The Value of Stochastic Modeling in Two-Stage Stochastic Programs with Cost Uncertainty

Erick Delage Sharon Arroyo Yinyu Ye

July 15, 2013

Abstract

Although stochastic programming is probably the most effective frameworks for handling decision problems that involve uncertain variables, it is always a costly task to formulate the stochastic model that accurately embodies our knowledge of these variables. In practice, this might require one to collect a large amount of observations, to consult with experts of the specialized field of practice, or to make simplifying assumptions about the underlying system. When none of these options seem feasible, a common heuristic has been to simply seek the solution of a version of the problem where each uncertain variable takes on its expected value (otherwise known as the solution of the mean value problem). In this paper, we show that when 1) the stochastic program takes the form of a two-stage mixed-integer stochastic linear programs, and 2) the uncertainty is limited to the objective function, the solution of the mean value problem is in fact robust with respect to the selection of a stochastic model. We also propose tractable methods that will bound the actual value of stochastic modeling: i.e., how much improvement can be achieved by investing more efforts in the resolution of the stochastic model. Our framework is applied to an airline fleet composition problem. In the three cases that are considered, our results indicate that resolving the stochastic model can not lead to more than a 7% improvement of expected profits thus providing arguments against the need to develop these more sophisticated models.

1 Introduction

The stochastic programming framework can effectively account for uncertainty in a decision problem. Unfortunately, in practice it is often a costly task to fully resolve the stochastic model that characterizes this uncertainty. Such a task might require one to collect a large amount of observations, to consult with different experts in the field, or make simplifying assumptions about the underlying system. In fact, it might even be impossible to reach a consensus about the reliability of a model. One might recall for example the recent controversy related to the use of Gaussian copula to model default "correlations" when pricing credit derivatives (see [25], [20], and [32]). Indeed, history has taught us that we often regret the choice of stochastic model that supported the decisions that were made. While a

subjective estimation of probability is easily biased (see [38, 12]), there are also obvious limitations to the reliability of models that are resolved using historical data. Yet, the strength of stochastic programming relies on its aptitude to identify solutions that handle or adapt best to the set of potential outcomes, relative to their probability of occurring. It is therefore not surprising that if an inaccurate stochastic model is used, we might regret having relied on a particular stochastic program once the decision that we applied is evaluated against a more legitimate model.

In this paper, we show how one can protect himself against the risk of committing to a stochastic model that might reveal to be inaccurate in the long run. In fact, we argue that for many instances of two-stage mixed-integer stochastic linear programs, it is actually wasteful to compose a detailed description of the underlying stochastics of the problem. For instance, when uncertainty is limited to the objective function of the problem, we show that the solution obtained by simply replacing the stochastic parameters with their expected values, also referred to as the mean value problem (MVP) solution, provides already a robust alternative. Given such a candidate solution, we will also show how to quantify the gains that could be achieved by investing more efforts in the resolution (or confirmation) of the stochastic model prior to taking the decision. Hence, this should allow practitioners to measure whether there is any value in dedicating more resources to the modeling of uncertainty.

The question of whether a decision model captures reality with enough accuracy to draw some conclusions has been haunting decision makers for many years. Most textbooks on decision analysis (see for example [31, 14]) recommend to always perform sensitivity (or post-optimality) analysis to identify the need for extra development or validation of a decision model. Tools for such analysis are especially abundant in the case of linear programs (c.f., [6]). For example, one can easily identify for the MVP model the range of values that a cost coefficient can take without affecting the optimal solution. One can therefore use this information to evaluate the importance of considering this coefficient as random. Unfortunately, in the case of two-stage problems, [39] observes that a naive sensitivity analysis of the MVP problem will disregard the temporal relation between decisions and parameters of the problem. Such an analysis is unable to take into account the fact that only the sensitivity of first stage decisions is relevant, moreover it fails to provide any guidance for identifying the decisions that adapt well to the potential outcomes. To address stochastic programs, a better strategy consists of analyzing the sensitivity of the first stage solution to changes in the stochastic model. Yet, currently available methods can only be applied when the true stochastic model is known to lie in a particular parametrized space. This is obviously not the case when the form of the distributions involved in the model is subject to disagreements.

Our work is related to the concept of value of stochastic solution (VSS), which was first introduced by [7] to measure the potential benefit from solving the stochastic program over solving a deterministic program. In contexts where the stochastic model is well defined, as shown in [8] and [21], bounding this value can provide arguments for investing the necessary computational resources toward finding the optimal solution of a stochastic program. Indeed, solving a stochastic program with mixed-integer variables has remained to this day a real

computational challenge (see [27] for a survey of available methods). Unfortunately, when a decision maker is hesitant about which probabilistic model to use, there is no guarantees that the *VSS* achieved under an assumed stochastic model will be representative of the gains achieved by deriving a more accurate model and solving it. We argue that the sum of both gains (i.e., from accurate modeling and solving) should be properly estimated before even starting to develop a stochastic programming. One might say that together these gains compose the value of stochastic modeling (*VSM*).

Our work also follows in spirit the efforts of [3] who showed that the "stochasticity gap" (i.e., the *VSM*) for a robust solution is at most equal to the optimal expected cost if the uncertainty is limited to the right-hand side of the constraints and the stochastic program satisfies a set of strong "positivity" conditions. In this paper, we consider uncertainty in the cost coefficient and impose no condition on the underlying model. Our bounds for *VSM*, which are evaluated numerically, scale with the level of uncertainty in the parameters, thus have the potential of taking on a lower value than the optimal expected cost. In particular, the work of [3] cannot shed any light on the stochasticity gap associated to the fleet mix optimization problem studied in Section 5.

This paper is organized as follows. In Section 2, we introduce the the two-stage stochastic program with cost uncertainty and its associated mean value problem. Section 3 identifies conditions under which the MVP solution can be considered a robust solution to implement with respect to the available knowledge of the distribution of uncertain parameters. We then turn ourselves in Section 4 to the problem of evaluating the *VSM*. After demonstrating that an exact evaluation is typically out of reach, we propose new methods for bounding its value. Section 5 discusses implication of these results for a stochastic fleet mix optimization problem. Finally, numerical results for three instances of this problem are presented in Section 6. In our three cases, bounding the *VSM* reveals that, once the MVP solution is obtained, there is little added value from investing more efforts in the development of a stochastic model.

2 Stochastic Programming with Cost Uncertainty

We are interested in the following stochastic linear program:

$$\begin{array}{lll} \text{(SLP)} & \underset{\boldsymbol{x}}{\text{minimize}} & \boldsymbol{c}_1^{\mathsf{T}}\boldsymbol{x} + \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})] \\ & \text{subject to} & \boldsymbol{A}_1\boldsymbol{x} \leq \boldsymbol{b}_1 \\ & & \boldsymbol{x} \in \mathbb{R}^{n_1-p_1} \times \mathbb{Z}^{p_1}\ , \end{array}$$

where $\boldsymbol{A}_1 \in \mathbb{R}^{m_1 \times n_1}$, and $\boldsymbol{b}_1 \in \mathbb{R}^{m_1}$. The vector $\boldsymbol{x} \in \mathbb{R}^{n_1}$ refers to a set of decisions that must be made prior to the realization of $\boldsymbol{\xi}$, a random vector in some probability space $(\mathbb{R}^d, \mathcal{B}, F)$ with $\mathcal{B}$ the Borel σ-algebra on $\mathbb{R}^d$. The function $h(\boldsymbol{x},\boldsymbol{\xi})$ is the cost incurred in a second-stage

once $\boldsymbol{\xi}$ is revealed and a recourse action $\boldsymbol{y} \in \mathbb{R}^{n_2}$ is taken. Formally,

$$\begin{array}{rlll} h(\boldsymbol{x}, \boldsymbol{\xi}) := & \underset{\boldsymbol{y}}{\text{minimize}} & \boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{C}_2 \boldsymbol{y} \\ & \text{subject to} & \boldsymbol{A}_2 \boldsymbol{x} + \boldsymbol{B}_2 \boldsymbol{y} \leq \boldsymbol{b}_2 \\ & & \boldsymbol{y} \in \mathbb{R}^{n_2 - p_2} \times \mathbb{Z}^{p_2} \ . \end{array}$$

Note that while the feasibility region for $\boldsymbol{y}$ is defined with certainty through $\boldsymbol{A}_2 \in \mathbb{R}^{m_2 \times n_1}$, $\boldsymbol{B}_2 \in \mathbb{R}^{m_2 \times n_2}$, and $\boldsymbol{b}_2 \in \mathbb{R}^{m_2}$, it is the cost incurred for choosing the recourse $\boldsymbol{y}$, measured through $\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{C}_2 \boldsymbol{y}$ where $\boldsymbol{C}_2 \in \mathbb{R}^{d \times n_2}$, that is unknown initially.

When one replaces the random vector of parameters in the SLP problem by its expected value, the problem reduces to what is known as the mean value problem :

$$\begin{array}{lrll} \text{(MVP)} & \underset{\boldsymbol{x}, \boldsymbol{y}}{\text{minimize}} & \boldsymbol{c}_1^{\mathsf{T}} \boldsymbol{x} + \boldsymbol{\mu}^{\mathsf{T}} \boldsymbol{C}_2 \boldsymbol{y} \\ & \text{subject to} & \boldsymbol{A}_1 \boldsymbol{x} \leq \boldsymbol{b}_1 \\ & & \boldsymbol{A}_2 \boldsymbol{x} + \boldsymbol{B}_2 \boldsymbol{y} \leq \boldsymbol{b}_2 \\ & & \boldsymbol{x} \in \boldsymbol{x} \in \mathbb{R}^{n_1 - p_1} \times \mathbb{Z}^{p_1} \\ & & \boldsymbol{y} \in \mathbb{R}^{n_2 - p_2} \times \mathbb{Z}^{p_2} \ , \end{array}$$

where $\boldsymbol{\mu} = \mathbb{E}_F[\boldsymbol{\xi}]$. To simplify our discussion, in what follows we refer to the feasible set for $\boldsymbol{x}$ as $\mathcal{X}$ and to the feasible set for $\boldsymbol{y}$ as $\mathcal{Y}(\boldsymbol{x})$. Specifically, $\mathcal{X} := \{\boldsymbol{x} \in \mathbb{R}^{n_1 - p_1} \times \mathbb{Z}^{p_1} \mid \boldsymbol{A}_1 \boldsymbol{x} \leq \boldsymbol{b}_1\}$, and $\mathcal{Y}(x) := \{\boldsymbol{y} \in \mathbb{R}^{n_2 - p_2} \times \mathbb{Z}^{p_2} \mid \boldsymbol{A}_2 \boldsymbol{x} + \boldsymbol{B}_2 \boldsymbol{y} \leq \boldsymbol{b}_2\}$.

Although not always explicitly stated, it has been common practice to rely on the MVP problem as a heuristic to get solutions to stochastic programs mainly for computational reasons. Indeed, while it is well-known that multi-stage stochastic program are generally intractable (see [35]), the difficulty also arise in two-stage stochastic linear programming problem when they involve integer decision variables (see [18] for instance). Recently, [28] even argued that since the realistic SLP problems are almost always intractable, one should instead try to identify, for a given class of models, whether there are properties of the MVP solution that can be exploited. We will soon suggest good reasons for implementing the MVP solution instead of the solution of the SLP when there is ambiguity in the choice of a distribution.

3 Robustness of the Mean Value Problem's Solution

As argued in Section 1, in practice it is often the case that a decision maker cannot fully resolve the distribution involved in the stochastic program at the time of making his decision. Instead, he might only have collected partial information about this distribution: *e.g.*, information that allows him to locate some of its moments. In what follows, we assume that the decision maker has gathered sufficient information about the random vector $\boldsymbol{\xi}$ to identify its support, locate its expected value, and has gathered some information about other moments of the form $\mathbb{E}_F[\psi(\boldsymbol{\xi})]$ for some convex mapping $\psi(\cdot)$. While we initially

demonstrate that the MVP solution is an optimal robust solution to the SLP problem when one only imposes upper bounds on these other moments, we also provide conditions under which this solution remains an optimal robust choice even when the other moments, such as the covariance matrix, are exactly known. This appears to indicate that while the MVP is a robust alternative, it might not be such a conservative one since it remains optimally robust even when a large amount of information about the distribution of $\boldsymbol{\xi}$ is known.

Remark 3.1. *Note that it is only for the sake of the clarity of exposition that we assume throughout the paper that the expected value of $\boldsymbol{\xi}$ is known exactly. If one only knows that $\mathbb{E}_F[\boldsymbol{\xi}] \in \mathcal{U}$, then one could use similar arguments as those that will be presented shortly to argue that it is the solution to the Robust MVP*

$$\begin{array}{ll} \underset{\boldsymbol{x},\boldsymbol{y}}{\text{minimize}} & \boldsymbol{c}_1^{\mathsf{T}}\boldsymbol{x} + \sup\limits_{\boldsymbol{\mu}\in\mathcal{U}} \ \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{C}_2\boldsymbol{y} \\ \text{subject to} & \boldsymbol{A}_1\boldsymbol{x} \leq \boldsymbol{b}_1 \\ & \boldsymbol{A}_2\boldsymbol{x} + \boldsymbol{B}_2\boldsymbol{y} \leq \boldsymbol{b}_2 \\ & \boldsymbol{x} \in \boldsymbol{x} \in \mathbb{R}^{n_1-p_1} \times \mathbb{Z}^{p_1} \\ & \boldsymbol{y} \in \mathbb{R}^{n_2-p_2} \times \mathbb{Z}^{p_2} \ , \end{array}$$

that is an optimal robust solution to the SLP problem.

3.1 The Case of Known Mean and Bounded Moments

Let us consider the family of distributional sets that can be represented by the form

$$\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi) = \left\{ F \in \mathcal{M} \,\middle|\, \begin{array}{l} \mathbb{P}_F(\boldsymbol{\xi} \in \mathcal{S}) = 1 \\ \mathbb{E}_F[\boldsymbol{\xi}] = \boldsymbol{\mu} \\ \mathbb{E}_F[\psi(\boldsymbol{\xi})] \leq 0 \ , \ \forall\, \psi \in \Psi \end{array} \right\}$$

where $\boldsymbol{\mu} \in \mathcal{S}$, $\mathcal{S} \subseteq \mathbb{R}^d$, $\mathcal{M}$ is the set of all probability measures on the measurable space $(\mathbb{R}^d, \mathcal{B})$, with $\mathcal{B}$ the Borel σ-algebra on $\mathbb{R}^d$, and Ψ is a set of convex mappings from $\mathbb{R}^m$ to $\mathbb{R}$. Intuitively, the constraints of the type $\mathbb{E}_F[\psi(\boldsymbol{\xi})] \leq 0$ reflects the fact that the probability measure achieves below a certain level of dispersion around $\boldsymbol{\mu}$. Indeed, as shown in Appendix A, when $\mathcal{S}$ is a convex set, one can verify that if a random vector $\boldsymbol{\xi}$ has a distribution that lies in $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$ with, then for any $\alpha \leq 1$ the distribution of $\zeta = \alpha(\boldsymbol{\xi} - \boldsymbol{\mu}) + \boldsymbol{\mu}$ also lies in this set. Note that this property of the distributional set inherently prevents one from capturing complete information about a distribution since its structure prevents it from discarding distributions that are "less dispersed". This will be addressed in section 3.3.

Overall, we believe the set $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$ captures well the type of information one would have in hand in early steps of uncertainty assessment: after identifying what type of realization might occur, a decision maker will normally try to bound how far most realization might be from the mean. The later can be done using historical samples by estimating the upper bound b of a confidence interval for the moment of a convex function $\phi(\cdot)$, and considering that $\psi(\boldsymbol{z}) := \phi(\boldsymbol{z}) - b$. In the following examples, we describe how knowledge about the covariance matrix, or about semi-variance, can be used to construct such a set.

Example 3.1. *Due to estimation errors that corrupt the evaluation of moments of high order, knowledge of a distribution is often limited to mean and covariance information. In [15], the authors suggest using first and second order moment information to construct the following distributional set :*

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

where the matrix $\boldsymbol{\Sigma}$ acts as an upper bound on the covariance matrix of F through a linear matrix inequality. The authors motivate the uncertainty region for the second order moment matrix by constructing it based on historical data in a way that guarantees with high probability that it contains the true values of these moments. The set $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is an example of $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$ where

$$\Psi := \{\psi : \mathbb{R}^d \to \mathbb{R} \mid \exists \boldsymbol{z} \in \mathbb{R}^d,\, \psi(\boldsymbol{\xi}) = (\boldsymbol{z}^\mathsf{T}(\boldsymbol{\xi} - \boldsymbol{\mu}))^2 - \boldsymbol{z}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{z}\}\,.$$

Furthermore, we can verify that Ψ is indeed a set of convex mappings.

Example 3.2. *Although Ψ cannot represent a bound on skewness, it is still possible to bound statistics that can capture the asymmetry of the distribution. For instance,*

$$\psi(\boldsymbol{\xi}; \boldsymbol{z}) := \max(0, \boldsymbol{z}^T(\boldsymbol{\xi} - \boldsymbol{\mu}))^2 - b(z)$$

bounds the maximum semi-variance in the $\boldsymbol{z}$ direction since we have that

$$\mathbb{E}_F[\psi(\boldsymbol{\xi}; \boldsymbol{z})] \leq 0 \quad \Rightarrow \quad \mathbb{E}_F[(\max(0, \boldsymbol{z}^T(\boldsymbol{\xi} - \boldsymbol{\mu}))^2] \leq b(z).$$

Remark 3.2. *Note that while we do not impose that $\mathcal{S}$ be a convex set, if it is the case then $\boldsymbol{\mu} \in \mathcal{S}$ happens naturally. The assumption that $\boldsymbol{\mu} \in \mathcal{S}$ becomes more limiting if $\mathcal{S}$ is a list of discrete scenarios since it requires that $\boldsymbol{\mu}$ be part of this list. The assumption is for instance not satisfied if $\boldsymbol{\xi}$ is a vector of Bernoulli variables for instance; however, if one has not resolved yet whether a random vector is discrete or continuous, then the assumption that $\boldsymbol{\mu}$ is a potential realization should be natural to make.*

Our first result indicates that the MVP solution is a robust solution when the distribution information can be described through the form $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$.

Proposition 3.1. *The solution to the MVP problem is optimal with respect to the distributionally robust problem*

$$\underset{\boldsymbol{x}}{\text{minimize}} \qquad \sup_{F \in \mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)} \; \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} + \mathbb{E}_F[h(\boldsymbol{x}, \boldsymbol{\xi})]. \tag{1}$$

Distributionally robust optimization, also referred as minimax stochastic programming, was first introduce in [33]. Since then, it has attracted much attention and especially recently due to the emergence of more effective resolution methods (see [15] and references therein).

The motivation behind this solution concept resides in the observation that decision makers are typically ambiguity averse (see Ellsberg paradox in [19]) in the sense that they prefer to be exposed to known risks then to unknown ones. Proposition 3.1 states that the MVP solution achieves the best guarantees on the magnitude of expected cost given the existing distribution ambiguity.

