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

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

Erick Delage
Department of Electrical Engineering, Stanford University, Stanford, California, USA
edelage@stanford.edu, http://www.stanford.edu/∼edelage

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 return of assets.

Subject classifications: Programming: stochastic, Statistics: estimation, Finance: portfolio.
Area of review: Optimization.
History: Draft created February 20, 2008. First Revision submitted September, 2008

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

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \;\; h(\mathbf{x},\xi) \;\; ,$$

where $\mathcal{X}$ is a convex set of feasible solutions and $h(\mathbf{x},\xi)$ is a convex cost function in $\mathbf{x}$ that depends on some parameters ξ, it is often the case that at the time of optimizing, the parameters have not yet been fully resolved. For examples, an investment manager cannot know the exact return for any available securities; or in a different context, a manufacturing producer cannot know the exact size of future demand.

If one chooses to represent his uncertainty about ξ through a distribution f_ξ, one can instead resort to minimizing the expected cost. This leads to solving a stochastic program:

$$\text{(SP)} \qquad \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \;\; \mathbb{E}_\xi[h(\mathbf{x},\xi)] \;\; ,$$

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

In an effort to address these issues, a robust formulation for stochastic programming was proposed in Scarf (1958). In this model, after defining a set $\mathcal{D}$ of possible probability distributions that is assumed to include the true f_ξ, 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:

$$\text{(DRSP)} \qquad \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \quad \left(\max_{f_\xi \in \mathcal{D}} \mathbb{E}_\xi[h(\mathbf{x},\xi)] \right) \quad .$$

Since its introduction, this model has gain 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)), computing upper bounds on the optimal value of a stochastic program (e.g. in Birge and Wets (1987) and in Kall (1988)), or providing robust decisions in contexts where 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 chooses the distributional set to be one that contains distributions that put all of their weight at a single point anywhere in the parameter support set $\mathcal{S}$, then the DRSP reduces to a so-called robust optimization problem with respect to the worst realization of ξ in $\mathcal{S}$ (*e.g.*, in Ben-Tal and Nemirovski (1998) and in Bertsimas et al. (2007)). Otherwise, in Lagoa and Barmish (2002) and in Shapiro (2006), the authors consider a set that contains unimodal distributions that satisfy some given support constraints; under some conditions on $h(x,\xi)$, they characterize the worst distribution as being a 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 and that they lead to linear equality or inequality constraints. For example, in his original model, Scarf considered a one dimensional decision variable x representing how much inventory one should hold, and a single parameter ξ representing a random demand with known mean and variance. The return function had the form $h(x,\xi) = -\min\{r\xi - cx\ ,\ r\xi - cx\}$. To solve this model, Scarf exploited the fact that the worst case distribution of demand could be chosen to be one with all 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 functions take more interesting forms, they all assume known first and second moments of the stochastic demand and their solution methods all rely on characterizing the worst case distribution as a point distribution.

The computational difficulties related to dealing with ξ of a larger dimension and with richer objective functions have limited the practical application of the DRSP model. More specifically, although in some cases the worst case moment expression can be simplified analytically, like in the linear chance constraint problem considered in Calafiore and El Ghaoui (2006), it is typical that the model becomes intractable (or NP-Hard to solve) and that only global optimization methods can be employed 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. For instance, in many data-driven problems, one must estimate these moments based on limited historical data assumed to be generated from f_ξ. 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 knowledge about the distribution's support and a *confidence region* for its mean and its covariance 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 the current exact-moment approaches (see Example 1 of Section 2.3 for more details). As a second contribution, in Section 3, after deriving a new confidence region for covariance matrices, we show how our proposed distributional set is well justified for addressing data-driven problems (*i.e.*, problems where the knowledge of ξ 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, our model performs better in practice on the distribution that drives the daily return of popular stocks compared to other DRSP formulations.

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_ξ 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 μ_0 and covariance matrix Σ_0 of the distribution: *e.g.*, empirical estimates. However we believe that in such problems, it is also rarely the case that one is entirely confident in these moments. For this reason, we propose representing this uncertainty with the following set parameterized by $\gamma_1 \geq 0$ and $\gamma_2 \geq 1$:

$$(\mathbb{E}[\xi] - \mu_0)^\mathsf{T} \Sigma_0^{-1} (\mathbb{E}[\xi] - \mu_0) \leq \gamma_1 \tag{1a}$$
$$\mathbb{E}[(\xi - \mu_0)(\xi - \mu_0)^\mathsf{T}] \preceq \gamma_2 \Sigma_0 \ . \tag{1b}$$

While Constraint (1a) assumes that the mean of ξ lies in an ellipsoid of size γ_1 centered at the estimate μ_0, Constraint (1b) forces the covariance matrix $\mathbb{E}[(\xi - \mu_0)(\xi - \mu_0)^\mathsf{T}]$ to lie in a positive semi-definite cone bounded by a matrix inequality. In other words, it describes how likely ξ is close to μ_0 in terms of the correlations expressed in Σ_0. Note that the parameters γ_1 and γ_2 provide natural means of quantifying the size of one's confidence in μ_0 and Σ_0 respectively.

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

$$\mathcal{D}_1(\mathcal{S}, \mu_0, \Sigma_0, \gamma_1, \gamma_2) = \left\{ f_\xi \in \mathcal{M} \;\middle|\; \begin{array}{l} \mathbb{P}(\xi \in \mathcal{S}) = 1 \\ (\mathbb{E}[\xi] - \mu_0)^\mathsf{T} \Sigma_0^{-1} (\mathbb{E}[\xi] - \mu_0) \leq \gamma_1 \\ \mathbb{E}[(\xi - \mu_0)(\xi - \mu_0)^\mathsf{T}] \preceq \gamma_2 \Sigma_0 \end{array} \right\} \ ,$$

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_ξ. The set $\mathcal{D}_1(\mathcal{S}, \mu_0, \Sigma_0, \gamma_1, \gamma_2)$, 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, $\mathcal{D}_1(\mathcal{S}, \mu_0, \mathbf{I}, 0, \infty)$ imposes exact mean and support constraints as is studied in Dupacová (1987) and in Bertsimas and Popescu (2005). Similarly, $\mathcal{D}_1(\mathbb{R}^m, \mu_0, \Sigma_0, 0, 1)$ 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 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 $\mathcal{D}_1(\mathcal{S}, \mu_0, \Sigma_0, \gamma_1, \gamma_2)$ which constrains all three types of statistics: support, mean and covariance matrix.

REMARK 1. While our proposed uncertainty model cannot be used to express arbitrarily large confidence in the second order statistics of ξ, in Section 3 we will show how in practice there are natural ways of assigning μ_0, Σ_0, γ_1 and γ_2 based on historical data. Of course, in some situations it might be interesting to add the following constraint:

$$\gamma_3 \Sigma_0 \preceq \mathbb{E}[(\xi - \mu_0)(\xi - \mu_0)^\mathsf{T}] \ , \tag{2}$$

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) unnecessary. For example, an instance of the portfolio optimization problem presented in Section 4 will have this characteristic.

2.1. The Inner Moment Problem with Moment Uncertainty

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

DEFINITION 1. Given a fixed x, let $\Psi(\mathbf{x};\gamma_1,\gamma_2)$ be the optimal value of the moment problem:

$$\underset{f_\xi \in \mathcal{D}_1}{\text{maximize}}\, \mathbb{E}_\xi[h(\mathbf{x},\xi)] \quad . \tag{3}$$

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

$$\begin{aligned}
\underset{f_\xi}{\text{maximize}} \quad & \int_{\mathcal{S}} h(\mathbf{x},\xi) df_\xi(\xi) && \text{(4a)}\\
\text{subject to} \quad & \int_{\mathcal{S}} df_\xi(\xi) = 1 && \text{(4b)}\\
& \int_{\mathcal{S}} (\xi-\mu_0)(\xi-\mu_0)^\mathsf{T} df_\xi(\xi) \preceq \gamma_2\Sigma_0 && \text{(4c)}\\
& \int_{\mathcal{S}} \begin{bmatrix} \Sigma_0 & (\xi-\mu_0) \\ (\xi-\mu_0)^\mathsf{T} & \gamma_1 \end{bmatrix} df_\xi(\xi) \succeq 0 && \text{(4d)}\\
& f_\xi \in \mathcal{M} \quad . && \text{(4e)}
\end{aligned}$$

As it is often done with moment problems, since we are strictly interested in the optimal value of this problem, we can shortcut the difficulties in solving 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 general moment problems).

