Math. Program., Ser. A
DOI 10.1007/s10107-017-1125-8

FULL LENGTH PAPER

# Data-driven robust optimization

**Dimitris Bertsimas**[1] · **Vishal Gupta**[2] · **Nathan Kallus**[3]

Received: 13 September 2015 / Accepted: 14 February 2017


**Abstract** The last decade witnessed an explosion in the availability of data for operations research applications. Motivated by this growing availability, we propose a novel schema for utilizing data to design uncertainty sets for robust optimization using statistical hypothesis tests. The approach is flexible and widely applicable, and robust optimization problems built from our new sets are computationally tractable, both theoretically and practically. Furthermore, optimal solutions to these problems enjoy a strong, finite-sample probabilistic guarantee whenever the constraints and objective function are concave in the uncertainty. We describe concrete procedures for choosing an appropriate set for a given application and applying our approach to multiple uncertain constraints. Computational evidence in portfolio management and queueing confirm that our data-driven sets significantly outperform traditional robust optimization techniques whenever data are available.

**Keywords** Robust optimization · Data-driven optimization · Chance-constraints · Hypothesis testing

✉ Vishal Gupta
guptavis@usc.edu

Dimitris Bertsimas
dbertsim@mit.edu

Nathan Kallus
kallus@cornell.edu

[1] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2] Marshall School of Business, University of Southern California, Los Angeles, CA 90029, USA

[3] School of Operations Research and Information Engineering, Cornell University and Cornell Tech, New York, NY 10011, USA

Published online: 25 February 2017

**Mathematics Subject Classification** 80M50 (Optimization: Operations research, mathematical programming) · 62H15 (Multivariate Analysis: Hypothesis Testing)

## 1 Introduction

Robust optimization is a popular approach to optimization under uncertainty. The key idea is to define an uncertainty set of possible realizations of the uncertain parameters and then optimize against worst-case realizations within this set. Computational experience suggests that with well-chosen sets, robust models yield tractable optimization problems whose solutions perform as well or better than other approaches. With poorly chosen sets, however, robust models may be overly-conservative or computationally intractable. Choosing a good set is crucial. Fortunately, there are several theoretically motivated and experimentally validated proposals for constructing good uncertainty sets [3,6,10,16]. These proposals share a common paradigm; they combine a priori reasoning with mild assumptions on the uncertainty to motivate the construction of the set.

On the other hand, the last decade witnessed an explosion in the availability of data. Massive amounts of data are now routinely collected in many industries. Retailers archive terabytes of transaction data. Suppliers track order patterns across their supply chains. Energy markets can access global weather data, historical demand profiles, and, in some cases, real-time power consumption information. These data have motivated a shift in thinking—away from a priori reasoning and assumptions and towards a new data-centered paradigm. A natural question, then, is how should robust optimization techniques be tailored to this new paradigm?

In this paper, we propose a general schema for designing uncertainty sets for robust optimization from data. We consider uncertain constraints of the form $f(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0$ where $\mathbf{x} \in \mathbb{R}^k$ is the optimization variable, and $\tilde{\mathbf{u}} \in \mathbb{R}^d$ is an uncertain parameter. We model this constraint by choosing a set $\mathcal{U}$ and forming the corresponding robust constraint

$$f(\mathbf{u}, \mathbf{x}) \leq 0 \quad \forall \mathbf{u} \in \mathcal{U}. \tag{1}$$

We assume throughout that $f(\mathbf{u}, \mathbf{x})$ is concave in $\mathbf{u}$ for any $\mathbf{x}$.

In many applications, robust formulations decompose into a series of constraints of the form (1 through an appropriate transformation of variables, including uncertain linear optimization and multistage adaptive optimization (see, e.g., [6]). In this sense, (1) is the fundamental building block of many robust optimization models.

Many approaches [6,16,22] to constructing uncertainty sets for (1) assume $\tilde{\mathbf{u}}$ is a random variable whose distribution $\mathbb{P}^*$ is not known except for some assumed structural features. For example, they may assume that $\mathbb{P}^*$ has independent components but unknown marginal distributions. Furthermore, instead of insisting the given constraint hold almost surely with respect to $\mathbb{P}^*$, they instead authorize a small probability of violation. Specifically, given $\epsilon > 0$, these approaches seek sets $\mathcal{U}_\epsilon$ that satisfy two key properties:

(P1) The robust constraint (1) is *computationally tractable.*

(P2) The set $\mathcal{U}_\epsilon$ *implies a probabilistic guarantee for* $\mathbb{P}^*$ *at level* $\epsilon$, that is, for any $\mathbf{x}^* \in \mathbb{R}^k$ and for every function $f(\mathbf{u}, \mathbf{x})$ that is concave in $\mathbf{u}$ for every $\mathbf{x}$, we have the implication:

$$\text{If } f\left(\mathbf{u}, \mathbf{x}^*\right) \leq 0 \ \ \forall \mathbf{u} \in \mathcal{U}_\epsilon, \quad \text{then } \mathbb{P}^*\left(f\left(\tilde{\mathbf{u}}, \mathbf{x}^*\right) \leq 0\right) \geq 1 - \epsilon. \tag{2}$$

(P2) ensures that a feasible solution to the robust constraint will also be feasible with probability $1 - \epsilon$ with respect to $\mathbb{P}^*$, despite not knowing $\mathbb{P}^*$ exactly. Existing proposals achieve (P2) by leveraging the a priori structural features of $\mathbb{P}^*$. Some of these approaches, e.g., [16], only consider the special case when $f(\mathbf{u}, \mathbf{x})$ is bi-affine, but one can generalize them to (2) using techniques from [5] (see also Sect. 2.1).

Like previous proposals, we also assume $\tilde{\mathbf{u}}$ is a random variable whose distribution $\mathbb{P}^*$ is not known exactly, and seek sets $\mathcal{U}_\epsilon$ that satisfy these properties. Unlike previous proposals—and this is critical—we assume that we have data $\mathcal{S} = \{\hat{\mathbf{u}}^1, \ldots, \hat{\mathbf{u}}^N\}$ drawn i.i.d. according to $\mathbb{P}^*$. By combining these data with the a priori structural features of $\mathbb{P}^*$, we can design new sets that imply similar probabilistic guarantees, but which are much smaller with respect to subset containment than their traditional counterparts. Consequently, robust models built from our new sets yield less conservative solutions than traditional counterparts, while retaining their robustness properties.

The key to our schema is using the confidence region of a statistical hypothesis test to quantify what we learn about $\mathbb{P}^*$ from the data. Specifically, our schema depends on three ingredients: a priori assumptions on $\mathbb{P}^*$, data, and a hypothesis test. By pairing different a priori assumptions and tests, we obtain distinct data-driven uncertainty sets, each with its own geometric shape, computational properties, and modeling power. These sets can capture a variety of features of $\mathbb{P}^*$, including skewness, heavy-tails and correlations.

In principle, there are many possible pairings of a priori assumptions and tests. We focus on pairings we believe are most relevant to practitioners for their tractability and applicability. Our list is non-exhaustive; there may exist other pairings that yield effective sets. Specifically, we consider situations where:

- $\mathbb{P}^*$ has known, finite discrete support (Sect. 4).
- $\mathbb{P}^*$ may have continuous support, and the components of $\tilde{\mathbf{u}}$ are independent (Sect. 5).
- $\mathbb{P}^*$ may have continuous support, but data are drawn from its marginal distributions asynchronously (Sect. 6). This situation models the case of missing values.
- $\mathbb{P}^*$ may have continuous support, and data are drawn from its joint distribution (Sect. 7). This is the general case.

Table 1 summarizes the a priori structural assumptions, hypothesis tests, and resulting uncertainty sets that we propose. Each set is convex and admits a tractable, explicit description; see the referenced equations.

For each of our sets, we provide an explicit, equivalent reformulation of (1). The complexity of optimizing over this reformulation depends both on the function $f(\mathbf{u}, \mathbf{x})$ and the set $\mathcal{U}$. For each of our sets, we show that this reformulation is polynomial time tractable for a large class of functions $f$ including bi-affine functions, separable functions, conic-quadratic representable functions and certain sums of uncertain

**Table 1** Summary of data-driven uncertainty sets proposed in this paper. SOC, EC and LMI denote second-order cone representable sets, exponential cone representable sets, and linear matrix inequalities, respectively

| Assumptions on $\mathbb{P}^*$ | Hypothesis test | Geometric description | Eqs. | Inner problem |
|---|---|---|---|---|
| Discrete support | $\chi^2$-test | SOC | (13, 15) | |
| Discrete support | G-test | Polyhedral* | (13, 16) | |
| Independent marginals | KS Test | Polyhedral* | (21) | Line search |
| Independent marginals | K Test | Polyhedral* | (76) | Line search |
| Independent marginals | CvM Test | SOC* | (76, 69) | |
| Independent marginals | W Test | SOC* | (76, 70) | |
| Independent marginals | AD Test | EC | (76, 71) | |
| Independent marginals | Chen et al. [23] | SOC | (27) | Closed-form |
| None | Marginal Samples | Box | (31) | Closed-form |
| None | Linear Convex Ordering | Polyhedron | (34) | |
| None | Shawe-Taylor and Cristianini [46] | SOC | (39) | Closed-form |
| None | Delage and Ye [25] | LMI | (41) | |