Proof: This proposition is a consequence of the fact that $h(\boldsymbol{x}, \boldsymbol{\xi})$ is a concave function of $\boldsymbol{\xi}$ (see proof in Appendix B). Thus, by Jensen's inequality we know that for any distribution F, it is the case that $\mathbb{E}_F[h(\boldsymbol{x}, \boldsymbol{\xi})] \leq h(\boldsymbol{x}, \mathbb{E}_F[\boldsymbol{\xi}])$, and that equality is achieved if the distribution of $\boldsymbol{\xi}$ is the Dirac measure[1] $\delta_{\mathbb{E}_F[\boldsymbol{\xi}]}$. If we can show that $\delta_{\boldsymbol{\mu}}$ is a feasible probability measure, then it is necessarily an optimal worst-case distribution since

$$\mathbb{E}_{\delta_{\boldsymbol{\mu}}}[h(\boldsymbol{x}, \boldsymbol{\xi})] = h(\boldsymbol{x}, \boldsymbol{\mu}) = h(\boldsymbol{x}, \mathbb{E}_F[\boldsymbol{\xi}]) \geq \mathbb{E}_F[h(\boldsymbol{x}, \boldsymbol{\xi})] \,, \ \forall\, F \in \mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi) \,.$$

The fact that $\delta_{\boldsymbol{\mu}}$ is a feasible probability measure is guaranteed by the existence of a feasible measure $F_0 \in \mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$, which we can assume exists since otherwise the proposition is true for any $\boldsymbol{x}$. First, it is easy to verify that $\mathbb{E}_{\delta_{\boldsymbol{\mu}}}[\boldsymbol{\xi}] = \boldsymbol{\mu}$ and that $\mathbb{P}_{\delta_{\boldsymbol{\mu}}}(\boldsymbol{\xi} \in \mathcal{S}) = 1$ since it was assumed that $\boldsymbol{\mu} \in \mathcal{S}$. By Jensen's inequality, it is also the case that for all $\psi(\cdot) \in \Psi$

$$\mathbb{E}_{\delta_{\boldsymbol{\mu}}}[\psi(\boldsymbol{\xi})] = \psi(\boldsymbol{\mu}) = \psi(\mathbb{E}_{F_0}[\boldsymbol{\xi}]) \leq \mathbb{E}_{F_0}[\psi(\boldsymbol{\xi})] \leq 0 \,,$$

since $\psi(\cdot)$ is a convex mapping. This completes our proof. □

Since the solution to robust problem (1) is insensitive to the members of the set Ψ, Proposition 3.1 provides theoretical arguments to support the common belief that error in estimating the mean of a distribution is often more dramatic than other types of distributional misspecification (see for instance [13] for numerical evidence in a portfolio selection problem). Intuitively, in the context of an SLP, this is due to the fact that the Dirac measure $\delta_{\boldsymbol{\mu}}$ (as defined in Endnote 1) becomes a worst-case distribution when one does not observe (or impose) a minimum amount of dispersion around $\boldsymbol{\mu}$.Having this in mind, section 3.3 will actually show conditions under which the Dirac measure $\delta_{\boldsymbol{\mu}}$ remains a close approximation of the worst-case distribution even when a certain level of dispersion needs to be met.

Although the focus of this paper is two-stage stochastic programming, it might not come as a surprise that Proposition 3.1 can be extended to multi-stage problems (see Appendix C for the proof).

Proposition 3.2. *Consider the multi-stage stochastic programming problem*

$$\underset{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T}{\text{minimize}} \quad \mathbb{E}_F\left[\sum_{t=1}^{T} \xi_t^{\mathsf{T}} C_t \boldsymbol{x}_t(\xi_{[1:t]})\right] \tag{2a}$$

$$\text{subject to} \quad \big(\boldsymbol{x}_1(\xi_1), \boldsymbol{x}_2(\xi_{[1:2]}), \ldots, \boldsymbol{x}_T(\xi_{[1:T]})\big) \in \mathcal{X}_{[1:T]} \tag{2b}$$

where $\xi_t \in \mathbb{R}^{m_t}$ is the vector of random parameters associated the t-th stage, $\boldsymbol{x}_t : \prod_{t=1}^{T} \mathbb{R}^{m_t} \to \mathbb{R}^{n_t}$ is a decision vector that can adapt to the random parameters observed in the t-th first

[1]Recall that the Dirac measure $\delta_{\boldsymbol{a}}$ is the measure of mass one at the point $\boldsymbol{a}$.

stages, and $\mathcal{X}_{[1:T]}$ is a convex set of feasible sequence of decision values. The solution to the mean value problem

$$\underset{\bar{\boldsymbol{x}}_1,\bar{\boldsymbol{x}}_2,...,\bar{\boldsymbol{x}}_T}{\text{minimize}} \quad \sum_{t=1}^{T} \boldsymbol{\mu}_t^{\mathsf{T}} C_t \bar{\boldsymbol{x}}_t \tag{3a}$$

$$\text{subject to} \quad (\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, ..., \bar{\boldsymbol{x}}_T) \in \mathcal{X}_{[1:T]} \tag{3b}$$

is optimal with respect to the distributionally robust version of this multi-stage stochastic programming problem that accounts for uncertainty in the distribution of $\zeta := [\xi_1, \xi_2, ..., \xi_T]$ through the distributional set $\mathcal{D}(\mathcal{S}, [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_T], \Psi)$.

Remark 3.3. *The idea of approximating a stochastic program using bounds that are based on Jensen's inequality and its generalizations has received considerable attention in the stochastic programming community. In Chapter 8 of [9], the authors describe how one may circumvent the difficulty of integrating the recourse function over a continuous outcome space by partitioning this space and replacing $\mathbb{E}_F[h(\boldsymbol{x}, \boldsymbol{\xi})] = \sum_i \mathbb{P}_F(\boldsymbol{\xi} \in \mathcal{S}_i)\mathbb{E}_F[h(\boldsymbol{x}, \boldsymbol{\xi})|\boldsymbol{\xi} \in \mathcal{S}_i]$ with a tractable lower bound $\sum_i \mathbb{P}_F(\boldsymbol{\xi} \in \mathcal{S}_i)h(\boldsymbol{x}, \mathbb{E}_F[\boldsymbol{\xi}|\boldsymbol{\xi} \in \mathcal{S}_i])$ given that $h(\boldsymbol{x}, \boldsymbol{\xi})$ is convex in $\boldsymbol{\xi}$. While this bound is obtained using a straightforward application of Jensen's inequality, more sophisticated bounds have been proposed that are based on the solution of moment problems (see [17] for a review). These are especially useful when solving the stochastic program using a decomposition scheme such as the L-shaped method. Although the idea that the MVP problem can provide a bound for a stochastic program is obviously not new, the idea that its solution can be distributionally robust appears to have been unexplored to this date for cases where distribution information includes more than just mean and covariance information.*

3.2 The Case of Known Support, Mean and Covariance Matrix

We start by paying special attention to the case of known support, mean, and covariance matrix (i.e. the case studied in Example 3.1) for two reasons. It is definitively a case that is interesting in its own right given that it has recently attracted a lot of attention in the literature on distributionally robust optimization (see [11, 2, 15]). Furthermore, the ideas that are used to prove the result will become valuable for deriving the similar results for the more general case. In particular, we assume that the information that is available takes the shape of the following distributional set

$$\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left\{ F \in \mathcal{M} \;\middle|\; \begin{array}{l} \mathbb{P}_F(\boldsymbol{\xi} \in \mathcal{S}) = 1 \\ \mathbb{E}_F[\boldsymbol{\xi}] = \boldsymbol{\mu} \\ \mathbb{E}_F[(\boldsymbol{z}^T(\boldsymbol{\xi} - \boldsymbol{\mu}))^2 - \boldsymbol{z}^T\boldsymbol{\Sigma}\boldsymbol{z}] = 0 \;,\; \forall\, \boldsymbol{z} \in \mathbb{R}^d \end{array} \right\},$$

where the last constraint is equivalent to $\mathbb{E}_F[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$ since

$$\forall \boldsymbol{z} \in \mathbb{R}^d,\; \boldsymbol{z}^{\mathsf{T}}\mathbb{E}_F[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^T]\boldsymbol{z} = \boldsymbol{z}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{z} \;\;\Leftrightarrow\;\; \mathbb{E}_F[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}\;,$$

and where again we assume that $\boldsymbol{\mu} \in \mathcal{S}$. Note how $\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ now imposes that the moments of functions indexed by $\boldsymbol{z} \in \mathbb{R}^d$ be exactly equal to 0 instead of smaller or equal.

Unfortunately, one can easily show based on [5] that problem (1) with $\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ contains instances that are NP-hard to solve.[2] Since we cannot expect to guarantee that the MVP solution is robust for those instances without claiming that the MVP problem can be used to solve the hardest problems in the NP-complete class, we will instead focus on a perturbed version of this distributional set: i.e.,

$$\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon) := \left\{ F \in \mathcal{M} \left| \begin{array}{l} \mathbb{P}_F(\boldsymbol{\xi} \in \mathcal{S}) \geq 1 - \epsilon \\ \mathbb{E}_F[\boldsymbol{\xi}] = \boldsymbol{\mu} \\ \mathbb{E}_F[\psi(\boldsymbol{\xi})] = 0 \ , \ \forall \, \psi \in \Psi \end{array} \right. \right\} ,$$

for some arbitrarily small value $\epsilon > 0$. Informally, we can pretend for all practical purposes that this perturbed set is equivalent to $\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ since it is usually impossible to have complete confidence about the support (unless $\mathcal{S} = \mathbb{R}^d$).

Assumption 3.3. *There exists a radius R such that, for all $\boldsymbol{x} \in \mathcal{X}$, the feasible region $\mathcal{Y}(\boldsymbol{x})$ is contained inside a ball of this radius centered at zero.*

This assumption imposes that the set of feasible recourse actions be bounded. Although this might appear restrictive, in practice both first stage and recourse variable are usually non-negative and their respective magnitude is typically bounded due to physical limitations: for example, resource availability will limit the total number of items that can be produced, credit risk will limit the size of the loan a firm can obtain from a bank, etc. From a technical point of view, this assumption is needed to guarantee that $h(\boldsymbol{x}, \boldsymbol{\xi})$ is Lipschitz continuous in $\boldsymbol{\xi}$.

Lemma 3.4. *Given that $\mathcal{Y}(\boldsymbol{x})$ satisfies Assumption 3.3, the function $h(\boldsymbol{x}, \cdot)$ is Lipschitz continuous with constant $R\|\boldsymbol{C}_2\|$, where $\|\boldsymbol{C}_2\|$ is the spectral norm of $\boldsymbol{C}_2$.*

Proof: For any two vectors $\boldsymbol{z}_1 \in \mathbb{R}^d$ and $\boldsymbol{z}_2 \in \mathbb{R}^d$, we see that

$$\begin{aligned} h(\boldsymbol{x}, \boldsymbol{z}_2) - h(\boldsymbol{x}, \boldsymbol{z}_1) &= \min_{\boldsymbol{y}_2 \in \mathcal{Y}(\boldsymbol{x})} \boldsymbol{y}_2^\intercal \boldsymbol{C}_2^\intercal \boldsymbol{z}_2 - \min_{\boldsymbol{y}_1 \in \mathcal{Y}(\boldsymbol{x})} \boldsymbol{y}_1^\intercal \boldsymbol{C}_2^\intercal \boldsymbol{z}_1 \\ &\leq \boldsymbol{y}_1^{*\intercal} \boldsymbol{C}_2 (\boldsymbol{z}_2 - \boldsymbol{z}_1) \leq \|\boldsymbol{y}_1^*\| \|\boldsymbol{C}_2\| \|\boldsymbol{z}_2 - \boldsymbol{z}_1\| \leq R \|\boldsymbol{C}_2\| \|\boldsymbol{z}_2 - \boldsymbol{z}_1\| \ , \end{aligned}$$

where $\boldsymbol{y}_1^* \in \operatorname{argmin}_{\boldsymbol{y}_1 \in \mathcal{Y}(\boldsymbol{x})} \boldsymbol{y}_1^\intercal \boldsymbol{C}_2^\intercal \boldsymbol{z}_1$. Similarly, we have that

$$h(\boldsymbol{x}, \boldsymbol{z}_2) - h(\boldsymbol{x}, \boldsymbol{z}_1) \geq \boldsymbol{y}_2^{*\intercal} \boldsymbol{C}_2 (\boldsymbol{z}_2 - \boldsymbol{z}_1) \geq -\|\boldsymbol{y}_2^*\| \|\boldsymbol{C}_2\| \|\boldsymbol{z}_2 - \boldsymbol{z}_1\| \geq -R \|\boldsymbol{C}_2\| \|\boldsymbol{z}_2 - \boldsymbol{z}_1\| \ . \ \square$$

We follow with a proposition which establishes that the Dirac measure $\delta_{\boldsymbol{\mu}}$ continues to give an accurate estimate of the worst-case $\mathbb{E}_F[h(\boldsymbol{x}, \boldsymbol{\xi})]$ given the available information about the distribution, hence implying that the MVP solution remains an optimal robust solution. This might come as a surprise since, for any $\epsilon > 0$, the conditions imposed by the set $\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)$ actually reject this probability measure from the set of plausible ones for $\boldsymbol{\xi}$.

[2] It is worth mentioning that based on Proposition 3.1, the MVP solution is a safe approximate solution of problem (1) with $\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ yet we seek conditions under which the MVP solution can be considered robust optimal.

Proposition 3.5. *Given that $\mathcal{Y}(\boldsymbol{x})$ satisfies Assumption 3.3, we have that for any $\mathcal{S}$, any $\boldsymbol{\mu} \in \mathcal{S}$, and any $\boldsymbol{\Sigma} \succeq 0$ and any $\epsilon > 0$*

$$\sup_{F \in \bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)} \mathbb{E}_F[h(\boldsymbol{x}, \boldsymbol{\xi})] = h(\boldsymbol{x}, \boldsymbol{\mu}) \,,$$

hence, the solution of the MVP problem is optimal with respect to the distributionally robust problem (1) *under $\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)$.*

This result can be considered an extension of a result presented in [2], which noted that the MVP solution is a robust solution when only mean and covariance matrix information is known. Indeed, we now know that the MVP solution remains optimal even when the support of the distribution is "almost certainly" known.

The result presented here has even a broader significance on solving some general NP-hard problems. That is, by relaxing the constraint set a bit, one can turn an NP-Hard problem into a polynomial-time solvable problem. Most early NP-hard approximations algorithms focus on finding an absolutely feasible solution to achieve a sub-optimal solution. Here we find an absolutely optimal solution for a slightly perturbed difficult problem. This is quite helpful since, in practice, some constraints can be "soft", such as we argue is the case for many distributionally robust optimization models.

Proof: The proof hinges on using the sequence of distributions $F_1, F_2, ...$ with

$$F_k(\boldsymbol{\xi}) = (1 - \beta_k^{-1})\delta_{\boldsymbol{\mu}}(\boldsymbol{\xi}) + \beta_k^{-1} G_k(\boldsymbol{\xi}) \,,$$

for an increasing sequence of β_k, where G_k is random vector distributed according to the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\beta_k \boldsymbol{\Sigma}$. For k large enough, we will show that the sequence F_k lies in $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)$ and makes the expected value $\mathbb{E}_{F_k}[h(\boldsymbol{x}, \boldsymbol{\xi})]$ become arbitrarily close to $h(\boldsymbol{x}, \boldsymbol{\mu})$.

First, given that $\beta_k \geq 1/\epsilon$, one can confirm that $\mathbb{P}_{F_k}(\boldsymbol{\xi} \in \mathcal{S}) \geq (1 - \beta_k^{-1}) \geq 1 - \epsilon$. Secondly, by construction $\mathbb{E}_{F_k}[\boldsymbol{\xi}] = (1 - \beta_k^{-1})\boldsymbol{\mu} + \beta_k^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}$ for all k. Thirdly, for all $\boldsymbol{z} \in \mathbb{R}^d$, we have that

$$\mathbb{E}_{F_k}[(\boldsymbol{z}^T(\boldsymbol{\xi} - \boldsymbol{\mu}))^2 - \boldsymbol{z}^T \boldsymbol{\Sigma} \boldsymbol{z}] = (1 - \beta_k^{-1}) \cdot 0 + \beta_k^{-1} \cdot \beta_k \boldsymbol{z}^T \boldsymbol{\Sigma} \boldsymbol{z} - \boldsymbol{z}^T \boldsymbol{\Sigma} \boldsymbol{z} = 0 \,.$$

We therefore have that for k large enough, $F_k \in \mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)$.

Based on the Lipschitz property of $h(x, \cdot)$ demonstrated in Lemma 3.4, we can verify that

$$\begin{aligned}
\sup_{F \in \bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)} \mathbb{E}_F[h(\boldsymbol{x}, \boldsymbol{\xi})] &\geq \sup_{\{k | F_k \in \bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)\}} \mathbb{E}_{F_k}[h(\boldsymbol{x}, \boldsymbol{\xi})] \\
&= \sup_{\{k | F_k \in \bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)\}} (1 - \beta_k^{-1}) h(\boldsymbol{x}, \boldsymbol{\mu}) + \beta_k^{-1} \mathbb{E}_{G_k}[h(\boldsymbol{x}, \boldsymbol{\xi})] \\
&\geq \sup_{\{k | F_k \in \bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)\}} (1 - \beta_k^{-1}) h(\boldsymbol{x}, \boldsymbol{\mu}) + \beta_k^{-1} \mathbb{E}_{G_k}[h(\boldsymbol{x}, \boldsymbol{\mu}) - R\|\boldsymbol{C}_2\| \|\boldsymbol{\xi} - \boldsymbol{\mu}\|] \\
&= \sup_{\{k | F_k \in \bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)\}} h(\boldsymbol{x}, \boldsymbol{\mu}) - \beta_k^{-1} R\|\boldsymbol{C}_2\| \|\boldsymbol{\mu}\| - \beta_k^{-1} R\|\boldsymbol{C}_2\| \, O(\sqrt{\beta_k}) \\
&= h(\boldsymbol{x}, \boldsymbol{\mu}) \,,
\end{aligned}$$

where we used the fact that

$$\mathbb{E}_{G_k}[\|\boldsymbol{\xi}-\boldsymbol{\mu}\|]^2 \leq \mathbb{E}_{G_k}[\|\boldsymbol{\xi}-\boldsymbol{\mu}\|^2] = \sum_{i=1}^{m} \mathbb{E}_{G_k}[(\xi_i-\mu_i)^2] = \beta_k \, \mathbf{trace}(\boldsymbol{\Sigma}) \ ,$$

so that $\mathbb{E}_{G_k}[\|\boldsymbol{\xi}-\boldsymbol{\mu}\|] = O(\sqrt{\beta_k})$. We can then simply conclude that

$$h(\boldsymbol{x},\boldsymbol{\xi}) \leq \sup_{F \in \bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\boldsymbol{\Sigma},\epsilon)} \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})] \leq \sup_{F \in \mathcal{D}(\mathbb{R}^d,\boldsymbol{\mu},\boldsymbol{\Sigma})} \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})] \leq h(\boldsymbol{x},\boldsymbol{\mu}) \ . \ \square$$

3.3 The Case of Known Mean and Known Moments

Next, we show that there are conditions under which the conclusions of Proposition 3.1 and 3.5 also apply when the distributional set imposes that F satisfy exactly some prescribed moments for a range of convex functions. While we would like to define conditions under which the MVP solution is robust for a distributional set of the form

$$\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi) = \left\{ F \in \mathcal{M} \;\middle|\; \begin{array}{l} \mathbb{P}_F(\boldsymbol{\xi} \in \mathcal{S}) = 1 \\ \mathbb{E}_F[\boldsymbol{\xi}] = \boldsymbol{\mu} \\ \mathbb{E}_F[\psi(\boldsymbol{\xi})] = 0 \ , \ \forall \, \psi \in \Psi \end{array} \right\} \ ,$$

we can expect that the issues identified in section 3.2 might arise and therefore we will instead focus on a perturbed version of this distributional set: i.e.,

$$\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi,\epsilon) := \left\{ F \in \mathcal{M} \;\middle|\; \begin{array}{l} \mathbb{P}_F(\boldsymbol{\xi} \in \mathcal{S}) \geq 1-\epsilon \\ \|\mathbb{E}_F[\boldsymbol{\xi}] - \boldsymbol{\mu}\| \leq \epsilon \\ \mathbb{E}_F[\psi(\boldsymbol{\xi})] = 0 \ , \ \forall \, \psi \in \Psi \end{array} \right\} \ ,$$

for some arbitrarily small value $\epsilon > 0$. Here the relaxation is applied both to the support constraint and the constraint on the location of the mean which is now restricted to a sphere of arbitrarily small radius centered at $\boldsymbol{\mu}$. In practice, both of these relaxations can often be considered of little significance.