LEMMA 1. *For a fixed $\mathbf{x}\in\mathbb{R}^n$, given that $\gamma_1 \geq 0$, $\gamma_2 \geq 1$, $\Sigma_0 \succ 0$, and $h(\mathbf{x},\xi)$ being f_ξ-integrable for all $f_\xi \in \mathcal{D}_1$, $\Psi(\mathbf{x};\gamma_1,\gamma_2)$ is finite and equal to the value of the following dual problem:*

$$\begin{aligned}
\underset{\mathbf{Q},\mathbf{q},r,t}{\text{minimize}} \quad & r+t && \text{(5a)}\\
\text{subject to} \quad & r \geq h(\mathbf{x},\xi) - \xi^\mathsf{T}\mathbf{Q}\xi - \xi^\mathsf{T}\mathbf{q} \;\; \forall\, \xi\in\mathcal{S} && \text{(5b)}\\
& t \geq \left(\gamma_2\Sigma_0 + \mu_0\mu_0^\mathsf{T}\right)\bullet\mathbf{Q} + \mu_0^\mathsf{T}\mathbf{q} + \sqrt{\gamma_1}\,\|\Sigma_0^{1/2}(\mathbf{q}+2\mathbf{Q}\mu_0)\| && \text{(5c)}\\
& \mathbf{Q}\succeq 0 \;\; , && \text{(5d)}
\end{aligned}$$

where $(A\bullet B)$ refers to the Frobenius inner product between matrices, $\mathbf{Q}\in\mathbb{R}^{m\times m}$ is a symmetric matrix, the vector $\mathbf{q}\in\mathbb{R}^m$ and $r,t\in\mathbb{R}$. In addition, there exists a set of real-valued assignments for $(\mathbf{Q},\mathbf{q},r,t)$ that achieves optimality for Problem (5)*.*

We defer the proof of this Lemma to the appendix since it results from the application of 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 convex set separation.

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

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

with linear objective and convex feasible set $\mathcal{Z}$. Given that the 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 the size of the problem by using the ellipsoid method if and only if $\mathcal{Z}$ satisfies the following two conditions :

1. *for any $\bar{z}$, $\bar{z}\in\mathcal{Z}$ can be verified in time polynomial in the dimension of z;*
2. *for any infeasible $\bar{z}$, a hyperplane that separates $\bar{z}$ from the feasible set $\mathcal{Z}$ can be generated in time polynomial in the dimension of 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 $\xi \in \mathbb{R}^m$ either confirm that $\xi \in \mathcal{S}$ or provide a hyperplane that separates ξ from $\mathcal{S}$ in time polynomial in m.*

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

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

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

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

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

where $\nabla_\xi h(\mathbf{x}, \xi)$ is a super-gradient of $h(\mathbf{x}, \cdot)$. It follows from the application of 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 Problem (4) and Problem (5) under a more general form of $h(\mathbf{x}, \xi)$.

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

1. *evaluate the value of $h_k(\mathbf{x}, \xi)$*
2. *find a super-gradient of $h_k(\mathbf{x}, \xi)$ in ξ.*

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

PROPOSITION 1. *Given that $\mathcal{S}$ satisfies Assumption 1 and that $h(\mathbf{x}, \xi)$ satisfies Assumption 2, Problem* (5) *is a convex optimization problem and can be solved to any ϵ precision in time polynomial in $\log(1/\epsilon)$ and the size of the problem.*

Proof: First, the feasible set of the problem is convex in $(\mathbf{Q}, \mathbf{q}, r, t)$ since $(\gamma_2\Sigma_0 + \mu_0\mu_0^\mathsf{T}) \bullet \mathbf{Q} + \mu_0^\mathsf{T}\mathbf{q} + \sqrt{\gamma_1}\ \|\Sigma_0^{1/2}(\mathbf{q} + 2\mathbf{Q}\mu_0)\|$ is a convex function in $(\mathbf{Q}, \mathbf{q})$. By Lemma 1, we are assured that the optimal solution set of Problem (5) is non-empty. In order to apply Lemma 2, we verify the two conditions for each constraint of Problem (5). 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. Otherwise, based on an infeasible assignment $(\bar{\mathbf{Q}}, \bar{\mathbf{q}}, \bar{r}, \bar{t})$, a separating hyperplane can be constructed in polynomial time:

$$\left(\gamma_2\Sigma_0 + \mu_0\mu_0^\mathsf{T} + \nabla_\mathbf{Q} g(\bar{\mathbf{Q}}, \bar{\mathbf{q}})\right) \bullet \mathbf{Q} + (\mu_0 + \nabla_\mathbf{q} g(\bar{\mathbf{Q}}, \bar{\mathbf{q}}))^\mathsf{T}\mathbf{q} - t \leq \nabla_\mathbf{q} g(\bar{\mathbf{Q}}, \bar{\mathbf{q}})^\mathsf{T}\bar{\mathbf{q}} + \nabla_\mathbf{Q} g(\bar{\mathbf{Q}}, \bar{\mathbf{q}}) \bullet \bar{\mathbf{Q}} - g(\bar{\mathbf{Q}}, \bar{\mathbf{q}}) \ ,$$

where $g(\mathbf{Q},\mathbf{q}) = \sqrt{\gamma_1}\,\|\Sigma_0^{1/2}(\mathbf{q}+2\mathbf{Q}\mu_0)\|$ and $\nabla_{\mathbf{Q}} g(\mathbf{Q},\mathbf{q})$ and $\nabla_{\mathbf{q}} g(\mathbf{Q},\mathbf{q})$ are its associated gradients in $\mathbf{Q}$ and $\mathbf{q}$ respectively. Finally, given the assumed structure of $h(\mathbf{x},\xi)$, Constraint (5b) can be decomposed into K sub-constraints

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

When considering the k-th sub-constraint, it was already shown in Lemma 3 that $\sup_{\xi\in\mathcal{S}} h_k(\mathbf{x},\xi) - \xi^{\mathsf{T}}\mathbf{Q}\xi - \xi^{\mathsf{T}}\mathbf{q}$ can be solved to an 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 ξ_* as follows:

$$(\xi_*\xi_*{}^{\mathsf{T}} \bullet \mathbf{Q}) + \xi_*{}^{\mathsf{T}}\mathbf{q} + r \geq h_{k^*}(\mathbf{x},\xi_*) \;\; .$$

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

2.2. The Distributionally Robust Stochastic Program with Moment Uncertainty

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

$$\underset{\mathbf{x}}{\text{minimize}} \;\; \left(\max_{f_\xi \,\in\, \mathcal{D}_1} \;\; \mathbb{E}_\xi[h(\mathbf{x},\xi)] \right) \tag{7a}$$
$$\text{subject to} \;\; \mathbf{x} \in \mathcal{X} \;\; . \tag{7b}$$

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 $\mathbf{x} \in \mathbb{R}^n$ either confirm that $\mathbf{x} \in \mathcal{X}$ or provide a hyperplane that separates $\mathbf{x}$ from $\mathcal{X}$ in time polynomial in n.*

ASSUMPTION 4. *The function $h(\mathbf{x},\xi)$ is convex in $\mathbf{x}$ and satisfies Assumption 2. In addition, it is assumed that one can find in polynomial time a sub-gradient of $h(\mathbf{x},\xi)$ in $\mathbf{x}$.*

PROPOSITION 2. *Given that Assumption 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 $\mathbf{x}$ and ξ.*

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 and use the fact that min-min operations can be performed jointly and that the constraint involving $h(\mathbf{x},\xi)$ decomposes. This leads to an equivalent convex optimization form for Problem (7):

$$\underset{\mathbf{x},\mathbf{Q},\mathbf{q},r,t}{\text{minimize}} \;\; r + t \tag{8a}$$
$$\text{subject to} \;\; r \geq h_k(\mathbf{x},\xi) - \xi^{\mathsf{T}}\mathbf{Q}\xi - \xi^{\mathsf{T}}\mathbf{q} \;\; , \;\; \forall\, \xi \in \mathcal{S}, \;\; k \in \{1,...,K\} \tag{8b}$$
$$t \geq \left(\gamma_2\Sigma_0 + \mu_0\mu_0^{\mathsf{T}}\right) \bullet \mathbf{Q} + \mu_0^{\mathsf{T}}\mathbf{q} + \sqrt{\gamma_1}\,\|\Sigma_0^{1/2}(\mathbf{q}+2\mathbf{Q}\mu_0)\| \tag{8c}$$
$$\mathbf{Q} \succeq 0 \tag{8d}$$
$$x \in \mathcal{X} \;\; . \tag{8e}$$