The additional "*" notation indicates a set of the above type with one additional, relative entropy constraint. $KS$, $K$, $CvM$, $W$, and $AD$ denote the Kolmogorov–Smirnov, Kuiper, Cramer-von Mises, Watson and Anderson-Darling goodness of fit tests, respectively. In some cases, we can identify a worst-case realization of $\mathbf{u}$ in (1) for bi-affine $f$ and a candidate $\mathbf{x}$ with a specialized algorithm. In these cases, the column "Inner Problem" roughly describes this algorithm

exponential functions. By exploiting special structure in some of our sets, we can provide specialized routines for identifying a worst-case realization of $\mathbf{u}$ in (1) for bi-affine $f$ and a candidate solution $\mathbf{x}$.[1] Utilizing this separation routine within a cutting-plane method may offer performance superior to approaches which attempt to solve (1) directly [13,38]. In these cases, the column "Inner Problem" in Table 1 roughly describes these routines.

We are not the first to consider using hypothesis tests in data-driven optimization; others have considered more specialized applications of hypothesis testing. Klabjan et al. [34] proposes a distributionally robust dynamic program based on Pearson's $\chi^2$-test for a particular inventory problem. Goldfarb and Iyengar [29] calibrate an uncertainty set for the mean and covariance of a distribution using linear regression and the $t$ test. It is not clear how to generalize these methods to other settings, e.g., distributions with continuous support in the first case or general parameter uncertainty in the second. By contrast, we offer a comprehensive study of the connection between hypothesis testing and uncertainty set design, addressing a number of cases with general machinery.

Recently, Ben-Tal et al. [9] proposed a class of data-driven uncertainty sets based on phi-divergences. Several classical hypothesis tests, like Pearson's $\chi^2$-test and the

[1] We say $f(\mathbf{u}, \mathbf{x})$ is bi-affine if the function $\mathbf{u} \mapsto f(\mathbf{u}, \mathbf{x})$ is affine for any fixed $\mathbf{x}$ and the function $\mathbf{x} \mapsto f(\mathbf{u}, \mathbf{x})$ is affine for any fixed $\mathbf{u}$.

G-test are based on phi-divergences (see also [32]). They focus on the case where the uncertain parameters $\tilde{\mathbf{u}}$, themselves, are a probability distribution with known, finite, discrete support. Robust optimization problems where the uncertainty is a probability distribution are typically called *distributionally robust optimization* (DRO) problems, and the corresponding uncertainty sets are called *ambiguity sets.* Although there have been a huge number of ambiguity sets proposed in the literature based on generalized moment constraints and probability metrics (see, e.g., [28,50] for recent work), to the best of our knowledge Ben-Tal et al. [9] is the first to connect an ambiguity set with a hypothesis test. In contrast to these DRO models for ambiguity sets, we design uncertainty sets for general uncertain parameters $\tilde{\mathbf{u}}$, such as future product demand, service times, and asset returns; these uncertain parameters need not represent probabilities. Methodologically, treating general uncertain parameters requires different techniques than those typically used in constructing ambiguity sets.

This distinction is not to suggest our work entirely unrelated to DRO. Our hypothesis testing perspective provides a unified view of ambiguity sets in DRO and many other data-driven methods from the literature. For example, Calafiore and El Ghaoui [18] and Delage and Ye [25] have proposed data-driven methods for chance-constrained and distributionally robust problems, respectively, without using hypothesis testing. We show how these works can be reinterpreted through the lens of hypothesis testing. Leveraging this viewpoint enables us to apply methods from statistics, such as the bootstrap, to refine these methods and improve their numerical performance. Moreover, applying our schema, we can design data-driven uncertainty sets for robust optimization based upon these methods. Although we focus on Calafiore and El Ghaoui [18] and Delage and Ye [25] in this paper, this strategy applies equally well to a host of other methods beyond DRO, such as the likelihood estimation approach of [49]. In this sense, we believe hypothesis testing and uncertainty set design provide a common framework in which to compare and contrast different approaches.