Assumption 3.6. *We assume that for all $\psi \in \Psi$, $\psi(\boldsymbol{\mu}) = -1$.*

This assumption can usually be satisfied by proper rescaling of the functions in Ψ if $\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi)$ is non-empty and $\psi(\boldsymbol{\mu}) \neq 0$ for all $\psi \in \Psi$, so that $\psi(\boldsymbol{\mu}) \leq \mathbb{E}_F[\psi(\boldsymbol{\xi})] = 0$ for any feasible F and $\psi'(\boldsymbol{z}) := -\psi(\boldsymbol{z})/\psi(\boldsymbol{\mu})$ can be used.

We follow with a proposition which indicates conditions under which it the Dirac measure $\delta_{\boldsymbol{\mu}}$ gives an approximately accurate estimate of the worst-case $\mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})]$ given the available information about the distribution, hence implying that the MVP solution is a nearly-optimal robust solution. The proof of Proposition 3.7 follows the spirit of our proof for the case with known covariance matrix and is deferred to Appendix D.

Proposition 3.7. *Given that $\mathcal{Y}(\boldsymbol{x})$ and Ψ satisfy assumptions 3.3 and 3.6 respectively, we have that $\sup_{F \in \bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi,\epsilon)} \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})] \in [h(\boldsymbol{x},\boldsymbol{\mu}), h(\boldsymbol{x},\boldsymbol{\mu}) + O(\epsilon)]$ for all $\epsilon > 0$ as long as the set Ψ satisfies the following conditions:*

1. *For an increasing sequence of* β_k *with* $\lim_{k\to\infty}\beta_k = \infty$*, there exists a series of distribution* G_k *such that for all* $\psi \in \Psi$*,* $\mathbb{E}_{G_k}[\psi(\boldsymbol{\xi})] = \beta_k$ *.*

2. *There is a function* ψ_0 *in the conical hull of* Ψ *that satisfies*

$$\psi_0(\boldsymbol{z}) \geq a\|\boldsymbol{z}\|^\gamma - b \;\; \forall\, \boldsymbol{z} \in \mathbb{R}^d \,,$$

for some $b \in \mathbb{R}$*,* $a > 0$*, and* $\gamma > 1$

Satisfying these conditions thus ensures that the solution to the mean value problem is $O(\epsilon)$*-suboptimal with respect to the distributionally robust problem* (3.1) *with* $\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, \Psi, \epsilon)$*.*

In simple words, the two conditions summarize the properties that were present in the case of known covariance matrix. Condition 1 requires to determine, for each of a list of increasing level of dispersion β_k, a distribution for which the dispersion level perceived by each moment measure coincides at β_k. This was the case for instance with the constructed series of multivariate normal distributions. On the other hand, Condition 2 will ensure that the tail of the distributions in this increasingly dispersed sequence only grows sublinearly in β_k. Specifically, that

$$\mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|^\gamma] \leq \mathbb{E}_{G_k}[(\psi_0(\boldsymbol{\xi}) + b)/a] = \beta_k b/a \;\; \Rightarrow \;\; \mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|] \leq \mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|^\gamma]^{1/\gamma} = O(\beta_k^{1/\gamma})$$
$$\Rightarrow \;\; \mathbb{P}_{G_k}(\|\boldsymbol{\xi}\| \geq \alpha) \leq \frac{\mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|]}{\mathbb{E}_{G_k}[\|\boldsymbol{\xi}\| \mid \|\boldsymbol{\xi}\| \geq \alpha]} \leq \frac{\mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|]}{\alpha} = O(\beta_k^{1/\gamma}/\alpha) \,.$$

Once again, this happened naturally with the multivariate normal distribution for which $\mathbb{P}_{G_k}(\|\boldsymbol{\xi} - \boldsymbol{\mu}\| \geq \alpha) = O(\beta_k^{1/2}/\alpha)$ for any k.

While Proposition 3.7 should help identify I wide range of situations where the MVP solution is nearly robust, we next attempt to make the implications of this proposition more practical. In particular, we identify a certain family of distribution sets, which impose norm-based moment constraints, that satisfy the two given conditions.

Proposition 3.8. *Given that* $\mathcal{Y}(\boldsymbol{x})$ *satisfies Assumption 3.3, we have that for any* $\epsilon > 0$*, the solution to the mean value problem is* $O(\epsilon)$*-suboptimal with respect to the distributionally robust problem* (3.1) *with* $\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, b, \gamma, \epsilon)$*, where*

$$\bar{\mathcal{D}}(\mathcal{S}, \boldsymbol{\mu}, G_0, \gamma, \epsilon) = \left\{ F \in \mathcal{M} \;\middle|\; \begin{array}{l} \mathbb{P}_F(\boldsymbol{\xi} \in \mathcal{S}) \geq 1 - \epsilon \\ \|\mathbb{E}_F[\boldsymbol{\xi}] - \boldsymbol{\mu}\| \leq \epsilon \\ \mathbb{E}_F[\|\boldsymbol{A}(\boldsymbol{\xi} - \boldsymbol{\mu})\|_\alpha^\gamma] = b(\boldsymbol{A}, \alpha) \;,\; \forall \alpha \geq 1,\, \forall\, \boldsymbol{A} \in \mathbb{R}^{d\times d} \end{array} \right\} \,,$$

where $b : \mathbb{R}^{d\times d} \times \mathbb{R}^+ \to \mathbb{R}$ *carries the moment information for each* $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ *and* $\alpha \geq 1$*, for some* $\boldsymbol{\mu} \in \mathcal{S}$*, some* $\mathcal{S} \subseteq \mathbb{R}^d$*, and some* $\gamma > 1$*.*

Note that in order to ensure that the moment problem is feasible, one might construct the moment information based on a reference distribution G_0 such that $b(\boldsymbol{A}, \alpha) := \mathbb{E}_{G_0}[\|\boldsymbol{A}(\boldsymbol{\xi} - \boldsymbol{\mu})\|_\alpha^\gamma]$. In practice, one could easily use an empirical distribution based on historical data. Note also that when $\gamma = 2$, this result reinforces the robustness of the MVP solution with

respect to the set of distribution $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon)$ since in that case $\alpha = 2$ and $\boldsymbol{A} \in \mathbb{R}^{1\times d}$. This signifies for instance that while adding moment information of the type $\mathbb{E}\left[(\max_i |\xi_i - \mu_i|)^2\right]$ to the set of distribution $\mathcal{D}(\mathbb{R}^{d\times d}, \boldsymbol{0}, \boldsymbol{I}, \epsilon)$ might successfully differentiate the standard multivariate normal distribution from a uniform distribution over the box $[-1/\sqrt{12}, 1/\sqrt{12}]^d$, it still somehow does not allow to prevent the Dirac measure from being an approximate worst-case distribution.

Before proving this result, we need to introduce the following lemma which proof is deferred to Appendix E.

Lemma 3.9. *Given any $\gamma \geq 1$ and any $\alpha \geq 1$, we have that*

$$\|\boldsymbol{z} - \boldsymbol{\mu}\|_\alpha^\gamma \geq \frac{1}{2^{\gamma-1}}\|\boldsymbol{z}\|_\alpha^\gamma - \|\boldsymbol{\mu}\|_\alpha^\gamma \ , \ \forall \ \boldsymbol{z} \in \mathbb{R}^d \ .$$

Proof of Proposition 3.8: Without loss of generality, we assume that $b(\boldsymbol{A}, \alpha) > 0$ for all $\boldsymbol{A}$ and α, otherwise the moment problem is infeasible and the conclusion follows trivially. We first identify the set Ψ as

$$\Psi := \{\psi : \mathbb{R}^d \to \mathbb{R} \mid \exists \alpha \geq 1, \boldsymbol{A} \in \mathbb{R}^{d\times d}, \psi(\boldsymbol{z}) = ((1/b(\boldsymbol{A}, \alpha))\|\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{\mu})\|_\alpha^\gamma - 1, \ \forall \, \boldsymbol{z} \in \mathbb{R}^d\} \, .$$

Each $\psi(\cdot)$ in Ψ is convex since $b(\boldsymbol{A}, \alpha) > 0$, the function y^γ is convex and increasing over the domain $y \geq 0$, and the function $\|\boldsymbol{z}\|_\alpha$ is convex and positive for all $\alpha \geq 1$. To satisfy Condition 1, one can identify any feasible distribution as G_0 and choose the sequence of G_k such that $G_k(\boldsymbol{\xi}) = G_0((\boldsymbol{\xi} - \boldsymbol{\mu})/(1 + \beta_k)^{1/\gamma} + \boldsymbol{\mu})$ so that

$$\begin{aligned} \mathbb{E}_{G_k}[\psi(\boldsymbol{\xi})] &= \mathbb{E}_{G_0}[\psi((1 + \beta_k)^{1/\gamma}(\boldsymbol{\xi} - \boldsymbol{\mu}) + \boldsymbol{\mu})] \\ &= (1/b(\boldsymbol{A}, \alpha))\mathbb{E}_{G_0}[\|\boldsymbol{A}((1 + \beta_k)^{1/\gamma}(\boldsymbol{\xi} - \boldsymbol{\mu}))\|_\alpha^\gamma] - 1 = (1 + \beta_k)(b(\boldsymbol{A}, \alpha)/b(\boldsymbol{A}, \alpha)) - 1 = \beta_k \ . \end{aligned}$$

To satisfy Condition 2, one can choose $\boldsymbol{A}$ to be the identity matrix $\boldsymbol{I}$ and any $\alpha \geq 1$ and confirm that

$$\begin{aligned} \psi_0(\boldsymbol{z}) &= (1/b(\boldsymbol{I}, \alpha))\|\boldsymbol{z} - \boldsymbol{\mu}\|_\alpha^\gamma - 1 \geq (1/b(\boldsymbol{I}, \alpha))((1/2^{\gamma-1})\|\boldsymbol{z}\|_\alpha^\gamma - \|\boldsymbol{\mu}\|_\alpha^\gamma) - 1 \\ &\geq (1/(b(\boldsymbol{I}, \alpha)d^{\gamma/2}2^{\gamma-1}))\|\boldsymbol{z}\|_2^\gamma - (1/b(\boldsymbol{I}, \alpha))\|\boldsymbol{\mu}\|_\alpha^\gamma - 1 \ , \end{aligned}$$

where we used Lemma 3.9 and the fact that $\|\boldsymbol{z}\|_\alpha \geq \|\boldsymbol{z}\|_\infty \geq d^{-1/2}\|\boldsymbol{z}\|_2$ to obtain the lower bound. □

The results that were just presented offer theoretical reasons to believe that the solution of the mean value problem is a valuable one in a class of stochastic programs when distribution information is limited. Initially, Proposition 3.1 demonstrated its robustness when one is only informed of upper bounds on the expected value of a set of convex functions. Later, Proposition 3.8 described conditions under which the robustness also prevails when the expected values of these functions are known exactly ; these conditions included the case where covariance matrix is known exactly. Since the conditions that we have presented are sufficient but in no way necessary, we expect that the MVP solution would remain the optimal robust solution in contexts where more distribution information would be available;

hence, it appears that the MVP solution is robust while not being too conservative. In order to prevent the MVP solution from being a worst-case (or approximately worst-case) distribution, there seems to be a need for moment information that is based on non-convex functions. It is clear for instance that if probability information (e.g. $\mathbb{P}_F(\|\boldsymbol{\xi} - \boldsymbol{\mu}\| \geq R) = \mathbb{E}_F[\mathbb{1}\{\|\boldsymbol{\xi} - \boldsymbol{\mu}\| \geq R\}] = p$) is included in $\bar{\mathcal{D}}$, then this should easily prevent the Dirac measure from being approximately optimal. Unfortunately, determining a distributionally robust solution in these cases is suspected to quickly become intractable given that evaluating the moment problem itself becomes challenging; see [4] for some NP-hardness results when moment are based on monomials.

Remark 3.4. *It is important to warn the reader that it is not generally the case that the MVP solution is a robust one. Indeed, the derived robustness property of the MVP solution relies heavily on the hypothesis that $\boldsymbol{\xi}$ is only involved in the objective function of the SLP. If we assumed instead that $\boldsymbol{\xi}$ affected the second stage feasible set, as in*

$$\begin{aligned} h(\boldsymbol{x}, \boldsymbol{\xi}) := \min_{\boldsymbol{y}} \quad & \boldsymbol{c}_2{}^{\mathsf{T}} \boldsymbol{y} \\ \text{subject to} \quad & \boldsymbol{A}_2 \boldsymbol{x} + \boldsymbol{B}_2 \boldsymbol{y} \leq \boldsymbol{\xi} \\ & \boldsymbol{y} \in \mathbb{R}^{n_2 - p_2} \times \mathbb{Z}^{p_2} \ , \end{aligned}$$

then we would observe using duality theory that $h(\boldsymbol{x}, \boldsymbol{\xi})$ is convex in $\boldsymbol{\xi}$. In this context, one can show that the MVP solution actually becomes an optimistic solution relative to ambiguity in the distribution:

$$\inf_{F \in \mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)} \mathbb{E}_F[h(\boldsymbol{x}, \boldsymbol{\xi})] \ \geq \ \inf_{F \in \mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)} h(\boldsymbol{x}, \mathbb{E}_F[\boldsymbol{\xi}]) \ = \ h(\boldsymbol{x}, \boldsymbol{\mu}) \ = \ \mathbb{E}_{\delta_{\boldsymbol{\mu}}}[h(\boldsymbol{x}, \boldsymbol{\xi})] \ .$$

The phenomenon of $h(\boldsymbol{x}, \boldsymbol{\xi})$ being non-concave, or even convex, in $\boldsymbol{\xi}$ occurs in a large number of contexts. It is therefore important to identify this property carefully before drawing hard conclusions about the robustness of the MVP solution.

4 The Value of Stochastic Modeling

In this section, we consider that one has in hand a candidate solution that was chosen using partial information about the random vector $\boldsymbol{\xi}$; such a candidate solution could either be obtained through the MVP model or any other approximation of the SLP. Our conjecture is that it is often wasteful to attempt to improve such a solution by developing a reliable stochastic model. Consequently, the objective of this section is to provide tools that can help identify such cases by estimating how much could potentially be gained from the investment of efforts in the resolution of this model.

Inspired by the foundations laid out in Section 3, we assume that the current information about $\boldsymbol{\xi}$ takes the form of a distributional set $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$ and that the MVP solution is the candidate we currently plan to implement, referred as $\boldsymbol{x}_1$. Based on our current knowledge of the distribution, we define the value of stochastic modeling as the maximum regret that might

be experienced after applying the current candidate $\boldsymbol{x}_1$ and finding out that the performance of $\boldsymbol{x}_1$ should actually have been measured using the "true" distribution F. Such a notion of regret can be computed using the form:

$$\text{VSM}(\boldsymbol{x}_1) := \sup_{F \in \mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)} \left\{ \boldsymbol{c}_1^\mathsf{T} \boldsymbol{x}_1 + \mathbb{E}_F[h(\boldsymbol{x}_1, \boldsymbol{\xi})] - \min_{\boldsymbol{x}_2 \in \mathcal{X}} \left\{ \boldsymbol{c}_1^\mathsf{T} \boldsymbol{x}_2 + \mathbb{E}_F[h(\boldsymbol{x}_2, \boldsymbol{\xi})] \right\} \right\} , \quad (4)$$

where we consider the worst-case scenario in terms of the true underlying distribution F. Intuitively, the regret that is experienced once F is known is measured as the difference between the expected cost obtained from applying our current decision and what could be achieved if the distribution information was at hand: i.e., the sub-optimality gap of applying the current decision in the true stochastic program. One should observe that the definition of *VSM* presented in equation (4) constitutes an optimistic view of what the value truly is. Actually, what we know is that the true value lies somewhere in the interval $[0, \text{VSM}(\boldsymbol{x}_1)]$.

From a modeling perspective, the problem of evaluating the *VSM* takes the shape of a semi-infinite linear program that is not too different from the inner problem involved in problem (1). Hence, our investment in distribution information should not exceed the *VSM*. Unfortunately, we are about to demonstrate that it is an intractable problem when the dimension of uncertainty becomes large. In an attempt to mitigate this issue, we will shortly propose tractable methods for finding both upper and lower bounds for the *VSM*.

4.1 NP-Hardness of *VSM* Evaluation

In order to understand better the computational difficulties associated with the evaluation of the *VSM*, we focus our attention on the following subclass of stochastic programs:

$$\underset{0 \leq x \leq 1}{\text{minimize}} \quad cx + \mathbb{E}_F[h(x, \boldsymbol{\xi})] , \quad (5)$$

where $x \in \mathbb{R}^+$, $c > 0$, $\boldsymbol{\xi}$ is known to be a random vector in $\mathbb{R}^d$ with zero mean and a covariance matrix that satisfies $\mathbb{E}_F[\boldsymbol{\xi}\boldsymbol{\xi}^\mathsf{T}] \preceq \mathbf{I}$ (see example 3.1 for proper motivation), and $h(x, \boldsymbol{\xi})$ is the optimal value of the second stage

$$\begin{aligned} \underset{\boldsymbol{y}}{\text{minimize}} \quad & \boldsymbol{\xi}^\mathsf{T} \boldsymbol{y} \\ \text{subject to} \quad & -x \leq y_i \leq x \ , \ \forall\, i \in \{1, 2, ..., d\} \\ & \boldsymbol{a}^\mathsf{T} \boldsymbol{y} = 0 \ . \end{aligned}$$

One can easily resolve that the solution of the mean value problem associated to this problem is $x_1 = 0$ since $cx + h(x, \mathbb{E}_F[\boldsymbol{\xi}]) = cx$. When measuring the *VSM* for this solution we are left with the evaluation of

$$\max_{0 \leq x_2 \leq 1} \sup_{F \in \mathcal{D}(\mathbb{R}^d, \mathbf{0}_d, \mathbf{I})} \{ c(x_1 - x_2) + \mathbb{E}_F[h(x_1, \boldsymbol{\xi})] - \mathbb{E}_F[h(x_2, \boldsymbol{\xi})] \} ,$$

where $\mathcal{D}(\mathbb{R}^d, \mathbf{0}_d, \mathbf{I})$ refers to the distributional set presented in Example 3.1. Since $x_1 = 0$ the problem reduces to solving

$$\underset{0 \le x_2 \le 1}{\text{maximize}} \quad \sup_{F \in \mathcal{D}(\mathbb{R}^d, \mathbf{0}_d, \mathbf{I})} \{\mathbb{E}_F[-cx_2 - h(x_2, \boldsymbol{\xi})]\} \,, \tag{6}$$

where we used the fact that $h(0, \boldsymbol{\xi}) = 0$ for all $\boldsymbol{\xi} \in \mathbb{R}^d$.

In our main result, we will make use of the following two lemmas which proofs are deferred to appendices F and G. The first one states that problem (6) is a linear program. The second lemma describes the difficulty related to evaluating the supremum operator at $x_2 = 1$.