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 because of Lemma 1, we are assured that the optimal solution set is non-empty. The arguments that were presented in the proof of Proposition 1 still apply for Constraint (8c) and (8d). However, the argument for Constraint (8b) needs to be revisited since $\mathbf{x}$ is now considered as an optimization variable. Feasibility of an assignment $(\bar{\mathbf{x}},\bar{\mathbf{Q}},\bar{\mathbf{q}},\bar{r})$ 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 ξ_* and $\nabla_{\mathbf{x}} h_k(\bar{\mathbf{x}}, \xi_*)$, a sub-gradient of $h_{k^*}(\cdot, \xi_*)$ at $\bar{\mathbf{x}}$:

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

Since by Assumption 4, a sub-gradient $\nabla_{\mathbf{x}} h_k(\bar{\mathbf{x}}, \xi_*)$ 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 $\min_{\mathbf{x}} \max_{\xi} h(\mathbf{x}, \xi)$ is a tractable robust optimization problem (*cf*., Ben-Tal and Nemirovski (1998) and Bertsimas et al. (2007)), then the less-conservative DRSP $\min_{\mathbf{x}} \max_{f_\xi \in \mathcal{D}_1} \mathbb{E}[h(\mathbf{x}, \xi)]$ is also tractable. In some cases, the inner moment problem might even be reducible analytically (see Section 4 for an example). Moreover, one still has access to the wide spectrum of methods that have been proposed for robust optimization problems: ranging from methods that use cutting planes more efficiently 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 $\mathbf{Q} \succeq 0$ plays an important role in making Problem (7) solvable in polynomial time. This constraint corresponds to the covariance matrix inequality in our distribution set $\mathcal{D}_1$ construction. If the inequality is replaced by an equality, then $\mathbf{Q}$ is "free" and Problem (6) is no longer a convex optimization problem. This explains why many DRSP problems under the exact-covariance knowledge actually become intractable.

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

2.3. Examples

Because our framework only imposes weak conditions on $h(\mathbf{x}, \xi)$ through Assumption 2 and 4, it is possible to revisit many well-known cases 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 $\mathbb{P}(\xi \in \mathcal{C})$ for a random vector ξ with known mean, and covariance matrix and some closed set $\mathcal{C}$. By formulating this problem as a semi-infinite linear program:

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

many have proposed methods to provide useful 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, *i.e.*, $\mathcal{S} = \mathbb{R}^m$, 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 be NP-hard. In fact, the hardness of this last problem arises already in finding a distribution that is feasible.

Our framework recommends relaxing the restrictions on the covariance of ξ and instead consider the distributional set $\mathcal{D}_1(\mathcal{S}, \mu_0, \Sigma_0, \gamma_1, \gamma_2)$. Such a distributional set constrains all three types of statistics: mean, covariance matrix and support. 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 a $f_\xi \in \mathcal{D}_1$ is already resolved using the Dirac measure δ_{μ_0}.[1] We can also construct an $h(\mathbf{x}, \xi)$ that satisfies Assumption 2 and 4 by choosing $h_0(\mathbf{x}, \xi) = 0$ and $h_k(\mathbf{x}, \xi) = \begin{cases} 1 & \text{, if } \xi \in \mathcal{C}_k \\ -\infty & \text{, otherwise} \end{cases}$. Then, clearly by the fact that

$$\mathbb{E}_\xi[h(\mathbf{x}, \xi)] = \mathbb{E}_\xi[\max_k \ h_k(\mathbf{x}, \xi)] = \mathbb{E}_\xi[\mathbb{1}\{\xi \in \mathcal{C}\}] = \mathbb{P}(\xi \in \mathcal{C}) \leq \max_{f_\xi \in \mathcal{D}_1} \ \mathbb{E}_\xi[h(\mathbf{x}, \xi)] \ ,$$

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

EXAMPLE 2. **Distributionally Robust Optimization with piecewise-linear convex costs.**

Assume that one is interested in solving the following DRSP model for a general piece-wise linear convex cost function of $\mathbf{x}$

$$\underset{\mathbf{x} \in \mathcal{X}}{\text{minimize}} \ \left(\max_{f_\xi \in \mathcal{D}_1} \ \mathbb{E}_\xi[\max_k \ \xi_k^\mathsf{T} \mathbf{x}] \right) \ ,$$

where each $\xi_k \in \mathbb{R}^n$ is a random vector. This is quite applicable since any convex cost function can be approximated by a piecewise linear function. By considering ξ to be a random matrix whose k-th column is the random vector ξ_k and taking $h_k(\mathbf{x}, \xi) = \xi_k^\mathsf{T} \mathbf{x}$, which is linear (hence concave) in ξ, the results presented earlier allows one to conclude that the DRSP can be solved efficiently. In fact, due to $h_k(\mathbf{x}, \xi) - \xi^\mathsf{T} Q \xi - \xi^\mathsf{T} q$ being a concave quadratic function of ξ, 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 in a case of portfolio optimization.

EXAMPLE 3. **Distributionally robust conditional value-at-risk.**

Conditional value-at-risk, also called mean excess loss, was introduced in the mathematical finance community as a new risk measure in 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 numerical properties. For instance, Rockafellar and Uryasev (2000) demonstrated that one can evaluate the ϑ-CVaR$_\xi[c(\mathbf{x}, \xi)]$ of a cost (or loss) function $c(\mathbf{x}, \xi)$ with random parameters distributed according to f_ξ by solving a minimization problem of convex form:

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

where $(y)^+ = \max\{y \ , \ 0\}$. While CVaR is a rich risk measure, it still requires the decision maker to commit to a distribution f_ξ. 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 in this section, 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:

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

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

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

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

$$\underset{\mathbf{x}\in\mathcal{X},\lambda\in\mathbb{R}}{\text{minimize}} \quad \left(\max_{f_\xi\in\mathcal{D}_1} \ \mathbb{E}_\xi[h(\mathbf{x},\lambda,\xi)]\right) \ ,$$

where $h(\mathbf{x},\lambda,\xi)=\lambda+\frac{1}{\vartheta}(c(\mathbf{x},\xi)-\lambda)^+$. Because of the argument that

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

it is clear that $h(\mathbf{x},\lambda,\xi)$ meets Assumption 2 and 4. Hence, Proposition 2 allows us to conclude that finding an optimal $\mathbf{x}$ (and its associated λ) with respect to the worst case conditional value-at-risk obtained 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, $\{\xi_i\}_{i=1}^M$, generated independently and randomly according to an unknown distribution f_ξ. 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 ξ. 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 ξ.

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

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

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

In practice, even when one does not have information about μ and Σ, 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 we derive R based on the bounded support of ξ. In this work, a confidence region for μ and Σ will be derived 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 $\{\xi_i\}_{i=1}^M$ be a set of independent random vectors ξ_i taking values in a set $\mathcal{S}_i$ for each i. Suppose that the real-valued function $g(\xi_1, \xi_2, ..., \xi_M)$ defined on $\mathcal{S}_1 \times \mathcal{S}_2 \times ... \times \mathcal{S}_M$ satisfies*

$$|g(\xi_1, \xi_2, ..., \xi_M) - g(\xi_1', \xi_2', ..., \xi_M')| \leq c_j \tag{10}$$

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

$$\mathbb{P}\left(g(\xi_1, \xi_2, ..., \xi_M) - \mathbb{E}[g(\xi_1, \xi_2, ..., \xi_M)] \leq -t\right) \leq \exp\left(\frac{-2t^2}{\sum_{j=1}^M c_j^2}\right) \quad .$$

3.1. Uncertainty Cone Centered at Empirical Mean

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

The following result was demonstrated from McDiarmid's theorem.

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

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

This result can in turn be used to derive a similar statement about the random vector ξ.