At the same time, Ben-Tal et al. [6] establish a one-to-one correspondence between uncertainty sets for linear optimization that satisfy (P2) and safe approximations to ambiguous linear chance constraints (see also Remark 1). Recall that an ambiguous, linear chance constraint in $\mathbf{x}$ is of the form $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(\mathbf{x}^T\tilde{\mathbf{u}} \leq 0) \geq 1-\epsilon$ for some ambiguity set $\mathcal{P}$, i.e., it is a specific instance of DRO. Thus, through this correspondence, all of our results can be recast as new *data-driven* constructions for safe-approximations to chance constraints. Whether one phrases our results in the language of ambiguous chance constraints or uncertainty sets for (classical) robust optimization is largely a matter of taste. In what follows, we prefer uncertainty sets since many existing robust optimization applications in engineering and operations research are formulated in terms of general uncertain parameters. Our new uncertainty sets can be *directly* substituted into these existing models with little additional effort.

Finally, we note that Campi and Garatti [21] propose a very different data-driven method for robust optimization not based on hypothesis tests. In their approach, one replaces the uncertain constraint $f(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0$ with $N$ sampled constraints over the data, $f(\hat{\mathbf{u}}^j, \mathbf{x}) \leq 0$, for $j = 1, \ldots, N$. For $f(\mathbf{u}, \mathbf{x})$ convex in $\mathbf{x}$ with arbitrary dependence in $\mathbf{u}$, they provide a tight bound $N(\epsilon)$ such that if $N \geq N(\epsilon)$, then, with high probability with respect to the sampling procedure $\mathbb{P}_{\mathcal{S}}$, any $\mathbf{x}$ which is feasible in the $N$ sampled constraints satisfies $\mathbb{P}^*(f(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0) \geq 1-\epsilon$. Various refinements of this

base method have been proposed yielding smaller bounds $N(\epsilon)$, including incorporating $\ell_1$-regularization [20] and allowing $\mathbf{x}$ to violate a small fraction of the constraints [19]. Compared to our approach, these methods are more generally applicable and provide a similar probabilistic guarantee. In the special case we treat where $f(\tilde{\mathbf{u}}, \mathbf{x})$ is concave in $\mathbf{u}$, however, our proposed approach offers some advantages. First, because it leverages the concave structure of $f(\mathbf{u}, \mathbf{x})$, our approach generally yields less conservative solutions (for the same $N$ and $\epsilon$) than [21] (see Sect. 3). Second, for fixed $\epsilon > 0$, our approach is applicable even if $N < N(\epsilon)$, while theirs is not. This distinction is important when $\epsilon$ is very small and there may not exist enough data. Finally, as we will show, our approach reformulates (1) as a series of (relatively) sparse convex constraints, while the Campi and Garatti's [21] approach will in general yield $N$ dense constraints which may be numerically challenging when $N$ is large.

We summarize our contributions:

1. We propose a new, systematic schema for constructing uncertainty sets from data using statistical hypothesis tests. When the data are drawn i.i.d. from an unknown distribution $\mathbb{P}^*$, sets built from our schema imply a probabilistic guarantee for $\mathbb{P}^*$ at any desired level $\epsilon$.
2. We illustrate our schema by constructing a multitude of uncertainty sets. Each set is applicable under slightly different a priori assumptions on $\mathbb{P}^*$ as described in Table 1.
3. We prove that robust optimization problems over each of our sets are generally tractable. Specifically, for each set, we derive an explicit robust counterpart to (1) and show that for a large class of functions $f(\mathbf{u}, \mathbf{x})$ optimizing over this counterpart can be accomplished in polynomial time using off-the-shelf software.
4. We unify several existing data-driven methods through the lens of hypothesis testing. Through this lens, we motivate the use of common numerical techniques from statistics such as bootstrapping and Gaussian approximation to improve their performance. Moreover, we apply our schema to derive new uncertainty sets for (1) inspired by the refined versions of these methods.
5. We illustrate how to model multiple uncertain constraints with our sets by optimizing the parameters chosen for each individual constraint. This approach is tractable and yields solutions which will satisfy all uncertain constraints simultaneously for any desired level $\epsilon$.
6. We illustrate how common cross-validation techniques from model selection in machine learning can be used to choose an appropriate set and calibrate its parameters.
7. Through applications in queueing and portfolio allocation, we assess the relative strengths and weaknesses of our sets. Overall, we find that although all of our sets shrink in size as $N \to \infty$, they differ in their ability to represent features of $\mathbb{P}^*$. Consequently, they may perform very differently in a given application. In the above two settings, we find that our model selection technique frequently identifies a good set choice, and a robust optimization model built with this set performs as well or better than other robust data-driven approaches.