Lemma 4.1. *Problem* (6) *is equivalent to the linear program*

$$\underset{x_2}{\text{maximize}} \quad \alpha x_2 \tag{7a}$$

$$\text{subject to} \quad 0 \le x_2 \le 1 \,, \tag{7b}$$

where $\alpha = \sup_{F \in \mathcal{D}(\mathbb{R}^d, \mathbf{0}_d, \mathbf{I})} \{\mathbb{E}_F[-c - h(1, \boldsymbol{\xi})]\}$.

Lemma 4.2. *It is NP-hard to find the supremum of the expression*

$$\sup_{F \in \mathcal{D}(\mathbb{R}^d, \mathbf{0}_d, \mathbf{I})} \mathbb{E}_F[-c - h(1, \boldsymbol{\xi})] \,.$$

Together, these two lemmas indicate that although problem (6) has the structure of a simple linear programming problem with optimal solution either at $x_2 = 0$ or $x_2 = 1$, evaluating the objective value at $x_2 = 1$ is a hard task and makes this problem computationally challenging.

Theorem 4.3. *Evaluating the VSM*$(\boldsymbol{x}_1)$*, as defined in equation* (4)*, is in general NP-hard in terms of the dimension of the uncertain vector* $\boldsymbol{\xi}$*.*

Proof: One can show that evaluating the *VSM* for stochastic program (5) is NP-hard. First, using Lemma 4.1 we study the equivalent problem (7) after rewriting it in the following form

$$\begin{aligned} \underset{t, x_2}{\text{maximize}} \quad & t \\ \text{subject to} \quad & t \le \alpha x_2 \\ & 0 \le x_2 \le 1 \,. \end{aligned}$$

Now, consider verifying the feasibility of $t = \beta$ and $x_2 = 1$ for any value of $\beta > 0$. This requires verifying that

$$\beta \le \sup_{F \in \mathcal{D}(\mathbb{R}^d, \mathbf{0}_d, \mathbf{I})} \mathbb{E}_F[-c - h(1, \boldsymbol{\xi})] \,.$$

Yet, Lemma 4.2 tells us that evaluating the right hand side expression is NP-hard. Therefore, by the equivalence of optimization and separation (as presented in [24]), solving problem (5) is NP-hard. This completes our proof. □

Note that the computational difficulties that were identified hold even when decision variables are all of the continuous type. Intuitively, the difficulty comes from seeking the most favorable distribution to exploit in the minimization of expected cost. This task is equivalent to solving a minimization problem where the objective function is not jointly convex in all decision variables. Since we cannot hope to find an exact value for this measure in a reasonable amount of time, in what follows we propose methods for bounding the *VSM*.

4.2 Lower Bounding the *VSM*

We start by proposing a lower bound for the value of stochastic modeling as expressed in equation (4). We first make an assumption that ensures that the set Ψ is one that is computationally feasible to work with.

Assumption 4.4. *The set of convex functions Ψ can be represented in the parameterized form*

$$\Psi := \{\psi : \mathbb{R}^d \to \mathbb{R} \mid \exists\, \boldsymbol{r} \in \mathcal{K}\,,\ \psi(\boldsymbol{\xi}) = \boldsymbol{r}^{\mathsf{T}}\boldsymbol{\psi}(\boldsymbol{\xi})\}\ ,$$

for some mapping $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^D$ and some convex cone $\mathcal{K} \subset \mathbb{R}^D$ for which there exists a tractable feasibility oracle, and for which $\boldsymbol{r}^T\boldsymbol{\psi}(\boldsymbol{z})$ is a convex function for all $\boldsymbol{r} \in \mathcal{K}$.

Note that this assumption is a natural one to make when Ψ contains an intractable number of functions if one wishes to be able to even verify computationally that a given candidate distribution is a feasible one.

In order to provide this bound, we choose to approximate problem (4) by focusing our attention to distributions for which all realizations have at most one parameter that diverges from its mean. Specifically, we limit our search to distributions in $\mathcal{D}(\mathcal{S}_0, \boldsymbol{\mu}, \Psi)$, where $\mathcal{S}_0 = \{\boldsymbol{\xi} \in \mathcal{S} \mid \exists\, i\,,\ \xi_j = \mu_j\, \forall\, j \neq i\}$. Indeed, by reducing the support of F we can only reduce the achievable regret. We then rely on a discretization of the space $\mathcal{S}_0$, which will grow at a tractable rate with respect to d, to get an estimate of how low VSM might be.

Definition 4.5. *Let $\mathcal{B}_\infty(\boldsymbol{\nu}, \boldsymbol{\tau}) = \{\boldsymbol{\xi} \in \mathbb{R}^d \mid |\xi_i - \nu_i| \leq \tau_i\ \forall\, i\}$ be any box known to contain $\mathcal{S}_0$. Given, for each parameter ξ_i, a discretization of K points $\{\xi_i^k\}_{k=1}^K$ of the interval $[\nu_i - \tau_i, \nu_i + \tau_i]$. Let $\{\boldsymbol{\xi}_i^k\}$ be a discretization of $\mathcal{S}_0$ constructed using $\boldsymbol{\xi}_i^k = \boldsymbol{\mu} + (\xi_i^k - \mu_i)\boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is the i-th column of the identity matrix. Given a first stage decision $\boldsymbol{x}_1$ and a region $\mathcal{X}_2 \subseteq \mathbb{R}^{n_1}$, we define $\mathcal{LB}(\boldsymbol{x}_1, \mathcal{X}_2, \{\boldsymbol{\xi}_i^k\})$ to be the optimal value of the convex optimization problem*

$$\underset{t, \boldsymbol{q}, \boldsymbol{r}, \{\boldsymbol{w}^k, \boldsymbol{v}^k\}_{k=1}^K}{\text{minimize}} \quad t \tag{8a}$$

$$\text{subject to} \quad t \geq w_i^k + \xi v_i^k - \alpha_i^k - \beta_i^k(\xi - \xi_i^k) - q_i(\xi - \mu_i) - \boldsymbol{r}^{\mathsf{T}}\boldsymbol{\psi}(\boldsymbol{\mu} + (\xi - \mu_i)\boldsymbol{e}_i)\ , \quad \begin{cases} \forall\, \xi \in [\nu_i - \tau_i, \nu_i + \tau_i] \\ \forall i \in \{1, 2, ..., d\} \\ \forall k\ \in \{1, 2, ..., K\} \end{cases} \tag{8b}$$

$$w_i^k + v_i^k \xi_i^m \geq \boldsymbol{c}_1^{\mathsf{T}}\boldsymbol{x}_1 + h(\boldsymbol{x}_1, \boldsymbol{\mu} + (\xi_i^m - \mu_i)\boldsymbol{e}_i)\ , \quad \begin{cases} \forall\, i \in \{1, 2, ..., d\} \\ \forall\, m,\ k\ \in \{1, 2, ..., K\} \end{cases} \tag{8c}$$

$$\boldsymbol{r} \in \mathcal{K}\ , \tag{8d}$$

where $t \in \mathbb{R}$, $\boldsymbol{q} \in \mathbb{R}^d$, $\boldsymbol{r} \in \mathbb{R}^D$, *and* $\boldsymbol{w}^k, \boldsymbol{v}^k \in \mathbb{R}^d$ *are the decision variables, while the constants* α_i^k *and* β_i^k *respectively refer to the optimal value of the associated scenario based optimization problem, and its super-gradient with respect to* ξ_i. *Namely, let* $\alpha_i^k = \min_{\boldsymbol{x}\in\mathcal{X}_2, \boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} \ \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} + (\boldsymbol{\xi}_i^k)^\mathsf{T}\boldsymbol{C}_2\boldsymbol{y}$, *and* $\beta_i^k = \boldsymbol{e}_i^\mathsf{T}\boldsymbol{C}_2\boldsymbol{y}^*$ *for any* $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}_2, \boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} \ \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} + (\boldsymbol{\xi}_i^k)^\mathsf{T}\boldsymbol{C}_2\boldsymbol{y}$.

We will soon see that when $\mathcal{LB}(\boldsymbol{x}_1, \{\hat{x}_2\}, \{\boldsymbol{\xi}_i^k\})$ gives a lower bound to the VSM, yet before doing so we can already recognize that to evaluate $\mathcal{LB}(\boldsymbol{x}_1, \mathcal{X}_2, \{\boldsymbol{\xi}_i^k\})$, it is first necessary to solve the MVP problem twice at every points of the discretization of $\mathcal{S}_0$. We can then rely on solving a robust optimization problem which in general might require the use of efficient cutting plane methods. In the particular case where the distribution information is limited to mean and covariance information, this robust optimization problem can further be reduced to a semi-definite program (SDP). We defer the proof of the following proposition to Appendix I.

Proposition 4.6. *Given some* $\hat{\boldsymbol{x}}_2 \in \mathcal{X}$ *and the fact the distributional set takes the form ,* $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \mathbf{I})$, *from Example 3.1, the value of* $\mathcal{LB}(\boldsymbol{x}_1, \mathcal{X}_2, \{\boldsymbol{\xi}_i^k\})$ *can be found using*

$$
\begin{aligned}
\underset{t,\boldsymbol{q},\boldsymbol{r},\{\boldsymbol{s}^k,\boldsymbol{w}^k,\boldsymbol{v}^k\}_{k=1}^K}{\text{minimize}} \quad & t + \boldsymbol{e}^\mathsf{T}\boldsymbol{r} \\
\text{subject to} \quad & w_i^k + v_i^k\xi_i^m \geq \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x}_1 + h(\boldsymbol{x}_1, \boldsymbol{\mu} + (\xi_i^m - \mu_i)\boldsymbol{e}_i) \ , \ \begin{cases} \forall\, i \in \{1,2,...,d\} \\ \forall\, m,\ k \ \in \{1,2,...,K\} \end{cases} \\
& \begin{bmatrix} \boldsymbol{r}_i & \frac{\beta_i^k - v_i^k + q_i - 2\mu_i\boldsymbol{r}_i}{2} \\ \frac{\beta_i^k - v_i^k + q_i - 2\mu_i\boldsymbol{r}_i}{2} & t - w_i^k + \alpha_i^k - \beta_i^k\xi_i^k - \mu_i q_i + \mu_i^2\boldsymbol{r}_i \end{bmatrix} \succeq -s_i \begin{bmatrix} 1 & -\nu_i \\ -\nu_i & \tau_i^2 \end{bmatrix} \ , \ \begin{cases} \forall\, i \in \{1,2,...,d\} \\ \forall\, k \in \{1,2,...,K\} \end{cases} \\
& \qquad \boldsymbol{r} \geq 0 \quad , \quad \boldsymbol{s}^k \geq 0 \ , \forall\, k \ \in \{1,2,...,K\} \ ,
\end{aligned}
$$

where $t \in \mathbb{R}$, $\boldsymbol{q}, \boldsymbol{r}, \boldsymbol{s} \in \mathbb{R}^d$, *and* $\boldsymbol{w}^k, \boldsymbol{v}^k \in \mathbb{R}^d$ *are the decision variables, while* $\sigma_i^2 = \Sigma_{i,i}$, *and the constants* α_i^k *and* β_i^k *are as defined in Definition 4.5. Moreover, this lower bound for the VSM can be obtained in* $O(K^5 d^{3.5} + Kd\,T_{MVP})$.

Note that it is trivial to generalize Proposition 4.6 to some arbitrary covariance matrix $\boldsymbol{\Sigma} \neq \mathbf{I}$, as long as $\boldsymbol{\Sigma} \succeq 0$, since in that case the VSM($\boldsymbol{x}_1$) problem can simply be reformulated in terms of $D(\mathcal{S}, \boldsymbol{\mu}, \mathbf{I})$ by replacing $\boldsymbol{\mu}$ and $\boldsymbol{C}_2$ by $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{1/2}\boldsymbol{C}_2$ respectively. The property of the discretization that is exploited in Proposition 4.6 is the fact that each ray of $\mathcal{S}_0$ is aligned with an eigenvector of $\boldsymbol{\Sigma}$.

Given that we confirmed that $\mathcal{LB}(\boldsymbol{x}_1, \mathcal{X}_2, \{\boldsymbol{\xi}_i^k\})$ is a tractable problem, we now confirm the relation of this problem to our objective of bounding the VSM.

Proposition 4.7. *Given some* $\hat{\boldsymbol{x}}_2 \in \mathcal{X}$, *the value* $\mathcal{LB}(\boldsymbol{x}_1, \{\hat{\boldsymbol{x}}_2\}, \{\boldsymbol{\xi}_i^k\})$ *is a lower bound for the VSM as expressed in problem* (4).

We refer the reader to Appendix H for a detailed proof of this result. Generally speaking, this result relies on initially fixing the decision $\boldsymbol{x}_2$ to $\hat{\boldsymbol{x}}_2$, approximating the expression $\boldsymbol{c}_1^\mathsf{T}(\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_2) + h(\boldsymbol{x}_1, \boldsymbol{\xi}) - h(\hat{\boldsymbol{x}}_2, \boldsymbol{\xi})$ on $\mathcal{S}_0$ with a piecewise-linear curve, and applying duality theory. Our choice of discretizing $\mathcal{S}_0$ instead of $\mathcal{S}$ is made to ensure that as the dimensionality of d

increases the number of points required to formulate $\mathcal{LB}(\boldsymbol{x}_1, \{\hat{\boldsymbol{x}}_2\}, \{\boldsymbol{\xi}_i^k\})$ to an "equivalent" level of accuracy does not increase exponentially.

To obtain a non-trivial bound, such as when setting $\hat{\boldsymbol{x}}_2 = \boldsymbol{x}_1$, one should optimize:

$$\underset{\hat{\boldsymbol{x}}_2 \in \mathcal{X}}{\text{maximize}} \quad \mathcal{LB}(\boldsymbol{x}_1, \{\hat{\boldsymbol{x}}_2\}, \{\boldsymbol{\xi}_i^k\}) \; .$$

Unfortunately, this is a non-convex optimization problem. We therefore suggest using a global optimization method such as branch and bound. Indeed, one can easily show that for any convex region $\hat{\mathcal{X}}$ associated to a node, $\mathcal{LB}(\boldsymbol{x}_1, \hat{\mathcal{X}}, \{\boldsymbol{\xi}_i^k\})$ will give an upper bound to the optimal value achievable over this region while $\mathcal{LB}(\boldsymbol{x}_1, \{\hat{\boldsymbol{x}}_2\}, \{\boldsymbol{\xi}_i^k\})$, with $\hat{x}_2 \in \hat{\mathcal{X}}$, gives a new lower bound to the problem. Given that the upper bound associated to a node is large enough, one can branch on any decision variable (i.e., $x_{2,i} \leq \hat{x}_{2,i}$ or $x_{2,i} \geq \hat{x}_{2,i}$. This branching process preserves the convexity of the sub-regions. In contexts where the size of the first stage decision vector is relatively smaller than the decision vector involved in the second stage, we can expect such a procedure to converge in a reasonable amount of time. Section 6 will provide a numerical example.

Remark 4.1. *Since our approach relies on the discretization of the support space, one could argue that a more straightforward approach would be to simply search for a distribution with support on these points. This would lead to solving the problem*

$$\begin{aligned}
\underset{\boldsymbol{w} \in \mathcal{P}}{\text{maximize}} \quad & \sum_{i,k} w_i^k \left(\boldsymbol{c}_1^\mathsf{T}(\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_2) + h(\boldsymbol{x}_1, \boldsymbol{\xi}_i^k) - h(\hat{x}_2, \boldsymbol{\xi}_i^k) \right) \\
& \sum_{i,k} w_i^k = 1 \qquad w_i^k \geq 0 \; , \; \forall \, i, k \\
& \sum_{i,k} w_i^k \boldsymbol{\xi}_i^k = \boldsymbol{\mu} \\
& \sum_{i,k} w_i^k (\boldsymbol{r}^\mathsf{T} \boldsymbol{\psi}(\boldsymbol{\xi}_i^k)) \leq 0 \, , \; \forall \, \boldsymbol{r} \in \mathcal{K} \, .
\end{aligned}$$

Yet, one can show that $\mathcal{LB}(\boldsymbol{x}_1, \{\hat{\boldsymbol{x}}_2\}, \{\boldsymbol{\xi}_i^k\})$ gives a tighter bound.

4.3 Upper Bounding the *VSM*

We now turn ourself to proposing an upper bound for the *VSM* as defined in equation (4). Our approach first relaxes the shape of the support set to a polygon with a number of vertices that is linear in the dimension of $\boldsymbol{\xi}$, and then discard all distributional information except for the mean and the support.

Definition 4.8. *Given a set $\bar{\mathcal{S}}_\rho = \{\boldsymbol{\xi} \in \mathbb{R}^d \,|\, \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|_1 \leq \rho\}$ such that $\mathcal{S} \subseteq \bar{\mathcal{S}}_\rho$, and a pair $(\boldsymbol{x}_1, \bar{\boldsymbol{y}}_1)$ of assignments for the first and second stage decisions, let $\mathcal{UB}(\bar{\mathcal{S}}_\rho, \boldsymbol{x}_1, \bar{\boldsymbol{y}}_1)$ be the*

optimal value of the problem

$$\begin{aligned} \underset{s,\boldsymbol{q}}{\text{minimize}} \quad & s + (\boldsymbol{\mu} - \boldsymbol{\xi}_0)^\mathsf{T}\boldsymbol{q} && \text{(10a)} \\ \text{subject to} \quad & s \geq \alpha(\boldsymbol{\xi}_0 + \rho\boldsymbol{e}_i) - \rho\boldsymbol{e}_i^\mathsf{T}\boldsymbol{q} \;,\; \forall\, i \in \{1,...,d\} && \text{(10b)} \\ & s \geq \alpha(\boldsymbol{\xi}_0 - \rho\boldsymbol{e}_i) + \rho\boldsymbol{e}_i^\mathsf{T}\boldsymbol{q} \;,\; \forall\, i \in \{1,...,d\}\,, && \text{(10c)} \end{aligned}$$

where $\boldsymbol{e}_i$ *is the* i*-th column of the* $d \times d$ *identity matrix, and* $\alpha(\boldsymbol{\xi}) = \max_{\boldsymbol{x}_2 \in \mathcal{X}, \boldsymbol{y}_2 \in \mathcal{Y}(\boldsymbol{x}_2)} \boldsymbol{c}_1^\mathsf{T}(\boldsymbol{x}_1 - \boldsymbol{x}_2) + \boldsymbol{\xi}^\mathsf{T}\boldsymbol{C}(\bar{\boldsymbol{y}}_1 - \boldsymbol{y}_2)$*.*

Proposition 4.9. *The value* $\mathcal{UB}(\bar{\mathcal{S}}_\rho, \boldsymbol{x}_1, \bar{\boldsymbol{y}}_1)$ *is an upper bound for VSM*$(\boldsymbol{x}_1)$*.*

Proof: We first choose to expand the set of distributions that we consider in measuring the value of stochastic modeling and only preserve the information about the relaxed support $\bar{\mathcal{S}}_\rho$ and the mean $\boldsymbol{\mu}$. We also relax the problem by allowing $\boldsymbol{x}_2$ to adapt to the realization of $\boldsymbol{\xi}$ in the spirit of evaluating the wait and see performance.

$$\begin{aligned} \text{VSM}(\boldsymbol{x}_1) \;&\leq\; \sup_{F \in \mathcal{D}(\bar{\mathcal{S}}_\rho, \boldsymbol{\mu}, \emptyset)} \max_{\boldsymbol{x}_2 \in \mathcal{X}} \mathbb{E}_F[\boldsymbol{c}_1^\mathsf{T}(\boldsymbol{x}_1 - \boldsymbol{x}_2) + h(\boldsymbol{x}_1, \boldsymbol{\xi}) - h(\boldsymbol{x}_2, \boldsymbol{\xi})] \\ &\leq\; \sup_{F \in \mathcal{D}_2(\bar{\mathcal{S}}_\rho, \boldsymbol{\mu}, \emptyset)} \mathbb{E}_F\left[\max_{\boldsymbol{x}_2 \in \mathcal{X}} \left\{\boldsymbol{c}_1^\mathsf{T}(\boldsymbol{x}_1 - \boldsymbol{x}_2) + h(\boldsymbol{x}_1, \boldsymbol{\xi}) - h(\boldsymbol{x}_2, \boldsymbol{\xi})\right\}\right]. \end{aligned}$$