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

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

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

Proof: This generalization for a ξ with arbitrary mean and covariance matrix is quite straightforward:

$$\begin{aligned}
\mathbb{P}\left((\hat{\mu} - \mu)^\mathsf{T}\Sigma^{-1}(\hat{\mu} - \mu) \leq \beta(\delta)\right) &= \mathbb{P}\left(\left\|\Sigma^{-1/2}\left(\frac{1}{M}\sum_{i=1}^M \xi_i - \mu\right)\right\|^2 \leq \beta(\delta)\right) \\
&= \mathbb{P}\left(\left\|\frac{1}{M}\sum_{i=1}^M \Sigma^{-1/2}(\xi_i - \mu)\right\|^2 \leq \beta(\delta)\right) \\
&= \mathbb{P}\left(\left\|\sum_{i=1}^M \zeta_i\right\|^2 \leq \beta(\delta)\right) \geq 1 - \delta \quad . \qquad \square
\end{aligned}$$

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

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

3.2. Uncertainty Cone Centered at Empirical Covariance

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

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

Note that the difficulty of this task relies heavily on the fact that one needs to derive a confidence interval for the eigenvalues of the stochastic matrix $\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}$, which is an important field of study in statistics. For the case that interests us, where $M \gg m$ with M finite and m fixed, prior work usually assumes ξ is a normally distributed random vector (see Anderson (1984) and Edelman (1989)). Under the Gaussian assumption, the sample 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 ξ takes a non-normal form, the asymptotic distribution of eigenvalues was studied by Waternaux (1976) and Fujikoshi (1980) among others. However, to the best of our knowledge, our work is the first to formulate an uncertainty sets 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 $\hat{\mathbf{I}} = M^{-1}\sum_i \zeta_i\zeta_i^\mathsf{T}$ for the mean and covariance matrix of ζ. Next, we will assume that the mean of ξ is exactly known and will formulate the confidence region for Σ in terms of $\hat{\Sigma}(\mu) = M^{-1}\sum_{i=1}^{M}(\xi_i-\mu)(\xi_i-\mu)^\mathsf{T}$. We conclude this section with our main result about a confidence region for μ and Σ which relies solely on M and on support information about the random vector ξ.

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

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

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

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

Proof: The proof of this theorem relies on applying Theorem 1 twice to show that both $\frac{1}{1+\alpha(\delta/2)}\hat{\mathbf{I}} \preceq \mathbf{I}$ and $\mathbf{I} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\mathbf{I}}$ 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 adjusted for the derivation of the lower bound.

When applying Theorem 1 to show that $\mathbf{I} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\mathbf{I}}$ occurs with probability $1-\delta/2$, the main step consists of defining $g(\zeta_1,\zeta_2,...,\zeta_M) = \min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z}$ and finding a lower bound for $\mathbb{E}[g(\zeta_1,\zeta_2,...,\zeta_M)]$. One can start by showing that Constraint (10) is met when $c_j = R^2/M$ for all j.

$$|g(\zeta_1,\zeta_2,...,\zeta_M) - g(\zeta_1',\zeta_2',...,\zeta_M')| = \left|\min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z} - \min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}'\mathbf{z}\right| \ ,$$

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

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

$$\begin{aligned} |g(\zeta_1,\zeta_2,...,\zeta_M) - g(\zeta_1',\zeta_2',...,\zeta_M'))| &= \min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z} - \mathbf{z}^{*\mathsf{T}}\hat{\mathbf{I}}'\mathbf{z}^* \\ &\leq \mathbf{z}^{*\mathsf{T}}(\hat{\mathbf{I}} - \hat{\mathbf{I}}')\mathbf{z}^* \end{aligned}$$

$$
\begin{aligned}
&= {\mathbf{z}^*}^\mathsf{T} \frac{1}{M}(\zeta_j\zeta_j^\mathsf{T} - \zeta_j'\zeta_j'^\mathsf{T})\mathbf{z}^* \\
&= \frac{1}{M}\left((\zeta_j^\mathsf{T}\mathbf{z}^*)^2 - (\zeta_j'^\mathsf{T}\mathbf{z}^*)^2\right) \\
&\leq \frac{\|\mathbf{z}^*\|^2\|\zeta_j\|^2}{M} \leq \frac{R^2}{M} \ .
\end{aligned}
$$

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

As for bounding $\mathbb{E}[g(\zeta_1, \zeta_2, ..., \zeta_M)]$, the task is a bit harder. We can instead try to find an upper bound on the maximum eigenvalue of $(\mathbf{I} - \hat{\mathbf{I}})$ since

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

Using Jensen's inequality and basic linear algebra, one can show that

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

where we used the fact that ζ_i are sampled independently thus making $\mathbb{E}[(\mathbf{I} - \zeta_i\zeta_i^\mathsf{T})(\mathbf{I} - \zeta_j\zeta_j^\mathsf{T})] = \mathbb{E}[\mathbf{I} - \zeta_i\zeta_i^\mathsf{T}]\mathbb{E}[\mathbf{I} - \zeta_j\zeta_j^\mathsf{T}] = 0$. By replacing this lower bound in Equation (15), we can now state that $\mathbb{E}[g(\zeta_1, \zeta_2, ..., \zeta_M)] \geq 1 - (R^2/\sqrt{M})\sqrt{1 - m/R^4}$. More importantly, Theorem 1 allows us to confirm the proposed upper bound using the following argument. Since the statement

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

implies that

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

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

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

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

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

The task of showing that $1/(1 + \alpha(\delta/2))\hat{\mathbf{I}} \preceq \mathbf{I}$ also occurs with probability $1 - \delta/2$ is very similar. One needs to apply Theorem 1, now defining $g(\zeta_1, \zeta_2, ..., \zeta_M) = -\min_{\|\mathbf{z}\|=1} \ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z}$, and to demonstrate that $\mathbb{E}[g(\zeta_1, \zeta_2, ..., \zeta_M)] \geq -1 - \alpha(\delta/2)$. The rest follows easily. $\square$

REMARK 4. Considering for simplicity the single dimension case where one is interested in a confidence region based on $\hat{\mathbf{I}} = \sum_{i=1}^{M} \zeta_i^2$, one can easily verify that $\alpha(\delta)$ is asymptotically of the right order in terms of M and R. Since $\mathbb{E}[\zeta^4]$ is bounded by R^4, the central limit theorem guarantees that $\sqrt{M}(\hat{\mathbf{I}} - \mathbb{E}[\zeta^2])$ converges in distribution to $\mathcal{N}(0, \mathbb{E}[\zeta^4] - 1)$. Thus, it follows that $(M/(\mathbb{E}[\zeta^4] - 1))\|\hat{\mathbf{I}} - \mathbb{E}[\zeta^2]\|^2$ converges in distribution to a χ^2-distribution with degree 1. For any $\delta > 0$, one can find $c(\delta)$ such that with probability greater than $1 - \delta$, $\|\hat{\mathbf{I}} - \mathbb{E}[\zeta^2]\| \leq \frac{c(\delta)\sqrt{\mathbb{E}[\zeta^4]-1}}{\sqrt{M}}$. Hence, asymptotically speaking the confidence region $-\frac{1}{1+\frac{c(\delta)R^2}{\sqrt{M}}}\hat{\mathbf{I}} \preceq \mathbf{I} \preceq \frac{1}{1-\frac{c(\delta)R^2}{\sqrt{M}}}\hat{\mathbf{I}}$ 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:

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

and similarly,

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

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

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

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

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

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 ξ with high probability.

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

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

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

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

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

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:

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

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 covariance matrix estimates 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 (16c) to hold.

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

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

$$\begin{aligned}\mathbf{x}^\mathsf{T}(\hat{\mu}-\mu)(\hat{\mu}-\mu)^\mathsf{T}\mathbf{x} &= (\mathbf{x}^\mathsf{T}(\hat{\mu}-\mu))^2 = (\mathbf{x}^\mathsf{T}\Sigma^{1/2}\,\Sigma^{-1/2}(\hat{\mu}-\mu))^2\\ &\leq \|\mathbf{x}^\mathsf{T}\Sigma^{1/2}\|^2\|\Sigma^{-1/2}(\hat{\mu}-\mu)\|^2 \leq \beta(\delta/2)\mathbf{x}^\mathsf{T}\Sigma\mathbf{x} \quad .\end{aligned}$$

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

$$\begin{aligned}(1+\alpha(\delta/4))\Sigma \succeq \hat{\Sigma}(\mu) &= \frac{1}{M}\sum_{i=1}^{M}(\xi_i-\mu)(\xi_i-\mu)^\mathsf{T}\\ &= \hat{\Sigma}+(\hat{\mu}-\mu)(\hat{\mu}-\mu)^\mathsf{T}\\ &\succeq \hat{\Sigma} \quad .\end{aligned}$$

□

3.3. Bounding the Support of ζ using Empirical Data

The above derivations assumed that one is able to describe a ball containing the support of the rather fictive random vector ζ. In fact, this assumption can be replaced by an assumption on the support of the more tangible random vector ξ as is presented in the following corollary.

