbol{\Sigma}_0$ are sufficient to ensure that the Dirac measure $\delta_{\boldsymbol{\mu}_0}$ (see Endnote 1 for definition) lies in the relative interior of the feasible set of Problem (3). Based on the weaker version of Proposition 3.4 in Shapiro (2001), we can conclude that there is no duality gap between the two problems and that if $\Psi(\boldsymbol{x};\gamma_1,\gamma_2)$ is finite then the set of optimal solutions to Problem (5) must be non-empty. □

Endnotes

1. Recall that the Dirac measure $\delta_{\boldsymbol{a}}$ is the measure of mass one at the point $\boldsymbol{a}$.
2. Note that if $\boldsymbol{\xi}$'s support set is unbounded, one can also derive bounds of similar nature either by considering that $\boldsymbol{\zeta}$ has bounded support with high probability, or by making use of partial knowledge of higher moments of the distribution. This last fact was recently confirmed in So (2008).
3. One should also verify that $\hat{\boldsymbol{\Sigma}} \succ 0$.
4. The list of assets that is used in our experiments was inspired by Goldfarb and Iyengar (2003). More specifically, the 30 assets are: AAR Corp., Boeing Corp., Lockheed Martin, United Technologies, Intel Corp., Hitachi, Texas Instruments, Dell Computer Corp., Palm Inc., Hewlett Packard, IBM Corp., Sun Microsystems, Bristol-Myers-Squibb, Applera Corp.-Celera Group, Eli Lilly and Co., Merck and Co., Avery Denison Corp., Du Pont, Dow Chemical, Eastman Chemical Co., AT&T, Nokia, Motorola, Ariba, Commerce One Inc., Microsoft, Oracle, Akamai, Cisco Systems, Northern Telecom, Duke Energy Company, Exelon Corp., Pinnacle West, FMC Corp., General Electric, Honeywell, Ingersoll Rand.
5. More specifically, given that one chooses 4 stocks randomly and selects a period of 60 days between 1992 and 2001 randomly, the values for γ_1 and γ_2 are chosen such that when using the first 30 days of the period to center the set $\mathcal{D}(\gamma_1,\gamma_2)$, the distributional set contains, with 99% probability, distributions with moments equal to the moments estimated from the last 30 days of the period.

Acknowledgments

The authors acknowledge the Fonds Québécois de la Recherche sur la Nature et les Technologies and The Boeing Company for their financial support. They wish to also thank Amir Dembo, Anthony Man-Cho So, Alexander Shapiro, Melvyn Sim and Benjamin Armbruster for valuable discussions.

References

Anderson, T. W. 1984. *An Introduction to Multivariate Analysis*. John Wiley & Sons.

Ben-Tal, A., A. Nemirovski. 1998. Robust convex optimization. *Math. Oper. Res.* **23**(4) 769–805.

Bertsimas, D., D. B. Brown, C. Caramanis. 2008. Theory and applications of robust optimization. Working paper.

Bertsimas, D., I. Popescu. 2005. Optimal inequalities in probability theory: A convex optimization approach. *SIAM J. Optim.* **15**(3) 780–804.

Bertsimas, D., I. Popescu, J. Sethuraman. 2000. Moment problems and semidefinite programming. H. Wolkowicz, R. Saigal, L. Vandenberghe, eds., *Handbook of Semidefinite Programming*. Kluwer Academic Publishers, 469–510.

Bertsimas, D., S. Vempala. 2004. Solving convex programs by random walks. *J. ACM* **51** 540–556.

Birge, J. R., R. J.-B. Wets. 1987. Computing bounds for stochastic programming problems by means of a generalized moment problem. *Math. Oper. Res.* **12**(1) 149–162.

Calafiore, G., M. C. Campi. 2005. Uncertain convex programs: Randomized solutions and confidence levels. *Math. Programming* **102** 25–46.

Calafiore, G., L. El Ghaoui. 2006. On distributionally robust chance-constrained linear programs. *J. Optim. Theory Appl.* **130**(1) 1–22.

de Farias, D. P., B. Van Roy. 2001. On constraint sampling for the linear programming approach to approximate dynamic programming. *Math. Oper. Res.* **29** 2004.

Dupacová, J. 1987. The minimax approach to stochastic programming and an illustrative application. *Stochastics* **20** 73–88.

Dupacová, J. 2001. Stochastic programming: Minimax approach. *Encyclopedia of Optimization* **5** 327–330.

Edelman, A. 1989. Eigenvalues and condition numbers of random matrices. Ph.D. thesis, MIT.

Ermoliev, Y., A. Gaivoronski, C. Nedeva. 1985. Stochastic optimization problems with partially known distribution functions. *Journal on Control and Optimization* **23** 696–716.