We are left with the task of computing an upper bound for a generalized moment problem. Following the strategies popularized by [16], [22], [10], and many others, one can simply apply weak duality, and get that the value of stochastic modeling is upper bounded by the optimal value of the problem

$$\begin{aligned} \underset{s,\boldsymbol{q}}{\text{minimize}} \quad & s + \boldsymbol{\mu}^\mathsf{T}\boldsymbol{q} \\ \text{subject to} \quad & s \geq \max_{\boldsymbol{x}_2 \in \mathcal{X}} \left\{\boldsymbol{c}_1^\mathsf{T}(\boldsymbol{x}_1 - \boldsymbol{x}_2) + \min_{\boldsymbol{y}_1 \in \mathcal{Y}(\boldsymbol{x}_1)} \boldsymbol{\xi}^\mathsf{T}\boldsymbol{C}_2\boldsymbol{y}_1 - \min_{\boldsymbol{y}_2 \in \mathcal{Y}(\boldsymbol{x}_2)} \boldsymbol{\xi}^\mathsf{T}\boldsymbol{C}_2\boldsymbol{y}_2 - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{q}\right\} \;,\; \forall\, \boldsymbol{\xi} \in \bar{\mathcal{S}}_\rho\,. \end{aligned}$$

Our last upper bounding step will be to reduce the feasible set of this problem by replacing $\min_{\boldsymbol{y}_1 \in \mathcal{Y}(\boldsymbol{x}_1)} \boldsymbol{\xi}^\mathsf{T}\boldsymbol{C}_2\boldsymbol{y}_1$ by the larger value $\boldsymbol{\xi}^\mathsf{T}\boldsymbol{C}_2\bar{\boldsymbol{y}}_1$ in the set of constraints. The more restrictive constraint takes the form:

$$s \geq \max_{\boldsymbol{\xi} \in \bar{\mathcal{S}}_\rho} \max_{\boldsymbol{x}_2 \in \mathcal{X}, \boldsymbol{y}_2 \in \mathcal{Y}(\boldsymbol{x}_2)} \left\{\boldsymbol{c}_1^\mathsf{T}(\boldsymbol{x}_1 - \boldsymbol{x}_2) + \boldsymbol{\xi}^\mathsf{T}\boldsymbol{C}_2\bar{\boldsymbol{y}}_1 - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{C}_2\boldsymbol{y}_2 - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{q}\right\}.$$

Since the expression that is maximized in terms of $\boldsymbol{\xi}$ is the maximization of a convex function over a convex set, we can conclude that their must be an optimal solution for $\boldsymbol{\xi}$ at one of the vertices of $\bar{\mathcal{S}}_\rho$; more specifically, in the set $\{\boldsymbol{\xi}_0 + \rho\boldsymbol{e}_1, \boldsymbol{\xi}_0 - \rho\boldsymbol{e}_1, \boldsymbol{\xi}_0 + \rho\boldsymbol{e}_2, \boldsymbol{\xi}_0 - \rho\boldsymbol{e}_2, ..., \boldsymbol{\xi}_0 + \rho\boldsymbol{e}_d, \boldsymbol{\xi}_0 - \rho\boldsymbol{e}_d\}$. This fact naturally leads to the optimization model used in the definition of $\mathcal{UB}(\bar{\mathcal{S}}_\rho, \boldsymbol{x}_1, \bar{\boldsymbol{y}}_1)$ and concludes this proof. □

Although we expect this bound to be a bit loose in general, it can be computed in an amount of time that is comparable to the time spent searching for the MVP solution;

specifically, in $O(d^{3.5} + d\,T_{\text{MVP}})$ where T_{MVP} is the amount of time required to solve the mean value problem. It therefore has the potential of quickly shedding some light on questions related to the need for supplementary information about the distribution. There is also some hope that one can exploit the structure of some subclasses of stochastic linear program to provide tighter VSM bounds. The stochastic fleet mix optimization problem discussed next is one such example.

5 Application to Stochastic Fleet Mix Optimization

A stochastic fleet mix optimization problem considers the decisions that an airline company must take when composing a fleet of aircraft for the service of some future flights. As these decisions are typically made ten to twenty years ahead of time, it involves a large amount of uncertainty with respect to the profits that will be generated from any given portfolio of aircraft. It is therefore natural to formulate this problem as a two-stage stochastic mixed integer linear program (see for example [26]). More specifically, the problem takes the following shape

$$\underset{\boldsymbol{x}\geq 0}{\text{maximize}} \quad -\boldsymbol{o}^{\mathsf{T}}\boldsymbol{x} + \mathbb{E}_{\,F}[\rho(\boldsymbol{x}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{c}}, \tilde{\boldsymbol{l}})] \quad , \tag{11}$$

where $\boldsymbol{x} \in \mathbb{R}^{n_1}$ is a vector describing how many aircraft of each type is acquired by the airline, o_k is the deterministic ownership cost for an aircraft of type k, $\rho(\boldsymbol{x}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{c}}, \tilde{\boldsymbol{l}})$ is a function that computes the future weekly profits generated from the chosen fleet, and $\tilde{\boldsymbol{p}}$, $\tilde{\boldsymbol{c}}$, $\tilde{\boldsymbol{l}}$ are some uncertain longer term profit parameters (described shortly). In estimating future revenues and costs, the model assumes that the airline company allocates N flights to aircraft in a way that is optimal with respect to actual expenses and revenues. This gives rise to a second stage optimization model of the form

$$\begin{aligned}
\rho(\boldsymbol{x}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{c}}, \tilde{\boldsymbol{l}}) \;:=\; & \underset{\boldsymbol{z}^+, \boldsymbol{z}^-, \boldsymbol{w}, \boldsymbol{u}}{\text{maximize}} \quad \sum_k \left(\sum_i \tilde{p}_i^k w_i^k - \tilde{c}_k z_k^+ + \tilde{l}_k z_k^- \right) && \text{(12a)} \\
& \text{subject to} \quad \sum_k w_i^k = 1 \;, \quad \forall\, i \in \{1, 2, ..., N\} && \text{(12b)} \\
& \qquad \sum_{g \in \text{in}(v)} u_g^k + \sum_{i \in \text{arr}(v)} w_i^k = \sum_{g \in \text{out}(v)} u_g^k + \sum_{i \in \text{dep}(v)} w_i^k \;, \quad \forall\, k \;, \quad \forall\, v && \text{(12c)} \\
& \qquad x_k + z_k^+ - z_k^- = \sum_{v \in \{v | \text{time}(v) = 0\}} \left(\sum_{g \in \text{out}(v)} v_g^k + \sum_{i \in \text{dep}(v)} w_i^k \right) \;, \quad \forall\, k && \text{(12d)} \\
& \qquad z_k^+ \geq 0 \;, \quad z_k^- \geq 0 \;, \quad u_g^k \geq 0 \;, \quad w_i^k \in \{0, 1\} \;, \quad \forall\, k \;, \quad \forall\, g \;, \quad \forall\, i \;. && \text{(12e)}
\end{aligned}$$

where each variable $w_i^k = 1$ describes whether or not an aircraft of type k is assigned to flight i. The vectors $\boldsymbol{z}^+ \in \mathbb{R}^{n_1}$ and $\boldsymbol{z}^- \in \mathbb{R}^{n_1}$ respectively count how many aircraft of each type need to be added, or can be removed, from the fleet under the optimal fleet assignment. The second stage profits decompose into the following sets of terms: the first set denotes profits made from each flight, the second set denotes rental costs for extra aircraft that are required by the assignment, and the last set denotes profits made from leasing out the part of the fleet which is unused. Specifically, $\tilde{p}_i^k$ is the profit generated by using an aircraft of type k for flight i, $\tilde{c}_k$ is the cost of renting an aircraft of type k, and $\tilde{l}_k$ is the revenue per aircraft of type k leased out. The general form of this problem considers $\tilde{\boldsymbol{p}}$, $\tilde{\boldsymbol{c}}$, and $\tilde{\boldsymbol{l}}$ to be uncertain at the time of composing the original fleet mix since they depend on factors that can not be determined at that time due to fluctuating demand for the flight, price of gas, ticket price, crew cost, etc. Note that these factors are assumed to be resolved when comes the time of making the final allocation of aircraft to flights. We will also assume here, as it is typically the case, that the rental price of an aircraft is assured to be higher that the original ownership cost and that there are no possibility of arbitrage; specifically, $\tilde{l}_k \leq o_k \leq \tilde{c}_k$ with probability equal to one for all k.

Flight ID	From	Departure Time	To	Arrival Time (+ turn time)
#1	A	t_0	B	t_1
#2	B	t_1	A	t_2
#3	B	t_1	C	t_2
#4	A	t_2	B	t_3
#5	C	t_2	A	t_3

Table 1: Example of flight schedule for a fleet mix optimization problem involving airports A, B, and C.

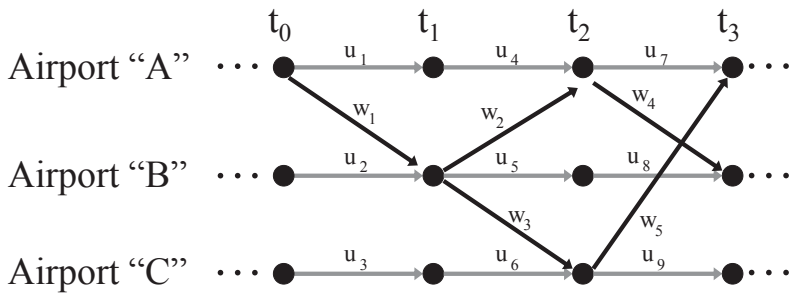

Figure 1: Flow graph associated with the flight schedule presented in Table 1.

The second stage aircraft allocation must satisfy some basic constraints. Constraint (12b) enforces that each flight is serviced by an aircraft. More importantly, constraints (12c) ensures that in an allocation scheme the aircraft needed for a flight is actually present in the

airport at departure time. This gives rise to some constraints reminiscent of flow constraints in network problems. The graph of flow constraints is derived from the schedule of flights. Some variables $\boldsymbol{u}$ are added to the problem formulation to allow an aircraft to stay in an airport garage between flights. Figure 1 presents the graph corresponding to the simple schedule presented in Table 1. The nodes of this graph are indexed by v. The sets of flight legs arriving to or departing from a node v are referred to as "arr(v)" and "dep(v)" respectively while the sets of ground legs incoming to and outgoing from an airport at time v are referred to as "in(v)" and "out(v)" respectively. Finally, constraint (12d) ensures that there are enough aircraft of each type to accommodate the schedule. In particular, while the left-hand side of the equation computes the number of aircraft available including the effect of rentals, the right-hand side sums over all nodes associated to time $t = 0$ how many aircraft are used in the network.

This version of the stochastic fleet mix optimization problem is a good example of a two-stage decision model where the uncertainty is limited to the second stage cost, while the second stage feasible set is deterministic. This is due to the assumption that the scheduling of flights has been completed prior to composing the original fleet mix. Specifically, one can actually rewrite the second stage problem as the following linear program:

$$\begin{aligned} \rho(\boldsymbol{x}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{c}}, \tilde{\boldsymbol{l}}) \;=\; \rho(\boldsymbol{x}, \boldsymbol{\xi}) \;:=\; \underset{\boldsymbol{y}}{\text{maximize}} \quad & \boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{y} \\ \text{subject to} \quad & \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}) \quad , \end{aligned}$$

where $\boldsymbol{\xi} = [\tilde{\boldsymbol{p}}^{\mathsf{T}},\ \tilde{\boldsymbol{c}}^{\mathsf{T}},\ \tilde{\boldsymbol{l}}^{\mathsf{T}},\ \boldsymbol{0}^{\mathsf{T}}]$, $\boldsymbol{y} = [\boldsymbol{w}^{\mathsf{T}},\ \boldsymbol{z}^{+\mathsf{T}},\ \boldsymbol{z}^{-\mathsf{T}},\ \boldsymbol{u}^{\mathsf{T}}]$, and $\mathcal{Y}(\boldsymbol{x})$ is the set of all assignments for the vector $\boldsymbol{y}$ that satisfy constraints (12b), (12c), (12d), and (12e) jointly. The following corollary is a direct consequence of Proposition 3.1.

Corollary 5.1. *Let $\hat{\boldsymbol{x}}$ be a solution of the mean value problem for the stochastic fleet mix optimization problem : i.e.,*

$$\hat{\boldsymbol{x}} \in \arg\max -\boldsymbol{o}^{\mathsf{T}}\boldsymbol{x} + \rho(\boldsymbol{x}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{l}}) \ ,$$

where $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{c}}$, $\hat{\boldsymbol{l}}$ are the mean of the random vectors $\tilde{\boldsymbol{p}}$, $\tilde{\boldsymbol{c}}$, and $\tilde{\boldsymbol{l}}$ respectively. The fleet composition $\hat{\boldsymbol{x}}$ is robust to ambiguity in the joint distribution F of $\tilde{\boldsymbol{p}}$, $\tilde{\boldsymbol{c}}$, and $\tilde{\boldsymbol{l}}$. Specifically,

$$\hat{\boldsymbol{x}} \in \arg\max_{\boldsymbol{x} \geq 0} \ \inf_{F \in \mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)} -\boldsymbol{o}^{\mathsf{T}}\boldsymbol{x} + \mathbb{E}_F[\rho(\boldsymbol{x}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{c}}, \tilde{\boldsymbol{l}})] \ ,$$

where $\boldsymbol{\mu} = [\hat{\boldsymbol{p}}^{\mathsf{T}},\ \hat{\boldsymbol{c}}^{\mathsf{T}},\ \hat{\boldsymbol{l}}^{\mathsf{T}},\ \boldsymbol{0}^{\mathsf{T}}]$, for any choice of $\mathcal{S}$ such that $\boldsymbol{\mu} \in \mathcal{S}$ and of set Ψ of convex functions.

To the best of our knowledge, the above results consider for the first time the distributionally robust version of this problem.

The structure of the stochastic fleet mix optimization allows us to propose an alternate method for computing an upper bound on the value of stochastic modeling based on the solution of the mean value problem.

Definition 5.2. *Given a fleet mix $\boldsymbol{x}_1$, a flight allocation $\hat{\boldsymbol{w}}$, a mean vector $\hat{p} = \mathbb{E}_F[\tilde{\boldsymbol{p}}]$, and a covariance matrix $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{p}}} = \mathbb{E}_F[(\tilde{\boldsymbol{p}} - \hat{\boldsymbol{p}})(\tilde{\boldsymbol{p}} - \hat{\boldsymbol{p}})^\mathsf{T}] \succ 0$, let $\mathcal{UB}_2(\boldsymbol{x}_1, \hat{\boldsymbol{w}}, \hat{\boldsymbol{p}}, \boldsymbol{\Sigma}) = \nu_0 + \sum_{i=1}^{N} \nu_i$, where ν_0 is the optimal value of*

$$\begin{array}{ll} \underset{\boldsymbol{x}_2, \boldsymbol{w}, \boldsymbol{u}}{\text{maximize}} & -\boldsymbol{o}^\mathsf{T}(\boldsymbol{x}_2 - \boldsymbol{x}_1) \\ \text{subject to} & [\boldsymbol{w}^\mathsf{T},\ \boldsymbol{0}^\mathsf{T},\ \boldsymbol{0}^\mathsf{T},\ \boldsymbol{u}^\mathsf{T}] \in \mathcal{Y}(\boldsymbol{x}_2) \quad . \end{array}$$

and, for $i \in \{1, 2, ..., N\}$, ν_i is the optimal value of

$$\begin{array}{ll} \underset{\boldsymbol{Q}_i, \boldsymbol{q}_i, t_i}{\text{minimize}} & t_i + \boldsymbol{S}_i \bullet \boldsymbol{Q}_i \\ \text{subject to} & \left[\begin{array}{cc} \boldsymbol{Q}_i & (\boldsymbol{q}_i - \boldsymbol{e}_j + \hat{\boldsymbol{w}}_i)/2 \\ (\boldsymbol{q}_i - \boldsymbol{e}_j + \hat{\boldsymbol{w}}_i)^\mathsf{T}/2 & t_i - \hat{\boldsymbol{p}}_i^\mathsf{T}(\boldsymbol{e}_j + \hat{\boldsymbol{w}}_i) \end{array} \right] \succeq 0 \ , \ \forall\, j \in \{1, 2, ..., K\} \ , \end{array}$$

where $\tilde{\boldsymbol{p}}_i \in \mathbb{R}^K$ is the random vector of profits associated with the assignment of each type of aircraft to flight i, $\hat{\boldsymbol{p}}_i$ is the expected value of $\tilde{\boldsymbol{p}}_i$, and $\boldsymbol{S}_i = \mathbb{E}_F[(\tilde{\boldsymbol{p}}_i - \hat{\boldsymbol{p}}_i)(\tilde{\boldsymbol{p}}_i - \hat{\boldsymbol{p}}_i)^\mathsf{T}]$, while $\boldsymbol{q}_i \in \mathbb{R}^K$ and $\boldsymbol{Q}_i \in \mathbb{R}^{K \times K}$ are new decision variables.