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

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

be a stochastic 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 $1-\delta$, Constraint (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 $\bar{\beta}(\bar{\delta}/2) = (\bar{R}^2/M)(2+\sqrt{2\ln(2/\bar{\delta})})^2$ respectively.*

Proof: Since we assumed that Σ was non-singular, the support of ξ being bounded by a ball of radius R_ξ implies that ζ is also bounded. Thus, there exists an R such that $\mathbb{P}(\|\zeta\| \leq R) = 1$. Given that ζ has a bounded support and the size of M, Theorem 4 guarantees us that with probability greater than $1-\bar{\delta}$, Constraint (16a), (16b) and (16c) are met. Thus

$$\begin{aligned}
R &= \sup_{\zeta\in\mathcal{S}_\zeta}\|\zeta\|_2 = \sup_{\xi\in\mathcal{S}_\xi}\|\Sigma^{-1/2}(\xi-\mu)\|_2 = \sup_{\xi\in\mathcal{S}_\xi}\|\Sigma^{-1/2}(\xi-\mu+\hat{\mu}-\hat{\mu})\|_2 \\
&\leq \sup_{\xi\in\mathcal{S}_\xi}\|\Sigma^{-1/2}(\xi-\hat{\mu})\|_2 + \|\Sigma^{-1/2}(\hat{\mu}-\mu)\|_2 \\
&\leq \sup_{\xi\in\mathcal{S}_\xi}\sqrt{1+\alpha(\bar{\delta}/4)}\|\hat{\Sigma}^{-1/2}(\xi-\hat{\mu})\|_2 + \sqrt{\beta(\bar{\delta}/2)} \\
&\leq \sqrt{1+\alpha(\bar{\delta}/4)}\hat{R} + \sqrt{\beta(\bar{\delta}/2)} \\
&\leq \hat{R}\sqrt{1+cR^2}+cR \ ,
\end{aligned}$$

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

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 Constraint (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

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

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

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

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 Constraint (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 most practical situations where one needs to deal with uncertainty in the parameters, it might not be clear how to define an uncertainty set for the mean and covariance matrix of the random vector of parameters ξ. It is more likely that one only has in hand a set of independent samples, $\{\xi_i\}_{i=1}^M$, drawn according to the distribution of ξ and wishes to solve a form of the DRSP model for which it is guaranteed that with high probability the solution is robust with respect to the unknown random vector ξ.

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

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

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^M \xi_i \ , \qquad \hat{\Sigma} = \frac{1}{M}\sum_{i=1}^M (\xi_i - \hat{\mu})(\xi_i - \hat{\mu})^\mathsf{T}$$
$$\bar{\gamma}_1 = \frac{\bar{\beta}(\bar{\delta}/2)}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)} \ , \qquad \bar{\gamma}_2 = \frac{1+\bar{\beta}(\bar{\delta}/2)}{1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2)} \ .$$

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; hence, $\bar{\gamma}_1 \to 0$ and $\bar{\gamma}_2 \to 1$ as M goes to infinity.

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

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

$$(1-\bar{\alpha}(\bar{\delta}/4)-\bar{\beta}(\bar{\delta}/2))(\hat{\mu}-\mu)\hat{\Sigma}^{-1}(\hat{\mu}-\mu) \leq (\hat{\mu}-\mu)\Sigma^{-1}(\hat{\mu}-\mu) \leq \bar{\beta}(\bar{\delta}/2) \ ,$$

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

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

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

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

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

We can now extend the results presented in sections 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 $\{\xi_i\}_{i=1}^M$ be a set of M samples generated independently at random according to the distribution f_ξ which support is contained in the set $\mathcal{S}$. For any $\delta > 0$, if Assumption 1, 2, 3 and 4 are satisfied then, given the set $\{\xi_i\}_{i=1}^M$, one can solve in polynomial time Problem* (7) *under the set $\mathcal{D}_1(\mathcal{S}, \hat{\mu}, \hat{\Sigma}, \bar{\gamma}_1, \bar{\gamma}_2)$ where $\hat{\mu}$, $\hat{\Sigma}$, $\bar{\gamma}_1$ 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 $\{\xi_i\}_{i=1}^M$, we have that any optimal solution $\mathbf{x}^*$ of the DRSP formed using these samples will satisfy the constraint*

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

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

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

where $\hat{\mu}$, $\hat{\Sigma}$, $\bar{\gamma}_1$ 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 $\{\xi_i\}_{i=1}^M$, we have that*

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

4. Application to Portfolio Optimization

We now turn ourselves to applying our framework to an instance of portfolio optimization. In such a problem, one is interested in maximizing his expected utility for the potential one step return of an investment portfolio. Given that n investment options are available, expected utility can be defined as $\mathbb{E}[u(\xi^\mathsf{T}\mathbf{x})]$, where $u(\cdot)$ is a non-decreasing function and $\xi \in \mathbb{R}^n$ 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_ξ and choose the portfolio which is optimal according to the following Distributionally Robust Portfolio Optimization model:

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

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

In Popescu (2007), the author addresses the case of Problem (18) where $\mathbb{E}[\xi]$ and $\mathbb{E}[\xi\xi^\mathsf{T}]$ are known exactly and one considers $\mathcal{D}$ to be the set of all distribution 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 function $u(\cdot)$. This approach is interesting as it provides a mean of taking into account uncertainty in the form of the distribution of returns. Unfortunately, our experiments will show that

in practice it is highly sensitive to the noise in the empirical estimation of these moments. Secondly, the proposed algorithm also relies on solving a one dimensional non-convex mathematical program. Thus, it is not guaranteed to converge to an 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 propose accounting for moment uncertainty in Markowitz models. Their motivation is closely aligned with ours and many of the proposed techniques can be applied in our context: *e.g.*, the use of factor models to reduce the dimensionality of ξ. Similarly, the results presented in Section 3 for the 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, the robust Markowitz model typically gives rise to problems that are easier to solve.

4.1. Portfolio Optimization with Moment Uncertainty

In order to apply our framework we need to assume that the utility function is piecewise linear concave, such that $u(y) = \min_{k\in\{1,2,...,K\}} \; a_k y + b_k$. This is not too constraining since in portfolio optimization the interesting utility functions are usually concave and such functions always have good piecewise linear approximation with finite K. We use historical knowledge of investment returns $\{\xi_1, \xi_2, ..., \xi_M\}$ to define a distributional uncertainty set for f_ξ. This can be done using the set $\mathcal{D}_1(\mathcal{S}, \hat{\mu}, \hat{\Sigma}, \gamma_1, \gamma_2)$ where $\hat{\mu}$ and $\hat{\Sigma}$ are assigned as the empirical estimates of the mean $\hat{\mu} = M^{-1}\sum_{i=1}^{M} \xi_i$ and covariance matrix $\hat{\Sigma} = M^{-1}\sum_{i=1}^{M}(\xi_i - \hat{\mu})(\xi_i - \hat{\mu})^\mathsf{T}$ of ξ respectively.[3] We consider two options for the choice of $\mathcal{S}$: either $\mathcal{S} = \mathbb{R}^n$, or an ellipsoidal set $\mathcal{S} = \{\xi | (\xi - \xi_0)^\mathsf{T}\mathbf{\Theta}(\xi - \xi_0) \leq 1\}$, with $\mathbf{\Theta} \succeq 0$.