Fujikoshi, Y. 1980. Asymptotic expansions for the distributions of the sample roots under nonnormality. *Biometrika* **67**(1) 45–51.

Gaivoronski, A. A. 1991. A numerical method for solving stochastic programming problems with moment constraints on a distribution function. *Ann. Oper. Res.* **31** 347–370.

Goffin, J. L., J. P. Vial. 1993. On the computation of weighted analytic centers and dual ellipsoids with the projective algorithm. *Math. Programming* **60**(1) 81–92.

Goldfarb, D., G. Iyengar. 2003. Robust portfolio selection problems. *Math. Oper. Res.* **28**(1) 1–38.

Grötschel, M., L. Lovász, A. Schrijver. 1981. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica* **1** 169–197.

Isii, K. 1963. On the sharpness of Chebyshev-type inequalities. *Annals of the Institute of Statistical Mathematics* **14** 185–197.

Kall, P. 1988. Stochastic programming with recourse: Upper bounds and moment problems A review. *Advances in Mathematical Optimization*. 86–103.

Lagoa, C. M., B. R. Barmish. 2002. Distributionally robust Monte Carlo simulation: A tutorial survey. *Proceedings of the International Federation of Automatic Control World Congress*. 1–12.

Landau, H. J. 1987. *Moments in Mathematics: Lecture Notes Prepared for the AMS Short Course*. American Mathematical Society.

Marshall, A., I. Olkin. 1960. Multivariate Chebyshev inequalities. *Annals of Mathematical Statistics* **31** 1001–1024.

McDiarmid, C. 1998. Concentration. M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, B. Reed, eds., *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer, 195–248.

Natarajan, K., M. Sim, J. Uichanco. 2008. Tractable robust expected utility and risk models for portfolio optimization. Accepted in *Mathematical Finance* .

Nesterov, Y., A. Nemirovski. 1994. *Interior-point polynomial methods in convex programming*, vol. 13. Studies in Applied Mathematics.

Pólik, I., T. Terlaky. 2007. A survey of the S-Lemma. *SIREV* **49**(3) 371–418.

Popescu, I. 2007. Robust mean-covariance solutions for stochastic optimization. *Oper. Res.* **55**(1) 98–112.

Prékopa, A. 1995. *Stochastic Programming*. Kluwer Academic Publishers.

Rockafellar, R. T., S. Uryasev. 2000. Optimization of conditional value-at-risk. *Journal of Risk* **2**(3) 21–41.

Rockafeller, R. T. 1970. *Convex Analysis*. Princeton University Press.

Rockafeller, R. T. 1974. *Conjugate Duality and Optimization*, Regional Conference Series in Applied Mathematics, vol. 16. SIAM.

Scarf, H. 1958. A min-max solution of an inventory problem. *Studies in The Mathematical Theory of Inventory and Production* 201–209.

Shapiro, A. 2001. On duality theory of conic linear problems. M. A. Goberna, M. A. López, eds., *Semi-Infinite Programming: Recent Advances*. Kluwer Academic Publishers, 135–165.

Shapiro, A. 2006. Worst-case distribution analysis of stochastic programs. *Math. Programming* **107**(1) 91–96.

Shapiro, A., T. Homem-de-Mello. 2000. On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. *SIAM J. Optim.* **11**(1) 70–86.

Shapiro, A., A. J. Kleywegt. 2002. Minimax analysis of stochastic problems. *Optimization Methods and Software* **17** 523–542.

Shawe-Taylor, J., N. Cristianini. 2003. Estimating the moments of a random vector with applications. J. Siemons, ed., *Proceedings of GRETSI 2003 Conference*. Cambridge University Press, 47–52.

So, A. M.-C. 2008. Private communication.

Sturm, J. F. 1999. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software* **11–12** 625–653.

Čerbáková, J. 2005. Worst-case VaR and CVaR. H. Haasis, H. Kopfer, J. Schnberger, eds., *Operations Research Proceedings*. Springer, 817–822.

Waternaux, C. 1976. Asymptotic distribution of the sample roots for the nonnormal population. *Biometrika* **63**(3) 639–645.

Ye, Y. 1997. Complexity analysis of the analytic center cutting plane method that uses multiple cuts. *Math. Programming* **78**(1) 85–104.

Yue, J., B. Chen, M.-C. Wang. 2006. Expected value of distribution information for the newsvendor problem. *Oper. Res.* **54**(6) 1128–1136.

Zhu, S. S., M. Fukushima. 2005. Worst-case conditional value-at-risk with application to robust portfolio management. Tech. rep., Kyoto University.

Zhu, Z., J. Zhang, Y. Ye. 2006. Newsvendor optimization with limited distribution information. Tech. rep., Stanford University.