Proposition 5.3. *Given that $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{w}}$ are optimal solutions of the mean value problem associated to problem* (11)*, the value $\mathcal{UB}_2(\hat{\boldsymbol{x}}, \hat{\boldsymbol{w}}, \hat{\boldsymbol{p}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{p}}})$ is an upper bound on VSM($\hat{\boldsymbol{x}}$) in problem* (11) *assuming that the true distribution is such that $\mathbb{E}_F[\tilde{\boldsymbol{p}}] = \hat{\boldsymbol{p}}$ and $\mathbb{E}_F[(\tilde{\boldsymbol{p}} - \hat{\boldsymbol{p}})(\tilde{\boldsymbol{p}} - \hat{\boldsymbol{p}})^\mathsf{T}] = \boldsymbol{\Sigma}_{\tilde{\boldsymbol{p}}}$.*

Proof: First, one can easily show that $\hat{\boldsymbol{z}}^+ = \boldsymbol{0}$ and $\hat{\boldsymbol{z}}^- = \boldsymbol{0}$ form with $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{w}}$, and some $\hat{\boldsymbol{u}}$ a feasible solution to the mean value problem for problem (11). The second stage solution $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{w}}^\mathsf{T},\ \boldsymbol{0}^\mathsf{T},\ \boldsymbol{0}^\mathsf{T},\ \hat{\boldsymbol{u}}^\mathsf{T}]^\mathsf{T}$ is actually optimal for the MVP since $\mathbb{E}_F[\tilde{\boldsymbol{c}}] \geq \mathbb{E}_F[\tilde{\boldsymbol{l}}]$. We can use $\hat{\boldsymbol{y}}$ and the wait and see trick to find an upper bound on the worst-case regret potentially experienced when using the fleet mix $\hat{\boldsymbol{x}}$.

$$\begin{aligned} \text{VSM}(\hat{\boldsymbol{x}}) &= \sup_{F \in \mathcal{D}(\mathbb{R}^m, \boldsymbol{\mu}, \boldsymbol{\Sigma})} \max_{\boldsymbol{x} \geq 0} -\boldsymbol{o}^\mathsf{T}(\boldsymbol{x} - \hat{\boldsymbol{x}}) + \mathbb{E}_F[\max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \boldsymbol{\xi}^\mathsf{T}\boldsymbol{y} - \max_{\boldsymbol{y}_1 \in \mathcal{Y}(\hat{\boldsymbol{x}})} \boldsymbol{\xi}^\mathsf{T}\boldsymbol{y}_1)] \\ &\leq \sup_{F \in \mathcal{D}(\mathbb{R}^m, \boldsymbol{\mu}, \boldsymbol{\Sigma})} \max_{\boldsymbol{x} \geq 0} -\boldsymbol{o}^\mathsf{T}(\boldsymbol{x} - \hat{\boldsymbol{x}}) + \mathbb{E}_F[\max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \boldsymbol{\xi}^\mathsf{T}(\boldsymbol{y} - \hat{\boldsymbol{y}})] \\ &\leq \sup_{F \in \mathcal{D}(\mathbb{R}^m, \boldsymbol{\mu}, \boldsymbol{\Sigma})} \mathbb{E}_F[\max_{\boldsymbol{x} \geq 0,\ \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} -\boldsymbol{o}^\mathsf{T}(\boldsymbol{x} - \hat{\boldsymbol{x}}) + \boldsymbol{\xi}^\mathsf{T}(\boldsymbol{y} - \hat{\boldsymbol{y}})] \\ &= \sup_{F_{\tilde{\boldsymbol{p}}} \in \mathcal{D}(\mathcal{S}_{\tilde{\boldsymbol{p}}}, \hat{\boldsymbol{p}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{p}}})} \mathbb{E}_{F_{\tilde{\boldsymbol{p}}}}[\max_{\boldsymbol{x} \geq 0,\ [\boldsymbol{w}^\mathsf{T},\ \boldsymbol{0}^\mathsf{T},\ \boldsymbol{0}^\mathsf{T}, \boldsymbol{u}^\mathsf{T}] \in \mathcal{Y}(\boldsymbol{x})} -\boldsymbol{o}^\mathsf{T}(\boldsymbol{x} - \hat{\boldsymbol{x}}) + \tilde{\boldsymbol{p}}^\mathsf{T}(\boldsymbol{w} - \hat{\boldsymbol{w}})] \\ &\leq \max_{\boldsymbol{x} \geq 0,\ [\boldsymbol{w}^\mathsf{T},\ \boldsymbol{0}^\mathsf{T},\ \boldsymbol{0}^\mathsf{T}, \boldsymbol{u}^\mathsf{T}]^\mathsf{T} \in \mathcal{Y}(\boldsymbol{x})} -\boldsymbol{o}^\mathsf{T}(\boldsymbol{x} - \hat{\boldsymbol{x}}) + \sup_{F_{\tilde{\boldsymbol{p}}} \in \mathcal{D}(\mathcal{S}_{\tilde{\boldsymbol{p}}}, \hat{\boldsymbol{p}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{p}}})} \mathbb{E}_{F_{\tilde{\boldsymbol{p}}}}[\max_{\boldsymbol{w} \in \mathcal{W}} \tilde{\boldsymbol{p}}^\mathsf{T}(\boldsymbol{w} - \hat{\boldsymbol{w}})] \ , \end{aligned}$$

where $\mathcal{W} = \{\boldsymbol{w} | w_i^k \in \{0, 1\},\ \sum_k w_i^k = 1\}$. Here, we used the fact the $\boldsymbol{o} \leq \tilde{\boldsymbol{c}}$ with probability one so that, in the wait and see formulation, one is always better off paying the original

ownership cost for the fleet. Thus,

$$\begin{aligned}
\text{VSM}(\hat{\boldsymbol{x}}) &\leq \nu_0 + \sup_{F_{\tilde{\boldsymbol{p}}} \in \mathcal{D}(\mathcal{S}_{\tilde{\boldsymbol{p}}}, \hat{\boldsymbol{p}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{p}}})} \mathbb{E}_{\tilde{\boldsymbol{p}}}[\max_{\boldsymbol{w} \in \mathcal{W}} \tilde{\boldsymbol{p}}^{\mathsf{T}}(\boldsymbol{w} - \hat{\boldsymbol{w}})] \\
&\leq \nu_0 + \inf_{\boldsymbol{q}, \boldsymbol{Q} \succeq 0} \sup_{\boldsymbol{p}, \boldsymbol{w} \in \mathcal{W}} (\hat{\boldsymbol{p}} + \boldsymbol{p})^{\mathsf{T}}(\boldsymbol{w} - \hat{\boldsymbol{w}}) - \boldsymbol{p}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{p} - \boldsymbol{p}^{\mathsf{T}}\boldsymbol{q} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{p}}} \bullet \boldsymbol{Q} \\
&\leq \nu_0 + \sum_i \inf_{\boldsymbol{q}_i, \boldsymbol{Q}_i \succeq 0} \sup_{\boldsymbol{p}_i \in \mathbb{R}^K, k \in \{1,2,\ldots,K\}} (\hat{\boldsymbol{p}}_i + \boldsymbol{p}_i)^{\mathsf{T}}(\boldsymbol{e}_k - \hat{\boldsymbol{w}}_i) - \boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{Q}_i\boldsymbol{p}_i - \boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{q}_i + \boldsymbol{S}_i \bullet \boldsymbol{Q}_i \\
&= \nu_0 + \sum_i \nu_i \ ,
\end{aligned}$$

where we first relaxed the support of F to $\mathbb{R}^d$, then used duality of the semi-infinite linear program, and finally reduced the set of feasible $\boldsymbol{Q}$. □

For simplicity, in what follows we consider that uncertainty is limited to the profit, $\tilde{\boldsymbol{p}}$, obtained from using the different aircraft to fly passengers of different flights.

6 Numerical Experiments

Our experiments involve data from models of three airlines. Specifically, the three test cases have the structure described in Table 2.

Test Case	Types of aircraft	Number of flights	Second stage decision variables	Second stage constraints
#1	3	84	3270	3107
#2	4	240	11781	11065
#3	6	535	22016	20238

Table 2: Comparison of model sizes for three instances of a fleet mix optimization problem.

For each of these test cases, we have the complete description of the flight schedule and we are interested in the task of choosing a fleet mix that can accommodate future demand optimally. In our experiments, we assume that it was impossible to adjust the fleet composition once demand is observed. We also consider that information about the distribution of future profits takes the form of a set of demand scenarios for each flight. Specifically, we simulate profit levels achieved under nine scenarios of demand: specifically, scenarios where demand differed by 0%, ±3%, ±5%, ±10%, and ±20% from the expected value for each flight. The demand is assumed independent between flights. This scenario based model obviously constitutes a naïve representation of the uncertainty related to the future realization of demand. Yet, we believe that they represent well some of the types of information easily available for an initial analysis of the benefits of moving forward with the design of a sophisticated stochastic model. In our analysis, we assume that they can be used to provide relatively good estimates of the mean and covariance matrix of the random vector of profits $\tilde{\boldsymbol{p}}$. We also assume that the support of this random vector of profits can be

approximated using the 90% confidence region of a normal random vector that matches these moments. These characteristics are the only ones necessary to evaluate the VSM. Next, we describe in details the two solutions methods that were used to shed light on the need of a refined stochastic model.

Naïve Stochastic Programming (NSP): This approach assumes that each scenario is equally likely to occur and searches for the fleet mix with highest total expected profit. When implementing this approach, because the distribution is defined on a large outcome space, we choose to solve the sample average approximation with 100 samples. We also choose to simplify computations by applying an analytic center cutting plane algorithm to a modified version of this problem where integrality constraints have been relaxed. The solution of this relaxed form is then used as is to generate an optimistic view of what might be achieved by a stochastic programming approach.

mean Value Problem (MVP): This approach strictly use the estimated expected profits. It assumes that these estimates are exact and solves the mean value problem for the stochastic program. The MVP solution can be obtained by solving an integer linear program of reasonable size and we know, based on Corollary 5.1, that it is robust both with respect to the form of the distribution and to the amount of correlation between the profit that will be achieved for each flight.

In what follows, we compare the application of the two methods. Their implementations used the commercial software CPLEX 12.2 for mixed-integer linear programming, and CVX [23] as a modeling language for semi-definite programming, which interfaces SDPT3 by [37]. We present in Table 3 the running time observed when applying the two approaches and measuring our proposed *VSM* bounds[3].

Test Case	MVP	NSP	$\mathcal{LB}$	$\mathcal{UB}_1$	$\mathcal{UB}_2$
#1	0.6 sec	3 min	1h	22 sec	12 sec
#2	1 sec	10 min	18h	6 min	40 sec
#3	5 sec	21 h	$>$ 48h	2 h	2 min

Table 3: Comparison of computation times for the different methods on three real instances of a fleet mix optimization problem. While all computations were done on a machine with 16 processors Intel Xeon CPU (64 bits), 2.93 GHz, 63 GB memory, only the algorithm for $\mathcal{LB}$ was parallelized over 8 CPUs.

Already from a computational perspective, the mean value problem associtated to the fleet mix problem is more attractive than the stochastic programming version. Indeed, state of the art softwares are readily available for solving integer programs such as the MVP problem. Although the sample average approximation can also be expressed as an equivalent deterministic integer linear program, the number of decision variables and constraints in

[3]Note that when determining $\mathcal{LB}$ for test case #3, the algorithm was interrupted after running for two days since it still had not improved on the lower bound obtained with the NSP problem.

this representation increases linearly with the number of scenarios. This quickly renders the model infeasible to solve. In fact, Table 3 shows that the approximate stochastic programming form is already computationally demanding. The table also shows us that measuring the proposed upper bounds for the VSM does scale well with the size of the problem. On the other hand, the efforts needed to compute our lower bound can become limiting for large problems due to the need to perform global optimization. Yet, the value of this lower bound mostly resides in its ability to identify a distribution in $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a fleet mix that together would be preferred to the MVP solution. In what follows, we will discuss the significance of the more tractable $\mathcal{UB}$ value.

Test Case	Relative "robust" expected profit		Relative average profit on NSP scenarios		Interval of relative *VSM*
	MVP	NSP	MVP	NSP	$[\mathcal{LB}, \min(\mathcal{UB}_1, \mathcal{UB}_2)]$
#1	0%	-0.2%	+0.040%	+0.041%	[0.02%, min(17%, 6%)]
#2	0%	0%	+0.01%	+0.01%	[0.002%, min(14%, 1%)]
#3	0%	-0.2%	+0.159%	+0.162%	[0.003%, min(46%, 7%)]

Table 4: Comparison between performance of the MVP solution and the NSP solution measured on instances of a fleet mix optimization problem. Each percentage expresses the relative value with respect to the optimal value of the MVP problem; i.e., the "robust" expected profit estimate for the MVP solution.

Table 4 compares the performance of MVP to the NSP approach. We choose to present the performances in relative term compared to the optimal total profit computed based on the MVP model. This is done both for ease of comparison and in order to preserve the anonymity of our test cases. Coincidently, in test case #2, the MVP problem leads to the same solution as the NSP approach while in test cases #1 and #3 the MVP and NSP solutions only differed by one or two aircraft. Incidentally, we see that, in test case #1, the MVP solution trades off 0.001% in average profit over the scenarios for 0.2% more expected profit under the worst-case distribution. These observations seems to indicate that there is not much to gain in developing a thorough stochastic model. However, because these observations are based on a naïve stochastic program, they could not serve as a formal argument. Evaluating the *VSM* allows us to confirm these observations more rigorously. The last column of Table 4 presents the intervals in which we can guarantee that the *VSM* lies for each test case, under the hypothesis that the estimated mean and covariance statistics are accurate. Taking a closer look at test case #1, the computed $\mathcal{UB}_2$ value says that it is impossible to improve the total expected profits by more than 6% of the current estimation of total profits. Indeed, we should conclude that it would be wasteful to invest more than 6% of our estimation of total profits on information that might help resolve the nature of this distribution. Given that developing an accurate definition for the joint distribution of demand for the 84 different flights is likely to constitute an expensive venture, our bound on *VSM* provides a tangible argument for simply implementing the MVP solution, especially considering that such an investment could be redirected toward a project that have a potential for higher returns. Note also that given the size of the gap between minimum upper bound and best lower bound, it is unlikely that a

gain as large as 6% can even be achieved under the most appealing distribution that satisfies our assumed characteristics. Based on Table 4, similar conclusions can be made for test cases #2 and #3. Overall, we believe that in these three test cases the MVP solution draws a lot of strength from the availability of recourse actions, which is enough, in a risk neutral setting, to protect the decision maker from the uncertainty of future demand. Finally, we leave as compelling subject for future work the challenge of developping numerical tools that might reduce the observed gap between lower and upper bound for the *VSM*.

Remark 6.1. *We need to mention that the message carried by our numerical results is in contradiction with the experiments presented in [26]. Indeed, in [26] the authors develop a stochastic model for a similar stochastic fleet mix optimization problem. Yet, their experiments instead provide evidence that significant gains can be achieved through the application of a fully developed stochastic program: actually around 15% in profit for the case they consider. Therefore, it does not appear to be generally the case that in fleet mix optimization problems it is always wasteful to develop a serious stochastic model of demand. Although we did not have access to the case studied in this experiment, we conjecture that measuring the VSM bounds would surely have encouraged more strongly the development of a stochastic model in this case. These observations further emphasize the need for tools like the VSM measure that can help identify such opportunities.*

7 Conclusion

In this work, we studied the value of stochastic modeling in a two stage stochastic linear programming problem with cost uncertainty. We first demonstrated that there are a variety of contexts in which the solution to the popular MVP problem is actually robust with respect to the current available knowledge of the distributions involved in the problem at hand. Section 3 also provided intuition about the type of distribution information that might contribute to improving the quality of the decision. For instance, Proposition 3.1 implied that finding out how concentrated a distribution is around its mean would not contribute strongly to improving the quality of the robust decision. Given that one considers investing resources to identify the distribution more precisely, the upper bounds on the *VSM* proposed in Theorem 4.9 and 5.3 can help quantify how much may potentially be gained by doing so. In our numerical experiments, we observed that for three fleet mix optimization problems, it would actually be wasteful to invest more than 7% of current expected revenues in performing a thorough market study of future flight demand. Although these number might represent potential gains of millions of dollars for a big airline company, one should consider that they are based on a **best-case scenario** analysis. Instead of undertaking a set of market studies that might help identify the distribution of future demande, it is important to verify whether these funds might be invested instead in an alternate project where the returns are more secure. Finally, our numerical results do seem to indicate that there is room for the development of tighter *VSM* bounds which we leave as a topic for future research.

A Measures in $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$ Are Less Dispersed

Lemma A.1. *Given a distributional set $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$ with $\mathcal{S}$ convex and a random vector $\boldsymbol{\xi}$ such that its distribution $F \in \mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$, for all $0 \leq \alpha \leq 1$ the random vector $\boldsymbol{\zeta} := \alpha(\boldsymbol{\xi} - \boldsymbol{\mu}) + \boldsymbol{\mu}$ also has a distribution that lies in $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$.*

Proof: We simply need to verify systematically the conditions imposed on the members of $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$. First,

$$\mathbb{P}_F(\boldsymbol{\zeta} \in \mathcal{S}) = \mathbb{P}_F(\alpha\boldsymbol{\xi} + (1-\alpha)\boldsymbol{\mu} \in \mathcal{S}) = 1 \, ,$$

since $\mathbb{P}_F(\boldsymbol{\xi} \in \mathcal{S}) = 1$, $\boldsymbol{\mu} \in \mathcal{S}$, and the fact that $\mathcal{S}$ is a convex set. Second, it is easy to see that

$$\mathbb{E}_F[\boldsymbol{\zeta}] = \mathbb{E}_F[\alpha\boldsymbol{\xi} + (1-\alpha)\boldsymbol{\mu}] = \boldsymbol{\mu} \, .$$

Finally, for any $\psi(\cdot) \in \Psi$, we have that

$$\begin{array}{rcccl} \mathbb{E}_F[\psi(\alpha\boldsymbol{\xi} + (1-\alpha)\boldsymbol{\mu})] & \leq & \mathbb{E}_F[\alpha\psi(\boldsymbol{\xi}) + (1-\alpha)\boldsymbol{\mu}] & = & \alpha\mathbb{E}_F[\psi(\boldsymbol{\xi})] + (1-\alpha)\psi(\boldsymbol{\mu}) \\ & \leq & \mathbb{E}_F[\psi(\boldsymbol{\xi})] & \leq & 0 \, , \end{array}$$

where we applied Jensen's inequality for the two first upper bounding steps. □

B Concavity of $h(\boldsymbol{x}, \xi)$ in ξ

Consider the second stage stochastic minimization problem where we

$$\begin{array}{rll} h(\boldsymbol{x}, \xi) := & \text{minimize}_{\boldsymbol{y}} & f(\boldsymbol{x}, \boldsymbol{y}, \xi) \\ & \text{s.t.} & \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}), \end{array}$$

where $\mathcal{Y}(\boldsymbol{x})$ is any given closed and bounded set function of $\boldsymbol{x}$. Here we make no assumption that set $\mathcal{Y}(\boldsymbol{x})$ is a polyhedral or even a convex set.

We assume that $f(\boldsymbol{x}, \boldsymbol{y}, \xi)$ is a concave function in the uncertain parameter-vector ξ. Note that this assumption has been made in most previous stochastic and robust optimization models (see [1]).

Lemma B.1. *The minimum value function $h(\boldsymbol{x}, \xi)$ is a concave function in ξ.*

This lemma implies that our Proposition 3.1 is also applicable to most linear and nonlinear two-stage optimization problems.

Proof: Since $\mathcal{Y}(\boldsymbol{x})$ is closed and bounded, for any given $\boldsymbol{x}$ and ξ, the above problem has a minimizer and $h(\boldsymbol{x}, \xi)$ attains a finite value.

Consider three possible objective functions: $f(\boldsymbol{x}, \boldsymbol{y}, \xi^1)$, $f(\boldsymbol{x}, \boldsymbol{y}, \xi^2)$, and $f(\boldsymbol{x}, \boldsymbol{y}, \alpha\xi^1 + (1-\alpha)\xi^2)$, where $0 \leq \alpha \leq 1$. Let $\boldsymbol{y}^1$, $\boldsymbol{y}^2$ and $\boldsymbol{y}^*$ be the minimizers respectively corresponding to the three objective functions in the problem for a given $\boldsymbol{x}$.

Then,

$$h(\boldsymbol{x}, \alpha\xi^1 + (1-\alpha)\xi^2) = f(\boldsymbol{x}, \boldsymbol{y}^*, \alpha\xi^1_(1-\alpha)\xi^2) \geq \alpha f(\boldsymbol{x}, \boldsymbol{y}^*, \xi^1) + (1-\alpha) f(\boldsymbol{x}, \boldsymbol{y}^*, \xi^2),$$

from the concavity of f in ξ.

However,

$$f(\boldsymbol{x}, \boldsymbol{y}^*, \xi^1) \geq f(\boldsymbol{x}, \boldsymbol{y}^1, \xi^1) \quad \text{and} \quad f(\boldsymbol{x}, \boldsymbol{y}^*, \xi^1) \geq f(\boldsymbol{x}, \boldsymbol{y}^2, \xi^2).$$

This is because that $\boldsymbol{y}^*$ is a feasible solution for the two respective problems. Thus,

$$h(\boldsymbol{x}, \alpha\xi^1 + (1-\alpha)\xi^2) \geq \alpha f(\boldsymbol{x}, \boldsymbol{y}^*, \xi^1) + (1-\alpha) f(\boldsymbol{x}, \boldsymbol{y}^*, \xi^2)$$

$$\geq \alpha f(\boldsymbol{x}, \boldsymbol{y}^1, \xi^1) + (1-\alpha) f(\boldsymbol{x}, \boldsymbol{y}^2, \xi^2) = \alpha h(\boldsymbol{x}, \xi^1) + (1-\alpha) h(\boldsymbol{x}, \xi^2),$$

that is, h is a concave function in ξ.