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 $\mathcal{X}$ satisfies Assumption 3, finding an optimal solution $\mathbf{x} \in \mathbb{R}^n$ to the DRPO model, Problem* (18)*, equipped with the set of distributions $\mathcal{D}_1(\mathcal{S}, \hat{\mu}, \hat{\Sigma}, \gamma_1, \gamma_2)$ can be done in $O(n^{6.5})$.*

Proof: We first reformulate the objective of Problem (18) in its minimization form :

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

After confirming that $\mathcal{S}$ satisfies the weaker version of Assumption 1 (see Remark 3) and that $h(\mathbf{x}, \xi) = \max_k \; -a_k\xi^\mathsf{T}\mathbf{x} - b_k$ satisfies Assumption 2 and 4, a straightforward application of Proposition 2 already confirms 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 and exploit the special structure of $h(\mathbf{x}, \xi)$ in Problem (18):

$$\begin{aligned}
\underset{\mathbf{x},\mathbf{Q},\mathbf{q},r,\mathbf{P},\mathbf{p},s}{\text{minimize}} \quad & \gamma_2(\Sigma_0 \bullet \mathbf{Q}) - \mu_0^\mathsf{T}\mathbf{Q}\mu_0 + r + (\Sigma_0 \bullet \mathbf{P}) - 2\mu_0^\mathsf{T}\mathbf{p} + \gamma_1 s && \text{(19a)}\\
\text{subject to} \quad & \begin{bmatrix} \mathbf{P} & \mathbf{p} \\ \mathbf{p}^\mathsf{T} & s \end{bmatrix} \geq 0 \;\; , \;\; \mathbf{p} = -\mathbf{q}/2 - \mathbf{Q}\hat{\mu} \;\; , \;\; \mathbf{Q} \succeq 0 && \text{(19b)}\\
& \xi^\mathsf{T}\mathbf{Q}\xi + \xi^\mathsf{T}\mathbf{q} + r \geq -a_k\xi^\mathsf{T}\mathbf{x} - b_k \;\; , \; \forall \xi \in \mathcal{S}, k \in \{1, 2, ..., K\} && \text{(19c)}\\
& \sum_{i=1}^{n} \mathbf{x}_i = 1 \;\; , \;\; \mathbf{x}_i \geq 0 \;\; , \; \forall i \;\; . && \text{(19d)}
\end{aligned}$$

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

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

While if $\mathcal{S}$ is an ellipsoid, the S-Lemma (*cf.*, Pólik and Terlaky (2007)) can be used to replace Constraint (19c)

$$\begin{bmatrix}\xi\\ 1\end{bmatrix}^\mathsf{T}\begin{bmatrix}\mathbf{\Theta} & -\mathbf{\Theta}\xi_0\\ -\xi_0^\mathsf{T}\mathbf{\Theta} & \xi_0^\mathsf{T}\mathbf{\Theta}\xi_0 - 1\end{bmatrix}\begin{bmatrix}\xi\\ 1\end{bmatrix} \leq 0 \rightarrow \begin{bmatrix}\xi\\ 1\end{bmatrix}^\mathsf{T}\begin{bmatrix}\mathbf{Q} & \mathbf{q}/2 + a_k\mathbf{x}/2\\ \mathbf{q}^\mathsf{T}/2 + a_k\mathbf{x}^\mathsf{T}/2 & r + b_k\end{bmatrix}\begin{bmatrix}\xi\\ 1\end{bmatrix} \geq 0 \ ,$$

with an equivalent constraint:

$$\begin{bmatrix}\mathbf{Q} & \mathbf{q}/2 + a_k\mathbf{x}/2\\ \mathbf{q}^\mathsf{T}/2 + a_k\mathbf{x}^\mathsf{T}/2 & r + b_k\end{bmatrix} \succeq -\tau_k\begin{bmatrix}\mathbf{\Theta} & -\mathbf{\Theta}\xi_0\\ -\xi_0^\mathsf{T}\mathbf{\Theta} & \xi_0^\mathsf{T}\mathbf{\Theta}\xi_0 - 1\end{bmatrix} \ , \ \tau_k \geq 0 \ ,$$

where τ_k, $k \in \{1, ..., K\}$, are extra slack variables. The problem can therefore also be reformulated as a semi-definite program:

$$\begin{array}{ll}\underset{\mathbf{x},\mathbf{Q},\mathbf{q},r,\mathbf{P},\mathbf{p},s,\tau}{\text{minimize}} & \gamma_2(\hat{\Sigma}\bullet\mathbf{Q}) - \hat{\mu}^\mathsf{T}\mathbf{Q}\hat{\mu} + r + (\hat{\Sigma}\bullet\mathbf{P}) - 2\hat{\mu}^\mathsf{T}\mathbf{p} + \gamma_1 s \\ \text{subject to} & \begin{bmatrix}\mathbf{P} & \mathbf{p}\\ \mathbf{p}^\mathsf{T} & s\end{bmatrix} \succeq 0 \ , \ \mathbf{p} = -\mathbf{q}/2 - \mathbf{Q}\hat{\mu} \ , \ \mathbf{Q} \succeq 0 \\ & \begin{bmatrix}\mathbf{Q} & \mathbf{q}/2 + a_k\mathbf{x}/2\\ \mathbf{q}^\mathsf{T}/2 + a_k\mathbf{x}^\mathsf{T}/2 & r + b_k\end{bmatrix} \succeq -\tau_k\begin{bmatrix}\mathbf{\Theta} & -\mathbf{\Theta}\xi_0\\ -\xi_0^\mathsf{T}\mathbf{\Theta} & \xi_0^\mathsf{T}\mathbf{\Theta}\xi_0 - 1\end{bmatrix} \ , \ \forall k \\ & \tau_k \geq 0 \ \ \forall k \\ & \sum_{i=1}^n \mathbf{x}_i = 1 \ , \ \mathbf{x}_i \geq 0 \ , \ \forall i \ .\end{array}$$

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

$$\begin{array}{ll}\underset{\mathbf{x}\in\mathbb{R}^{\tilde{n}}}{\text{minimize}} & c^\mathsf{T}\mathbf{x} \\ \text{subject to} & A_i(\mathbf{x}) \succeq 0 \ \ \forall i = 1, 2, ..., \tilde{K}\end{array}$$

in $O\left(\left(\sum_i^{\tilde{K}} \tilde{m}_i\right)^{0.5}\left(\tilde{n}^2\sum_i^{\tilde{K}} \tilde{m}_i^2 + \tilde{n}\sum_i^{\tilde{K}} \tilde{m}_i^3\right)\right)$, where $\tilde{m}_i$ stands for the dimension of the positive semi-definite cone (*i.e.*, $A_i(\mathbf{x}) \in \mathbb{R}^{\tilde{m}_i \times \tilde{m}_i}$) (see Nesterov and Nemirovski (1994)). In both SDP that interests us here, one can show that $\tilde{n} \leq n^2 + 4n + 2 + K$ and that the problem 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 Popescu and Bertsimas (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 our SDP models actually address the more involved problem of making robust decisions and don't result 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 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 structure in problems where subsets (or linear combinations of assets) are known to behave independently from each other.

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 the results presented in Section 4.2. 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 portfolio performance. Their approach is therefore susceptible, in practice, to the same deficiencies as Popescu's method when these moments are estimated using historical data.

4.2. A Case where the Worse Distribution has Largest Covariance Matrix

When presenting our distributionally robust framework, we argued in Remark 1 that a positive semi-definite lower bound on the covariance matrix was uninteresting. Actually, in the case of a portfolio optimization problem with piecewise concave utility function, the argument can be made more formally. The proof of the following proposition also provides valuable insight 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 infinite support constraint on the distribution is an instance of distributionally robust optimization where the upper positive semi-definite constraint on the covariance matrix is tight for a worst case distribution.*

Proof: Consider the inner problem of our robust portfolio optimization with unconstrained support for the distribution:

$$\max_{f_\xi \,\in\, \mathcal{D}_1(\mathbb{R}^m,\hat{\mu},\hat{\Sigma},0,\gamma_2)} \mathbb{E}_\xi[\max_k \; -a_k\xi^\mathsf{T}\mathbf{x} - b_k] \;\; . \tag{20}$$

For simplicity of our derivations, we consider that there is no uncertainty in the mean of the distribution (*i.e.*, $\gamma_1 = 0$). The dual of this problem can be formulated as:

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

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.

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

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

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

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

We can show that there always exists an optimal solution such that Constraint (21b) is satisfied with equality. Given an optimal assignment $X^* = \{(\Lambda_k^*, \lambda_k^*, \nu_k^*)\}_{k=1}^K$ such that $\Delta = \gamma_2\hat{\Sigma} + \hat{\mu}\hat{\mu}^\mathsf{T} - \sum_{k=1}^K \Lambda_k^* \succeq 0$, consider an alternate solution $X' = \{(\Lambda_k', \lambda_k', \nu_k')\}_{k=1}^K$ which is exactly the same as the original solution X^* except for $\Lambda_1' = \Lambda_1^* + \Delta$. Obviously the two solutions achieve the same objective values since $\{(\lambda_k^*, \nu_k^*)\}_{k=1}^K$ and $\{(\lambda_k', \nu_k;)\}_{k=1}^K$ are the same. If we can show that X' is also feasible then it is necessarily optimal. The only feasibility constraint that seriously needs to be verified is the following:

$$\begin{bmatrix} \Lambda_1' & \lambda_1' \\ \lambda_1'^\mathsf{T} & \nu_1' \end{bmatrix} = \begin{bmatrix} \Lambda_1^* & \lambda_1^* \\ \lambda_1^{*\mathsf{T}} & \nu_1^* \end{bmatrix} + \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \succeq 0 \ ,$$

and is necessarily satisfied since by definition X^* is feasible and that by construction Δ 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 $\sum_{k=1}^K a_k \mathbf{x}^\mathsf{T}\lambda_k^* + \nu_k^* b_k$ 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 $(\zeta_1, \zeta_2, ..., \zeta_K)$ that satisfy the following conditions:

$$\mathbb{E}[\zeta_k] = \frac{1}{\nu_k^*}\lambda_k^* \ , \ \mathbb{E}[\zeta_k\zeta_k^\mathsf{T}] = \frac{1}{\nu_k^*}\Lambda_k^* \ .$$

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

$$\begin{aligned}
\mathbb{E}[\zeta_k\zeta_k^\mathsf{T}] - \mathbb{E}[\zeta_k]\mathbb{E}[\zeta_k]^\mathsf{T} &= \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^\mathsf{T} \end{bmatrix}^\mathsf{T} \begin{bmatrix} \mathbb{E}[\zeta_k\zeta_k^\mathsf{T}] & \mathbb{E}[\zeta_k] \\ \mathbb{E}[\zeta_k]^\mathsf{T} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^\mathsf{T} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^\mathsf{T} \end{bmatrix}^\mathsf{T} \begin{bmatrix} \frac{1}{\nu_k}\Lambda_k^* & \frac{1}{\nu_k}\lambda_k^* \\ \frac{1}{\nu_k}\lambda_k^{*\mathsf{T}} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^\mathsf{T} \end{bmatrix} \\
&= \frac{1}{\nu_k^*}\begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^\mathsf{T} \end{bmatrix}^\mathsf{T} \begin{bmatrix} \Lambda_k^* & \lambda_k^* \\ \lambda_k^{*\mathsf{T}} & \nu_k^* \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^\mathsf{T} \end{bmatrix} \ \succeq \ 0 \ .
\end{aligned}$$

Hence, the random vectors $(\zeta_1, \zeta_2, ..., \zeta_K)$ exists. For instance, if $\mathbb{E}[(\zeta_k - \mathbb{E}[\zeta_k])(\zeta_k - \mathbb{E}[\zeta_k])^\mathsf{T}] \succ 0$, then ζ_k can take the form of a multivariate Gaussian distribution with such mean and covariance matrix. Otherwise, one can construct a lower dimensional random vector; for instance, if $\mathbb{E}[(\zeta_k - \mathbb{E}[\zeta_k])(\zeta_k - \mathbb{E}[\zeta_k])^\mathsf{T}] = 0$ then the random vector is the Dirac measure $\delta_{\mathbb{E}[\zeta_k]}$.

Let $\tilde{k}$ be an independent multinomial 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 $\xi = \zeta_{\tilde{k}}$. Since X^* satisfies Constraint (21b) and (21c) tightly, one can show that the distribution function of ξ^* lies in $\mathcal{D}(\mathbb{R}^m, \hat{\mu}, \hat{\Sigma}, 0, \gamma_2)$ and has largest covariance.

$$\begin{aligned}
\mathbb{E}[\xi^*] &= \sum_{k=1}^K \mathbb{E}[\zeta_k|\tilde{k} = k]\mathbb{P}(\tilde{k} = l) = \sum_{k=1}^K \frac{1}{\nu_k^*}\lambda_k^*\nu_k^* = \hat{\mu} \\
\mathbb{E}[\xi^*\xi^{*\mathsf{T}}] &= \sum_{k=1}^K \mathbb{E}[\zeta_k\zeta_k^\mathsf{T}|\tilde{k} = k]\mathbb{P}(\tilde{k} = l) = \sum_{k=1}^K \frac{1}{\nu_k^*}\Lambda_k^*\nu_k^* = \gamma_2\hat{\Sigma} + \hat{\mu}\hat{\mu}^\mathsf{T}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left[\max_l \ -a_l\mathbf{x}^\mathsf{T}\xi^* - b_l\right] &= \sum_{k=1}^K \mathbb{E}\left[\max_l \ -a_l\mathbf{x}^\mathsf{T}\zeta_{\tilde{k}} - b_l \middle| \tilde{k} = k\right] \mathbb{P}(\tilde{k} = k) \\
&\geq \sum_{k=1}^K \mathbb{E}[-a_k\mathbf{x}^\mathsf{T}\zeta_k - b_k]\mathbb{P}(\tilde{k} = k)
\end{aligned}$$

$$= \sum_{k=1}^{K} -a_k \mathbf{x}^\top \lambda_k^* - b_k \nu_k^*$$
$$= \max_{f_\xi \in \mathcal{D}_1(\mathbb{R}^m, \hat{\mu}, \hat{\Sigma}, 0, \gamma_2)} \mathbb{E}_\xi[\max_k -a_k \mathbf{x}^\top \xi - b_k]$$

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

An interesting consequence of Proposition 3 is that in the framework considered in Popescu (2007), if the utility function is piecewise concave, one can find the optimal portfolio in polynomial time using our semi-definite programming formulation with the distributional set $\mathcal{D}_1(\mathbb{R}^m, \hat{\mu}, \hat{\Sigma}, 0, 1)$. 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).

4.3. Experiments

We evaluate our portfolio optimization method with stock market investments. We use a historical dataset of 30 assets over a horizon of 15 years (1992-2007), obtained from the Yahoo! Finance.[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 limited to using 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 is not much to generate good empirical estimates of the mean and covariance matrix of returns; however, using a larger history would cause the assumption of independent and identical samples to be somewhat unrealistic.

In implementing our method, referred as the DRPO model, the distributional set is formulated as $\mathcal{D}_1(\mathbb{R}^4, \hat{\mu}, \hat{\Sigma}, 1.35, 8.32)$, where $\hat{\mu}$ and $\hat{\Sigma}$ are the empirical estimates of the mean and covariance of ξ respectively. Due to the sample size being too small to use $\bar{\gamma}_1$ and $\bar{\gamma}_2$ from 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 of the distribution f_ξ is assumed to be equal to the empirical estimates measured on the last 30 days period. The method is also compared to a naive approximation of the stochastic program, referred as the SP model, in which the selected portfolio is the one that maximizes the average utility over the last 30 days period. We believe that the statistics obtained over the set of 300 experiments 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.

Method	Single Day		2001-2004		2004-2007	
	Avg. utility	1-perc.	Avg. yearly return	10-perc.	Avg. yearly return	10-perc.
Our DRPO model	1.000	0.983	0.944	0.846	1.1017	1.025
Popescu's DRPO model	1.000	0.975	0.700	0.334	1.047	0.9364
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 1%-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 years period when managing a portfolio of 4 assets on a daily basis with either of the three methods. 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 level of economic growth. The figures also indicate periodically the 10% and 90% percentile of the wealth distribution over the set of 300 experiments. The statistics of the long term experiments demonstrate

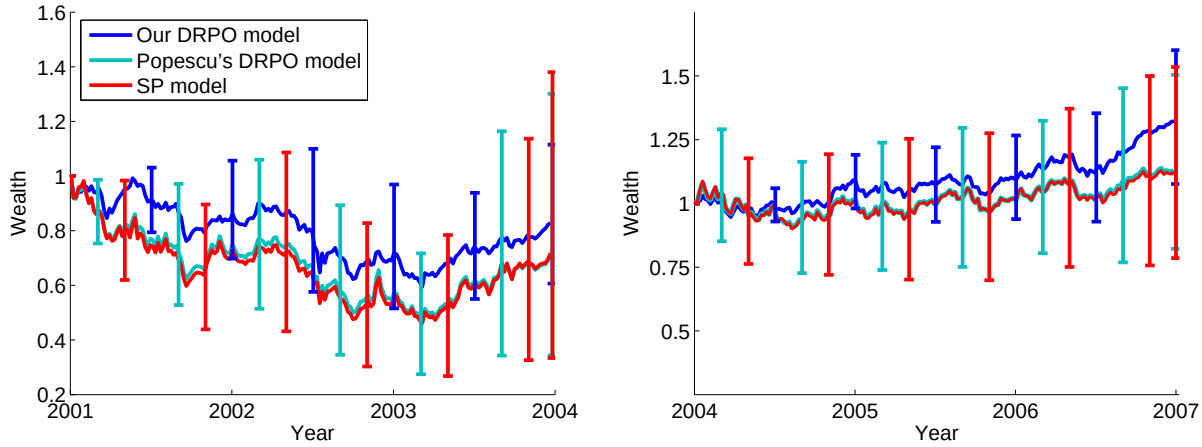

Figure 1 Comparison of wealth evolution in 300 experiments conducted over the years 2001-2007 using three different portfolio optimization models. For each model, the figures indicate periodically the 10% and 90% percentile of the wealth distribution in the set of experiments.

empirically that our method significantly outperforms the two other ones in terms of average return and risks 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 cumulated over the period 2001-2007 in 79.2% of our experiments (total set of 300 experiments). Also, it performed on average at least 1.67 times better than any competing models. Note that these experiments are purely illustrative of the strengths and weaknesses of the different models. For instance, the returns obtained in each experiment does 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 will prove that strong duality holds and that the solution $\Psi(\mathbf{x}, \gamma_1, \gamma_2)$ is bounded.