C Proof of Proposition 3.2

We follow similar steps as followed in the proof of Proposition 3.1. We first underline the fact that implementing the MVP solution, a policy that does not adapt to the sequence of observable ξ, leads to a worst-case expected cost that is equal to the optimal value of the MVP problem, which we refer to as V_{MVP}.

$$\begin{aligned} \sup_{F \in \mathcal{D}(\boldsymbol{\mu}_{[1:T]})} \mathbb{E}_F \left[\sum_{t=1}^{T} \xi_t^\mathsf{T} C_t \bar{\boldsymbol{x}}_t \right] &= \sup_{F \in \mathcal{D}(\boldsymbol{\mu}_{[1:T]})} \sum_{t=1}^{T} \mathbb{E}_F[\xi_t^\mathsf{T}] C_t \bar{\boldsymbol{x}}_t \\ &= \sum_{t=1}^{T} \boldsymbol{\mu}_t^\mathsf{T} C_t \bar{\boldsymbol{x}}_t, \end{aligned}$$

where $\mathcal{D}(\boldsymbol{\mu}_{[1:T]})$ is short for $\mathcal{D}(\mathcal{S}, [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_T], \Psi)$. Here, we used the fact that $\bar{x}_t$ does not adapt to the observed uncertain parameters and the fact that all the distributions in the set that is considered lead to the same expected value for the random vectors. We can therefore say that the optimal value of the distributionally robust multi-stage stochastic program must be smaller than V_{MVP}.

Secondly, after verifying that the Dirac measure $\delta_{\boldsymbol{\mu}_{[1:T]}}$ lies in the set $\mathcal{D}(\boldsymbol{\mu}_{[1:T]})$ (the argument being the same as in the proof of Proposition 3.1), one can show that V_{MVP} is actually also a lower bound for the same distributionally robust problem.

$$\begin{aligned} \min_{(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T) \in \mathcal{X} a.s.} \sup_{F \in \mathcal{D}(\boldsymbol{\mu}_{[1:T]})} \mathbb{E}_F \left[\sum_{t=1}^{T} \xi_t^\mathsf{T} C_t \boldsymbol{x}_t(\xi_{[1:t]}) \right] &\geq \min_{(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T) \in \mathcal{X} a.s.} \mathbb{E}_{\delta_{\boldsymbol{\mu}_{[1:T]}}} \left[\sum_{t=1}^{T} \xi_t^\mathsf{T} C_t \boldsymbol{x}_t(\xi_{[1:t]}) \right] \\ &= \min_{(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T) \in \mathcal{X} a.s.} \sum_{t=1}^{T} \boldsymbol{\mu}_t^\mathsf{T} C_t \boldsymbol{x}_t(\boldsymbol{\mu}_{[1:t]}) \\ &= V_{\text{MVP}}\,. \end{aligned}$$

Hence, we conclude that the MVP solution is an optimal solution for the distributionally robust multi-stage stochastic program under $\mathcal{D}(\boldsymbol{\mu}_{[1:T]})$. □

D Proof of Proposition 3.7

The proof consists of showing that the sequence of distributions $F_1, F_2, ...$ with

$$F_k(\boldsymbol{\xi}) = (1-(1+\beta_k)^{-1})\delta_{\boldsymbol{\mu}}(\boldsymbol{\xi}) + (1+\beta_k)^{-1}G_k(\boldsymbol{\xi})\,,$$

satisfies $\mathbb{E}_{F_k}[\psi(\boldsymbol{\xi})] = 0\ ,\ \forall\,\psi \in \Psi, \forall k$ and that given any $\epsilon > 0$, one can choose k large enough so that F_k satisfies:

$$\begin{aligned}\mathbb{P}_{F_k}(\boldsymbol{\xi} \in \mathcal{S}) &\geq 1-\epsilon \\ \|\mathbb{E}_{F_k}[\boldsymbol{\xi}] - \boldsymbol{\mu}\| &\leq \epsilon\,.\end{aligned}$$

Based on Condition 1, we can easily show the first part:

$$\begin{aligned}\mathbb{E}_{F_k}[\psi(\boldsymbol{\xi})] &= (1-(1+\beta_k)^{-1})\mathbb{E}_{\delta_{\boldsymbol{\mu}}}[\psi(\boldsymbol{\xi})] + (1+\beta_k)^{-1}\mathbb{E}_{G_k}[\psi(\boldsymbol{\xi})] \\ &= (1-(1+\beta_k)^{-1})\psi(\boldsymbol{\mu}) + (1+\beta_k)^{-1}\beta_k = 0\,.\end{aligned}$$

Based on Condition 2, we also have the property that $\mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|] = O(\beta_k^{1/\gamma})$, for some $\gamma > 1$. This is due to the fact that there exists an $a > 0$, $b \in \mathbb{R}$, and $\gamma > 1$:

$$\begin{aligned}\mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|]^{\gamma} &\leq \mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|^{\gamma}] = (1/a)(\mathbb{E}_{G_k}[b + a\|\boldsymbol{\xi}\|^{\gamma}] - b) \\ &\leq (1/a)(\mathbb{E}_{G_k}[\psi_0(\boldsymbol{\xi})] - b) = (1/a)(\beta_k - b)\,,\end{aligned}$$

where we used Jensen's inequality and the fact that $\psi_0(\boldsymbol{\xi}) = \sum_i \theta_i\psi_i(\boldsymbol{\xi})$ for some conical combination of $\psi_i \in \Psi$ allowing us to derive

$$\mathbb{E}_{G_k}[\psi_0(\boldsymbol{\xi})] = \mathbb{E}_{G_k}\left[\sum_i \theta_i\psi_i(\boldsymbol{\xi})\right] = \sum_i \theta_i\beta_k = \beta_k\,.$$

Hence, we have that $\mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|] = \mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|]^{\gamma/\gamma} \leq ((1/a)(\beta_k - b))^{1/\gamma} = O(\beta_k^{1/\gamma})$.

Thus, we can demonstrate that both $\mathbb{P}_{F_k}(\boldsymbol{\xi} \in \mathcal{S})$ and $\mathbb{E}_{F_k}[\boldsymbol{\xi}]$ will become feasible for k large enough:

$$\begin{aligned}\mathbb{P}_{F_k}(\boldsymbol{\xi} \in \mathcal{S}) &= (1-(1+\beta_k)^{-1})\mathbb{P}_{\delta_\mu}(\boldsymbol{\xi} \in \mathcal{S}) + (1+\beta_k)^{-1})\mathbb{P}_{G_k}(\boldsymbol{\xi} \in \mathcal{S}) \\ &\geq (1-(1+\beta_k)^{-1})\mathbb{P}_{\delta_\mu}(\boldsymbol{\xi} \in \mathcal{S}) = 1-(1+\beta_k)^{-1}\end{aligned}$$

$$\begin{aligned}\|\mathbb{E}_{F_k}[\boldsymbol{\xi}] - \boldsymbol{\mu}\| &\leq (1-(1+\beta_k)^{-1})\|\mathbb{E}_{\delta_{\boldsymbol{\mu}}}[\boldsymbol{\xi}] - \boldsymbol{\mu}\| + (1+\beta_k)^{-1}\|\mathbb{E}_{G_k}[\boldsymbol{\xi}] - \boldsymbol{\mu}\| \\ &\leq (1+\beta_k)^{-1}(\|\mathbb{E}_{G_k}[\boldsymbol{\xi}]\| + \|\boldsymbol{\mu}\|) \leq (1+\beta_k)^{-1}(\mathbb{E}_{G_k}[\|\boldsymbol{\xi}\|] + \|\boldsymbol{\mu}\|) = O(\beta_k^{-(1-1/\gamma)})\,.\end{aligned}$$

Ultimately, based on the Lipschitz property of $h(x,\cdot)$ demonstrated in Lemma 3.4, we

can verify that

$$\begin{aligned}
\sup_{F\in\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi,\epsilon)} \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})] &\geq \sup_{\{k|F_k\in\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi,\epsilon)\}} \mathbb{E}_{F_k}[h(\boldsymbol{x},\boldsymbol{\xi})] \\
&= \sup_{\{k|F_k\in\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi,\epsilon)\}} (1-(1+\beta_k)^{-1})h(\boldsymbol{x},\boldsymbol{\mu}) + (1+\beta_k)^{-1}\mathbb{E}_{G_k}[h(\boldsymbol{x},\boldsymbol{\xi})] \\
&\geq \sup_{\{k|F_k\in\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi,\epsilon)\}} (1-(1+\beta_k)^{-1})h(\boldsymbol{x},\boldsymbol{\mu}) + (1+\beta_k)^{-1}\mathbb{E}_{G_k}[h(\boldsymbol{x},\boldsymbol{\mu}) - R\|\boldsymbol{C}_2\|\|\boldsymbol{\xi}-\boldsymbol{\mu}\|] \\
&\geq \sup_{\{k|F_k\in\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi,\epsilon)\}} h(\boldsymbol{x},\boldsymbol{\mu}) - (1+\beta_k)^{-1}R\|\boldsymbol{C}_2\|\mathbb{E}_{G_k}[\|\boldsymbol{\xi}\| + \|\boldsymbol{\mu}\|] \\
&= \sup_{\{k|F_k\in\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi,\epsilon)\}} h(\boldsymbol{x},\boldsymbol{\mu}) - (1+\beta_k)^{-1}R\|\boldsymbol{C}_2\|\|\boldsymbol{\mu}\| - (1+\beta_k)^{-1}R\|\boldsymbol{C}_2\|\, O(\beta_k^{1/\gamma}) \\
&= h(\boldsymbol{x},\boldsymbol{\mu})\ .
\end{aligned}$$

We can then resolve an approximate upper bound using Proposition 3.1 and the Lipschitz property of $h(\boldsymbol{x},\cdot)$:

$$\begin{aligned}
\sup_{F\in\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi,\epsilon)} \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})] &\leq \sup_{\{\boldsymbol{z}|\|\boldsymbol{z}-\boldsymbol{\mu}\|\leq\epsilon\}} \sup_{F\in\mathcal{D}(\mathbb{R}^d,\boldsymbol{z},\Psi)} \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})] \\
&\leq \sup_{\{\boldsymbol{z}|\|\boldsymbol{z}-\boldsymbol{\mu}\|\leq\epsilon\}} h(\boldsymbol{x},\boldsymbol{z}) \ \leq\ h(\boldsymbol{x},\boldsymbol{\mu}) + O(\epsilon)\ .
\end{aligned}$$

We are left with demonstrating near-optimality of the MVP solution. This is easily done using the following argument, where for conciseness we let $\boldsymbol{x}_{\text{MVP}}$ refer to the optimal solution of the MVP problem, $g(\boldsymbol{x})$ be equal to $\sup_{F\in\bar{\mathcal{D}}(\mathcal{S},\boldsymbol{\mu},\Psi,\epsilon)} \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} + \mathbb{E}_F[h(\boldsymbol{x},\boldsymbol{\xi})]$, and $\boldsymbol{x}_g$ be the member of $\mathcal{X}$ that minimizes $g(\boldsymbol{x})$.

$$\begin{aligned}
g(\boldsymbol{x}_{\text{MVP}}) - g(\boldsymbol{x}_g) &\leq \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x}_{\text{MVP}} + h(\boldsymbol{x}_{\text{MVP}},\boldsymbol{\mu}) + O(\epsilon) - \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x}_g - h(\boldsymbol{x}_g,\boldsymbol{\mu}) \\
&= \min_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} + h(\boldsymbol{x},\boldsymbol{\mu}) - (\boldsymbol{c}_1^\mathsf{T}\boldsymbol{x}_g + h(\boldsymbol{x}_g,\boldsymbol{\mu})) + O(\epsilon) \ \leq\ O(\epsilon)\ ,
\end{aligned}$$

so that $g(\boldsymbol{x}_{\text{MVP}}) \leq g(\boldsymbol{x}_g) + O(\epsilon)$. □

E Proof of Lemma 3.9

We first re-parametrize as $z(\Delta,\delta) := 2\boldsymbol{\mu} + \delta\Delta$. We therefore need to show that for any $\Delta \in \mathbb{R}^d$

$$g(\delta) := \|\boldsymbol{\mu} + \delta\Delta\|_\alpha^\gamma - \frac{1}{2^{\gamma-1}}\|2\boldsymbol{\mu} + \delta\Delta\|_\alpha^\gamma + \|\boldsymbol{\mu}\|_\alpha^\gamma \geq 0\ ,\ \forall\ \delta\in\mathbb{R}\ .$$

This is done by showing that $g(\delta)$ is decreasing for negative δ's, achieves the value of zero at $\delta = 0$ and is increasing for positive δ's. The function $g(\delta)$ therefore achieves its minimum value of zero at $\delta = 0$ for any Δ. The simplest step is to show that $g(0) = 0$.

$$g(0) = \|\boldsymbol{\mu}\|_\alpha^\gamma - \frac{1}{2^{\gamma-1}}\|2\boldsymbol{\mu}\|_\alpha^\gamma + \|\boldsymbol{\mu}\|_\alpha^\gamma = 0\ .$$

The second step requires us to realize that $g(\delta)$ can be represented in terms of the convex function $g_2(\delta) = \|\mu + \delta\Delta\|_\alpha^\gamma$:

$$g(\delta) = g_2(\delta) - 2g_2(\delta/2) + \|\mu\|_\alpha^\gamma \, .$$

One can verify that $g_2(\delta)$ is convex using the fact that it is the composition of a function y^γ, which is convex and increasing over $y \geq 0$ for $\gamma \geq 1$, and of a convex function $\|\mu + \delta\Delta\|_\alpha$. The convexity of $g_2(\delta)$ tells us that $g_2'(\delta) \geq g_2'(\delta/2)$ for $\delta \geq 0$ and that $g_2'(\delta) \leq g_2'(\delta/2)$ for $\delta \leq 0$. Thus, we can easily verify the properties of the derivative of $g(\delta)$. While for $\delta \geq 0$, we have $g'(\delta) = g_2'(\delta) - g_2'(\delta/2) \geq 0$, we can also easily show that, for $\delta \leq 0$, we have $g'(\delta) = g_2'(\delta) - g_2'(\delta/2) \leq 0$. This completes our proof. □

F Proof of Lemma 4.1

To make this demonstration, we use the fact that $h(x_2, \boldsymbol{\xi}) = x_2 h(1, \boldsymbol{\xi})$. First, in the case where $x_2 = 0$, we already verified that $h(0, \boldsymbol{\xi}) = 0$. When $x_2 > 0$, one can easily show that $h(x_2, \boldsymbol{\xi}) = x_2 h(1, \boldsymbol{\xi})$ for all $\boldsymbol{\xi} \in \mathbb{R}^d$ by replacing the inner variable $\boldsymbol{y}$ by $\boldsymbol{z} = \boldsymbol{y}/x_2$. Thus, for all $x_2 \geq 0$, the objective of problem (6) reduces to

$$\begin{aligned} \sup_{F \in \mathcal{D}(\mathbb{R}^d, \mathbf{0}_d, \mathbf{I})} \{\mathbb{E}_F[-cx_2 - h(x_2, \boldsymbol{\xi})]\} &= \sup_{F \in \mathcal{D}(\mathbb{R}^d, \mathbf{0}_d, \mathbf{I})} \{\mathbb{E}_F[(-c - h(1, \boldsymbol{\xi}))x_2]\} \\ &= \sup_{F \in \mathcal{D}(\mathbb{R}^d, \mathbf{0}_d, \mathbf{I})} \{\mathbb{E}_F[(-c - h(1, \boldsymbol{\xi}))]\} \, x_2 \, . \quad \square \end{aligned}$$

G Proof of Lemma 4.2

Based on Lemma 1 of [15], considering that $h(1, \boldsymbol{\xi})$ is F-integrable for all $F \in \mathcal{D}(\mathbb{R}^d, \mathbf{0}_d, \mathbf{I})$ since

$$|\mathbb{E}_F[h(1, \boldsymbol{\xi})]| \leq \mathbb{E}_F[\|\boldsymbol{\xi}\|_1] \leq \sqrt{d}\mathbb{E}_F[\|\boldsymbol{\xi}\|_2] \leq \sqrt{d}\sqrt{\mathbb{E}_F[\|\boldsymbol{\xi}\|_2^2]} = d \, ,$$

by duality theory we can say that evaluating the supremum of such an expression is equivalent to finding the optimal value of

$$\begin{aligned} \underset{\boldsymbol{Q}, \boldsymbol{q}, t}{\text{minimize}} \quad & t + \mathbf{I} \bullet \boldsymbol{Q} \\ \text{subject to} \quad & t \geq \max_{\boldsymbol{y} \in \mathcal{Y}(1)} -c - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{y} - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{Q}\boldsymbol{\xi} - \boldsymbol{q}^\mathsf{T}\boldsymbol{\xi} \, , \forall \, \boldsymbol{\xi} \in \mathbb{R}^d \, , \end{aligned}$$

where $\mathcal{Y}(1) = \{\boldsymbol{y} \in \mathbb{R}^d | \boldsymbol{a}^\mathsf{T}\boldsymbol{y} = 0 \ \& \ -1 \leq y_i \leq 1 \, , \forall \, i\}$. This is necessarily a convex optimization problem with linear objective function since each constraint is jointly convex in $\boldsymbol{Q}$, $\boldsymbol{q}$, and t. We now show that the separation problem associated with the only constraint of this problem is NP-hard. Thus, by the equivalence of optimization and separation (see [24]) finding the optimal value of this problem can be shown to be NP-hard.

Consider separating the solution $\boldsymbol{Q} = (1/4)\mathbf{I}$, $\boldsymbol{q} = \mathbf{0}_m$, and $t = d - c$ from the feasible set. In this case, we must be able to verify its feasibility with respect to the only constraint. This can be shown to reduce to verifying if

$$\sup_{\boldsymbol{\xi}\in\mathbb{R}^d,\, \boldsymbol{y}\in\mathcal{Y}(1)} \; -\boldsymbol{\xi}^\mathsf{T}\boldsymbol{y} - (1/4)\boldsymbol{\xi}^\mathsf{T}\boldsymbol{\xi} \;\leq\; d\,.$$

After solving in terms of $\boldsymbol{\xi}$, we get that we need to verify if the optimal value of the problem

$$\begin{array}{ll} \underset{\boldsymbol{y}}{\text{maximize}} & \boldsymbol{y}^\mathsf{T}\boldsymbol{y} \\ \text{subject to} & -1 \leq y_i \leq 1\ ,\ \forall\, i \in \{1,2,...,d\} \\ & \boldsymbol{a}^\mathsf{T}\boldsymbol{y} = 0\ \ , \end{array}$$

is greater or equal to d or not. This is equivalent to showing if the set defined by

$$\Upsilon(\boldsymbol{a}) = \{\ \boldsymbol{y} \in \{-1,1\}^d \mid \boldsymbol{a}^\mathsf{T}\boldsymbol{y} = 0\ \}\ ,$$

for any $\boldsymbol{a} \in \mathbb{R}^d$, is empty or not since the extreme point of the unit box are the only ones for which $\|\boldsymbol{y}\|^2 \geq d$. Finally, one can easily confirm that the NP-complete Partition problem can be reduced to verifying whether some $\Upsilon(\boldsymbol{a})$ is empty or not.

Partition Problem: Given a set of positive integers $\{s_i\}_{i=1}^d$ with index set $\mathcal{A} = \{1,2,...,d\}$ is there a partition $(\mathcal{B}_1, \mathcal{B}_2)$ of $\mathcal{A}$ such that $\sum_{i\in\mathcal{B}_1} s_i = \sum_{i\in\mathcal{B}_2} s_i$?

The reduction is obtained by casting $a_i := s_i$ for all i, and considering that a feasible solution $\boldsymbol{y} \in \Upsilon(\boldsymbol{a})$ only exists if such a partition exists and identifies the partition through $\mathcal{B}_1 = \{i \in \mathcal{A} | y_i = 1\}$. This completes our proof. □

H Proof of Proposition 4.7

We first present without proof a Lemma that describes how to approximate a concave function on the real line to any level of accuracy by containing it between an outer and an inner piecewise linear concave functions.

Lemma H.1. *Given any set of points $\{z_i\}_{k=1}^K$, a concave function $g : \mathbb{R} \to \mathbb{R}$ is contained between the two piecewise linear concave functions. Specifically, $g_{inner}(z) \leq g(z) \leq g_{outer}(z)$, where*

$$g_{outer}(z) := \min_{k\in\{1,2,...,K\}} g(z_k) + (z - z_k)g'(z_k)\ ,$$

with $g'(z_k) \in \mathbb{R}$ as any super-gradient of $g(z)$ at z_k, and

$$\begin{array}{rll} g_{inner}(z) := & \min\limits_{w,v} & w + zv \\ & \text{subject to} & w + z_k v\ \geq g(z_k)\ ,\ \forall\, k \in \{1,2,...,K\}\ . \end{array}$$

We then reduce the intractable task of evaluating $\mathrm{VSM}(\boldsymbol{x}_1)$ to the search for a distribution in the more manageable set $\mathcal{D}(\mathcal{S}_0, \boldsymbol{\mu}, \Psi)$. The analysis is also limited to the potential regret of having committed to $\boldsymbol{x}_1$ instead of $\hat{\mathcal{X}}_2$. This naturally leads to a lower bound for the VSM which considers distributions in the larger set $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \Psi)$ and any alternative decision in $\mathcal{X}_2$

$$\mathrm{VSM}(\boldsymbol{x}_1) \;\geq\; \sup_{F \in \mathcal{D}(\mathcal{S}_0, \boldsymbol{\mu}, \Psi)} \mathbb{E}_F[\boldsymbol{c}_1^\mathsf{T}(\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_2) + h(\boldsymbol{x}_1, \boldsymbol{\xi}) - h(\hat{\boldsymbol{x}}_2, \boldsymbol{\xi})]$$

We then apply duality theory to the $\sup_{F \in \mathcal{D}(\mathcal{S}_0, \boldsymbol{\mu}, \Psi)}$ operation. Considering the primal problem as a semi-infinite linear conic problem

$$\begin{array}{lll} \sup\limits_{F \in \mathcal{M}} & \mathrm{E}_F[\boldsymbol{c}_1^\mathsf{T}(\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_2) + h(\boldsymbol{x}_1, \boldsymbol{\xi}) - h(\hat{\boldsymbol{x}}_2, \boldsymbol{\xi})] & \text{(13a)} \\ \text{subject to} & \mathbb{E}_F[\mathbb{1}\{\boldsymbol{\xi} \in \mathcal{S}_0\}] = 1 & \text{(13b)} \\ & \mathbb{E}_F[\boldsymbol{\xi}] = \boldsymbol{\mu} & \text{(13c)} \\ & \mathbb{E}_F[\boldsymbol{r}^\mathsf{T}\boldsymbol{\psi}(\boldsymbol{\xi})] \leq 0 \;,\; \forall\, \boldsymbol{r} \in \mathcal{K} \;, & \text{(13d)} \end{array}$$

the dual form takes the shape

$$\begin{array}{lll} \underset{t, \boldsymbol{q}, \boldsymbol{r}}{\text{minimize}} & t & \text{(14a)} \\ \text{subject to} & t \geq \boldsymbol{c}_1^\mathsf{T}(\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_2) + h(\boldsymbol{x}_1, \boldsymbol{\xi}) - h(\hat{\boldsymbol{x}}_2, \boldsymbol{\xi}) - (\boldsymbol{\xi} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{q} - \boldsymbol{r}^\mathsf{T}\boldsymbol{\psi}(\boldsymbol{\xi}) \;,\; \forall\, \boldsymbol{\xi} \in \mathcal{S}_0 & \text{(14b)} \\ & \boldsymbol{r} \in \mathcal{K} \,, & \text{(14c)} \end{array}$$

where t, $\boldsymbol{q}$, and $\boldsymbol{r}$ are the dual variables associated respectively with constraints (13b), (13c), and (13d). One can verify that there is no duality gap between primal and dual problems using the weaker version of Proposition 3.4 in [34] and the fact that the Dirac measure $\delta_{\boldsymbol{\mu}}$ lies in the relative interior of $\mathcal{D}(\mathcal{S}_0, \boldsymbol{\mu}, \Psi)$.

Exploiting the structure of $\mathcal{S}_0 \subseteq \bigcup_{i=1}^d \{\boldsymbol{\xi} \,|\exists\, \xi \in [\nu_i - \tau_i, \nu_i + \tau_i] \,,\; \boldsymbol{\xi} = \boldsymbol{\mu} + (\xi - \mu_i)\boldsymbol{e}_i\}$, we can decompose constraint (14b) into d simpler constraints

$$t \;\geq\; \max_{\xi \in [\nu_i - \tau_i, \nu_i + \tau_i]} \; g_i(\boldsymbol{x}_1, \xi) - g_i(\hat{\boldsymbol{x}}_2, \xi) - q_i(\xi - \mu_i) - \boldsymbol{r}^\mathsf{T}\boldsymbol{\psi}(\boldsymbol{\mu} + (\xi - \mu_i)\boldsymbol{e}_i) \;\forall\, i \in \{1, 2, \ldots, d\} \qquad \text{(15)}$$

where $g_i(\boldsymbol{x}, \xi) = \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} + h(\boldsymbol{x}, \boldsymbol{\mu} + (\xi - \mu_i)\boldsymbol{e}_i)$. Note that if $S_0 = S_0 \cap \mathcal{B}_\infty(\boldsymbol{\nu}, \boldsymbol{\tau})$, then this set of constraints is entirely equivalent to the original one.

For any fixed i, we now go through a sequence of relaxation steps for constraint (15).

$$\begin{array}{rcl} t & \geq & \max\limits_{\xi \in [\nu_i - \tau_i, \nu_i + \tau_i]} \; g_i(\boldsymbol{x}_1, \xi) - g_i(\hat{\boldsymbol{x}}_2, \xi) - q_i(\xi - \mu_i) - \boldsymbol{r}^\mathsf{T}\boldsymbol{\psi}(\boldsymbol{\mu} + (\xi - \mu_i)\boldsymbol{e}_i) \\ & \geq & \max\limits_{\xi \in [\nu_i - \tau_i, \nu_i + \tau_i]} \; g_i(\boldsymbol{x}_1, \xi) - \min\limits_k \{\alpha_i^k + (\xi - \xi_i^k)\beta_i^k\} - q_i(\xi - \mu_i) - \boldsymbol{r}^\mathsf{T}\boldsymbol{\psi}(\boldsymbol{\mu} + (\xi - \mu_i)\boldsymbol{e}_i) \\ & = & \max\limits_k \; \max\limits_{\xi \in [\nu_i - \tau_i, \nu_i + \tau_i]} \; g_i(\boldsymbol{x}_1, \xi) - \alpha_i^k - (\xi - \xi_i^k)\beta_i^k - q_i(\xi - \mu_i) - \boldsymbol{r}^\mathsf{T}\boldsymbol{\psi}(\boldsymbol{\mu} + (\xi - \mu_i)\boldsymbol{e}_i) \\ & \geq & \max\limits_k \; \max\limits_{\xi \in [\nu_i - \tau_i, \nu_i + \tau_i]} \; \min\limits_{(w,v) \in \mathcal{W}_i} \{w + \xi v\} - \alpha_i^k - \beta_i^k(\xi - \xi_i^k) - q_i(\xi - \mu_i) - \boldsymbol{r}^\mathsf{T}\boldsymbol{\psi}(\boldsymbol{\mu} + (\xi - \mu_i)\boldsymbol{e}_i) \\ & = & \max\limits_k \; \min\limits_{(w,v) \in \mathcal{W}_i} \; \max\limits_{\xi \in [\nu_i - \tau_i, \nu_i + \tau_i]} w + \xi v - \alpha_i^k - \beta_i^k(\xi - \xi_i^k) - q_i(\xi - \mu_i) - \boldsymbol{r}^\mathsf{T}\boldsymbol{\psi}(\boldsymbol{\mu} + (\xi - \mu_i)\boldsymbol{e}_i) \;, \end{array}$$

where $\mathcal{W}_i$ is short for $\{(w,v) \in \mathbb{R}^2 \mid w + \xi_m v \geq g_i(\boldsymbol{x}_1, \xi_m) \forall m \in \{1,2,...,K\}\}$. The two relaxation steps are obtained through the application of Lemma H.1 to first replace $g(\boldsymbol{x}_2, \boldsymbol{\xi})$ by its outer approximation, and then $g(x_1, \boldsymbol{\xi})$ by its inner approximation. The last equality is obtained by inversing the order of $\max_{\xi \in [\nu_i - \tau_i, \nu_i + \tau_i]}$ and $\min_{(w,v) \in \mathcal{W}_i}$ using Sion's minimax theorem (see [36]). Thus, we obtain the optimization problem presented in Definition 4.5.

I Proof of Proposition 4.6

We first represent $\mathcal{D}(\mathcal{S}, \boldsymbol{\mu}, \mathbf{I})$ in the form proposed in Example 3.1, i.e., using

$$\Psi = \{\psi : \mathbb{R}^d \to \mathbb{R} \mid \exists \boldsymbol{Q} \succeq 0\,,\ \psi(\boldsymbol{\xi}) = \boldsymbol{Q} \bullet ((\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^\mathsf{T} - \mathbf{I})\}$$

It is indeed the case that this set Ψ satisfies Assumption 4.4 since the positive semi-definite cone is one for which it is relatively easy to verify feasibility using singular value decomposition. In formulating problem (8), we get that constraint (8b) takes the shape:

$$t \geq w_i^k + \xi v_i^k - \alpha_i^k - \beta_i^k(\xi - \xi_i^k) - q_i(\xi - \mu_i) - \boldsymbol{Q} \bullet ((\xi - \mu_i)^2 \boldsymbol{e}_i \boldsymbol{e}_i^\mathsf{T} - \mathbf{I})\ ,\ \begin{cases} \forall\, \xi \in [\nu_i - \tau_i, \nu_i + \tau_i] \\ \forall i \in \{1,2,...,d\} \\ \forall k \in \{1,2,...,K\} \end{cases}$$

A simple replacement of t for $t - \mathbf{I} \bullet \boldsymbol{Q}$ leads to the equivalent formulation of problem (8)

$$\begin{array}{ll} \underset{t, \boldsymbol{q}, \boldsymbol{Q}, \{\boldsymbol{w}^k, \boldsymbol{v}^k\}_{k=1}^K}{\text{minimize}} & t + \mathbf{I} \bullet \boldsymbol{Q} \\ \text{subject to} & t \geq w_i^k + \xi v_i^k - \alpha_i^k - \beta_i^k(\xi - \xi_i^k) - q_i(\xi - \mu_i) - (\xi_i - \mu_i)^2 Q_{i,i}\ ,\ \begin{cases} \forall\, \xi \in [\nu_i - \tau_i, \nu_i + \tau_i] \\ \forall i \in \{1,2,...,d\} \\ \forall k \in \{1,2,...,K\} \end{cases} \\ & w_i^k + v_i^k \xi_i^m \geq \boldsymbol{c}_1^\mathsf{T} \boldsymbol{x}_1 + h(\boldsymbol{x}_1, \boldsymbol{\mu} + (\xi_i^m - \mu_i)\boldsymbol{e}_i)\ ,\ \begin{cases} \forall\, i \in \{1,2,...,d\} \\ \forall\, m,\ k \in \{1,2,...,K\} \end{cases} \\ & \boldsymbol{Q} \succeq 0\,. \end{array}$$

Since only the diagonal terms of $\boldsymbol{Q}$ are involved in both the objective function and the constraints, one can arbitrarily set all off diagonal terms of $\boldsymbol{Q}$ to zero. Thus, we are left with

$$\begin{array}{ll} \underset{t, \boldsymbol{q}, \boldsymbol{r}, \{\boldsymbol{w}^k, \boldsymbol{v}^k\}_{k=1}^K}{\text{minimize}} & t + \boldsymbol{e}^\mathsf{T} \boldsymbol{r} \\ \text{subject to} & t \geq w_i^k + \xi v_i^k - \alpha_i^k - \beta_i^k(\xi - \xi_i^k) - q_i(\xi - \mu_i) - (\xi_i - \mu_i)^2 r_i\ ,\ \begin{cases} \forall\, \xi \in [\nu_i - \tau_i, \nu_i + \tau_i] \\ \forall i \in \{1,2,...,d\} \\ \forall k \in \{1,2,...,K\} \end{cases} \\ & w_i^k + v_i^k \xi_i^m \geq \boldsymbol{c}_1^\mathsf{T} \boldsymbol{x}_1 + h(\boldsymbol{x}_1, \boldsymbol{\mu} + (\xi_i^m - \mu_i)\boldsymbol{e}_i)\ ,\ \begin{cases} \forall\, i \in \{1,2,...,d\} \\ \forall\, m,\ k \in \{1,2,...,K\} \end{cases} \\ & \boldsymbol{r} \geq 0\,. \end{array}$$

We finally apply the S-Lemma (cf., Theorem 2.2 in [30]) to replace, for each fixed i and k, the set of constraints indexed over the interval $[\nu_i - \tau_i, \nu_i + \tau_i]$ by an equivalent linear matrix inequality:

$$\left[\begin{array}{cc} r_i & \frac{\beta_i^k - v_i^k + q_i - 2\mu_i r_i}{2} \\ \frac{\beta_i^k - v_i^k + q_i - 2\mu_i r_i}{2} & t - w_i^k + \alpha_i^k - \beta_i^k \xi_i^k - \mu_i q_i + \mu_i^2 r_i \end{array} \right] \succeq -s_i^k \left[\begin{array}{cc} 1 & -\nu_i \\ -\nu_i & \tau_i^2 \end{array} \right] \quad , \quad s_i^k \geq 0 \, .$$

Now regarding the complexity of solving this semi-definite programming problem, it is well known that in the standard form

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

the problem can be solved in $O\left(\left(\sum_{i=1}^{\tilde{K}} \tilde{m}_i \right)^{0.5} \left(\tilde{n}^2 \sum_{i=1}^{\tilde{K}} \tilde{m}_i^2 + \tilde{n} \sum_{i=1}^{\tilde{K}} \tilde{m}_i^3 \right) \right)$, where $\tilde{m}_i$ stands for the dimension of the positive semi-definite cone (*i.e.*, $\boldsymbol{A}_i(\boldsymbol{x}) \in \mathbb{R}^{\tilde{m}_i \times \tilde{m}_i}$) (see [29]). In the SDP that interest us, one can show that $\tilde{n} = 1 + (3K + 2)d$ and that both problems can be solved in $O(K^5 d^{3.5})$ operations, with K being the number of scenarios for each random variable that composes $\boldsymbol{\xi}$. In calculating the total complexity of evaluating $\mathcal{LB}(\boldsymbol{x}_1, \mathcal{X}_2, \{\boldsymbol{\xi}_i^k\})$ under such a Ψ, we also need to account for $O(KdT_{\text{MVP}})$ operations for evaluating the required α_i^k and β_i^k. This completes our proof.

References

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

[2] D. Bertsimas, X. V. Doan, K. Natarajan, and C. P. Teo. Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research*, 35(3):580–602, 2010.

[3] D. Bertsimas and V. Goyal. On the power of robust solutions in two-stage stochastic and adaptive optimization problems. *Mathematics of Operations Research*, 35(2):284–305, 2010.

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

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

[6] D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1st edition, 1997.

[7] J. Birge. The value of the stochastic solution in stochastic linear programs with fixed recourse. *Mathematical Programming*, 24(3):314–325, 1982.

[8] J. Birge. Models and model value in stochastic programming. *Annals of Operations Research*, 59:1–18, 1995.

[9] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, New York, 2011.

[10] J. R. Birge and R. J.-B. Wets. Computing bounds for stochastic programming problems by means of a generalized moment problem. *Mathematics of Operations Research*, 12:149–162, 1987.

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

[12] E.C. Capen. The difficulty of assessing uncertainty. *Journal of Petroleum Technology*, 28(8):843–850, 1976.

[13] V. K. Chopra and W. T. Ziemba. The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management*, 19(2):6–11, 1993.

[14] R. T. Clemen and T. Reilly. *Making Hard Decisions with Decisiontools*. Duxbury, 2nd edition, 2001.

[15] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[16] J. Dupacová. Minimax stochastic programs with nonconvex nonseparable penalty functions. In A. Prékopa, editor, *Progress in Operations Research*, pages 303–316. North Holland, 1976.

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

[18] M. Dyer and L. Stougie. Computational complexity of stochastic programming problems. *Mathematical Programming*, 106(3):423–432, 2006.

[19] D. Ellsberg. Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, 75(4):643–669, 1961.

[20] P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependence in risk management: Properties and pitfalls. In *Risk Management: Value at Risk and Beyond*, pages 176–223. Cambridge University Press, 1999.

[21] L. F. Escudero, A. Garín, M. Merino, and G. Perez. The value of the stochastic solution in multistage problems. *Top*, 15:48–64, 2007.

[22] H. Gassmann and W. Ziemba. A tight upper bound for the expectation of a convex function of a multivariate random variable. *Mathematical Programming Studies*, 27:39–53, 1986.

[23] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. `http://cvxr.com/cvx`, February 2011.

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

[25] D.X. Li. On default correlation: A copula approach. *Journal of Fixed Income*, 9:43–54, 2000.

[26] O. Listes and R. Dekker. A scenario aggregation-based approach for determining a robust airline fleet composition for dynamic capacity allocation. *Transportation Science*, 39(3):367–382, 2005.

[27] F. V. Louveaux and R. Schultz. Stochastic integer programming. In *Handbooks in Operations Research and Management Science, Vol. 10, Stochastic Programming*, pages 213–266. Elsevier Science, 2003.

[28] F. Maggioni and S. Wallace. Analyzing the quality of the expected value solution in stochastic programming. *Annals of Operations Research*, pages 1–18, 2010.

[29] Y. Nesterov and A. Nemirovski. *Interior-point polynomial methods in convex programming*, volume 13. Studies in Applied Mathematics, Philadelphia, 1994.

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

[31] A. Ravindran, D. T. Phillips, and J. J. Solberg. *Operations Research : Principles and Practice*. John Wiley & Sons, 1987.

[32] F. Salmon. Recipe for disaster: The formula that killed wall street. *Wired Magazine*, 17(3), 2009.

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

[34] A. Shapiro. On duality theory of conic linear problems. In M. A. Goberna and M. A. López, editors, *Semi-Infinite Programming: Recent Advances*, pages 135–165, Dordrecht, 2001. Kluwer Academic Publishers.

[35] A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. In Vaithilingam Jeyakumar and Alexander Rubinov, editors, *Continuous Optimization*, volume 99 of *Applied Optimization*, pages 111–146. Springer US, 2005.

[36] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[37] R. H. Tütüncü, K. C. Toh, and M. J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming*, 95:189–217, 2003.

[38] A. Tversky and D. Kahneman. Judgement under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[39] S. W. Wallace. Decision making under uncertainty: Is sensitivity analysis of any use? *Operations Research*, 48:20–25, 2000.