STEP 1. One can first find through formulating the Lagrangian of Problem (3) that the dual can take the following form

$$\begin{aligned}
\underset{r,\mathbf{Q},\mathbf{P},\mathbf{p},s}{\text{minimize}} \quad & (\gamma_2\Sigma_0 - \mu_0\mu_0^\intercal) \bullet \mathbf{Q} + r + (\Sigma_0 \bullet \mathbf{P}) - 2\mu_0^\intercal \mathbf{p} + \gamma_1 s && \text{(22a)}\\
\text{subject to} \quad & \xi^\intercal \mathbf{Q}\xi + 2\xi^\intercal(\mathbf{p} - \mathbf{Q}\mu_0) + r - h(\mathbf{x},\xi) \geq 0 \ , \ \forall\, \xi \in \mathcal{S} && \text{(22b)}\\
& \mathbf{Q} \succeq 0 && \text{(22c)}\\
& \begin{bmatrix} \mathbf{P} & \mathbf{p} \\ \mathbf{p}^\intercal & s \end{bmatrix} \succeq 0 \ , && \text{(22d)}
\end{aligned}$$

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

We can further simplify This equivalence by finding analytical solutions for the variables $(\mathbf{P}, \mathbf{p}, s)$ in terms of some fixed $(\mathbf{Q}, \mathbf{q}, r)$. 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 $\mathbf{p}^* = 0$ otherwise ${\mathbf{p}^*}^\intercal \mathbf{p}^* > 0$ and

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

which contradicts Constraint (22d). Finally, $\mathbf{P}^* = 0$ is an optimal solution since it minimizes the objective. If $s^* = 0$ then, after replacing $\mathbf{q} = 2(\mathbf{p} - \mathbf{Q}\mu_0)$, Problem (22)'s objective does indeed reduces to

$$\gamma_2(\Sigma_0 \bullet \mathbf{Q}) - \mu_0^\mathsf{T}\mathbf{Q}\mu_0 + r \;\; = \;\; r + \gamma_2(\Sigma_0 \bullet \mathbf{Q}) + \mu_0^\mathsf{T}\mathbf{Q}\mu_0 + \mu_0^\mathsf{T}\mathbf{q} + \sqrt{\gamma_1}\|\Sigma_0^{1/2}(\mathbf{q} + 2\mathbf{Q}\mu_0)\| \;\; .$$

If instead one assumes that $s^* > 0$, then by applying Schur's complement, Constraint (22d) can be shown equivalent to $\mathbf{P} \succeq \frac{1}{s}\mathbf{p}\mathbf{p}^\mathsf{T}$. Since $\Sigma_0 \succeq 0$, $\mathbf{P}^* = \frac{1}{s}\mathbf{p}\mathbf{p}^\mathsf{T}$ is a valid optimal solution and can be replaced in the objective. It remains to solve for $s^* > 0$ in the one dimensional convex optimization problem $\text{minimize}_{s>0}\, \frac{1}{s}\mathbf{p}^\mathsf{T}\Sigma_0\mathbf{p} + \gamma_1 s$.By setting the derivative of the objective function to zero, we obtain that $s^* = \sqrt{\frac{1}{\gamma_1}\mathbf{p}^\mathsf{T}\Sigma_0\mathbf{p}}$. Thus, once again, after replacing $\mathbf{q} = 2(\mathbf{p} - \mathbf{Q}\mu_0)$, the optimal value of Problem (22) reduces to the form of Problem (5):

$$r + \gamma_2(\Sigma_0 \bullet \mathbf{Q}) + \mu_0^\mathsf{T}\mathbf{Q}\mu_0 + \mu_0^\mathsf{T}\mathbf{q} + \sqrt{\gamma_1}\|\Sigma_0^{1/2}(\mathbf{q} + 2\mathbf{Q}\mu_0)\| \;\; .$$

STEP 2. One can easily show that the conditions on γ_1, γ_2 and Σ_0 are sufficient to ensure that the Dirac measure δ_{μ_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. One can also show that $\Psi(\mathbf{x}; \gamma_1, \gamma_2)$ is bounded above by deriving a feasible assignment for the variables of Problem (5). Choosing $\mathbf{Q} = \mathbf{I}$ and $\mathbf{q} = 0$ ensures that the function $h(\mathbf{x}, \xi) - \xi^\mathsf{T}\mathbf{Q}\xi - \xi^\mathsf{T}\mathbf{q}$ is strictly concave thus enforcing $\sup_{\xi\in\mathcal{S}} h(\mathbf{x}, \xi) - \xi^\mathsf{T}\mathbf{Q}\xi - \xi^\mathsf{T}\mathbf{q}$ to be finite. It then follows that letting $r = \sup_{\xi\in\mathcal{S}} h(\mathbf{x}, \xi) - \xi^\mathsf{T}\mathbf{Q}\xi - \xi^\mathsf{T}\mathbf{q}$ and $t = (\gamma_2\Sigma_0 + \mu_0\mu_0^\mathsf{T}) \bullet \mathbf{Q} + \mu_0^\mathsf{T}\mathbf{q} + \sqrt{\gamma_1}\, \|\Sigma_0^{1/2}(\mathbf{q} + 2\mathbf{Q}\mu_0)\|$ constitutes, with $\mathbf{Q} = \mathbf{I}$ and $\mathbf{q} = 0$, a feasible solution that bounds Problem (5). We conclude that $\Psi(\mathbf{x}; \gamma_1, \gamma_2)$ must be finite and that the set of optimal solutions to Problem (5) must be non-empty. □

Notes

[1]Recall that the Dirac measure δ_a is the measure of mass one at the point a.

[2]Note that if ξ's support set is unbounded, one can also derive bounds of similar nature either by considering that ζ 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 $\hat{\Sigma} \succ 0$.

[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 $\mathcal{D}(\gamma_1, \gamma_2)$, the distributional set contains, with 99% probability, distributions with moments equal to the moments estimated from the second 30 days of the period.

Acknowledgments

The authors acknowledge the Fonds Québécois de la recherche sur la nature et les technologies and Boeing 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, New York, NY, USA.

Ben-Tal, A., A. Nemirovski. 1998. Robust convex optimization. *Mathematics of Operations Research* **23**(4) 769–805.

Bertsimas, D., D. B. Brown, C. Caramanis. 2007. Theory and applications of robust optimization.

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

Bertsimas, D., S. Vempala. 2004. Solving convex programs by random walks. *Journal of the 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. *Mathematics of Operations Research* **12**(1) 149–162.

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

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

De Farias, D. P., B. Van Roy. 2001. On constraint sampling for the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research* **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, Boston, MA, USA.

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. *Annals of Operations Research* **31** 347–370.

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

Goldfarb, D., G. Iyengar. 2003. Robust portfolio selection problems. *Mathematics of Operations Research* **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*. Akademie-Verla, Berlin, 86–103.

Lagoa, C. M., B. R. Barmish. 2002. Distributionally robust monte carlo simulation: A tutorial survey. *Proceedings of the IFAC World Congress*. Barcelona,Spain, 1–12.

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

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. Working Paper.

Nesterov, Y., A. Nemirovski. 1994. *Interior-point polynomial methods in convex programming*. SIAM, Philadelphia, PA, USA.

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

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

Popescu, I., D. Bertsimas. 2000. Moment problems via semidefinite programming: Applications in probability and finance.

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*. 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. 2000. Stochastic programming by monte carlo simulation methods. Stochastic Programming E-Print Series.

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. Program.* **107**(1) 91–96.

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 2005: Selected Papers of the Annual International Conference of the German Operations Research Society*. Springer Verlag, 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. *Mathematical Programming* **78**(1) 85–104.

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

Zhu, S. S., M. Fukushima. 2005. Worst-case conditional value-at-risk with application to robust portfolio management. Technical Report, Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University.

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