The remainder of the paper is structured as follows. Section 2 reviews background to keep the paper self-contained. Section 3 presents our schema for constructing uncer-

tainty sets. Sections 4–7 describe the various constructions in Table 1. Section 8 reinterprets several techniques in the literature through the lens of hypothesis testing and, subsequently, uses them to motivate new uncertainty sets. Section 9.1 and "Appendix 3" discuss modeling multiple constraints and choosing the right set for an application, respectively. The remainder of Sect. 9 presents numerical experiments, and Sect. 10 concludes. All proofs are in the electronic companion.

In what follows, we adopt the following notational conventions: Boldfaced lowercase letters ($\mathbf{x}, \boldsymbol{\theta}, \ldots$) denote vectors, boldfaced capital letters ($\mathbf{A}, \mathbf{C}, \ldots$) denote matrices, and ordinary lowercase letters ($x, \theta$) denote scalars. Calligraphic type ($\mathcal{P}, \mathcal{S} \ldots$) denotes sets. The $i$th coordinate vector is $\mathbf{e}_i$, and the vector of all ones is $\mathbf{e}$. We always use $\tilde{\mathbf{u}} \in \mathbb{R}^d$ to denote a *random* vector and $\tilde{u}_i$ to denote its components. $\mathbb{P}$ denotes a generic probability measure for $\tilde{\mathbf{u}}$, and $\mathbb{P}^*$ denotes its true (unknown) measure. Moreover, $\mathbb{P}_i$ denotes the marginal measure of $\tilde{u}_i$. We let $\mathcal{S} = \{\hat{\mathbf{u}}^1, \ldots, \hat{\mathbf{u}}^N\}$ be a sample of $N$ data points drawn i.i.d. according to $\mathbb{P}^*$, and let $\mathbb{P}^*_{\mathcal{S}}$ denote the measure of the sample $\mathcal{S}$, i.e., the $N$-fold product distribution of $\mathbb{P}^*$. Finally, $\hat{\mathbb{P}}$ denotes the empirical distribution with respect to $\mathcal{S}$, i.e., for any Borel set $\mathcal{A}$, $\hat{\mathbb{P}}(\tilde{\mathbf{u}} \in \mathcal{A}) \equiv \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}(\hat{\mathbf{u}}^j \in \mathcal{A})$. Here $\mathbb{I}(\cdot)$ denotes the usual indicator function.

## 2 Background

To keep the paper self-contained, we recall some results needed to prove our sets are tractable and imply a probabilistic guarantee.

### 2.1 Tractability of Robust Nonlinear Constraints

Ben-Tal et al. [5] study constraint (1) and prove that for nonempty, convex, compact $\mathcal{U}$ satisfying a mild, regularity condition,[2] (1) is equivalent to

$$\exists \mathbf{v} \in \mathbb{R}^d, t, s \in \mathbb{R} \text{ s.t. } \delta^*(\mathbf{v}|\,\mathcal{U}) \leq t,\ f_*(\mathbf{v}, \mathbf{x}) \geq s,\ t - s \leq 0. \tag{3}$$

Here, $f_*(\mathbf{v}, \mathbf{x})$ denotes the partial concave-conjugate of $f(\mathbf{u}, \mathbf{x})$ and $\delta^*(\mathbf{v}|\,\mathcal{U})$ denotes the support function of $\mathcal{U}$, defined respectively as

$$f_*(\mathbf{v}, \mathbf{x}) \equiv \inf_{\mathbf{u} \in \mathbb{R}^d} \mathbf{u}^T \mathbf{v} - f(\mathbf{u}, \mathbf{x}), \quad \delta^*(\mathbf{v}|\,\mathcal{U}) \equiv \sup_{\mathbf{u} \in \mathcal{U}} \mathbf{v}^T \mathbf{u}. \tag{4}$$

For many $f(\mathbf{u}, \mathbf{x})$, $f_*(\mathbf{v}, \mathbf{x})$ admits a simple, explicit description. For example, for bi-affine $f(\mathbf{u}, \mathbf{x}) = \mathbf{u}^T \mathbf{F} \mathbf{x} + \mathbf{f}_{\mathbf{u}}^T \mathbf{u} + \mathbf{f}_{\mathbf{x}}^T \mathbf{x} + f_0$, we have

$$f_*(\mathbf{v}, \mathbf{x}) = \begin{cases} -\mathbf{f}_{\mathbf{x}}^T \mathbf{x} - f_0 & \text{if } v = \mathbf{F}\mathbf{x} + \mathbf{f}_{\mathbf{u}} \\ -\infty & \text{otherwise,} \end{cases}$$

[2] An example of a sufficient regularity condition is that $ri(\mathcal{U}) \cap ri(dom(f(\cdot, \mathbf{x}))) \neq \emptyset$, $\forall \mathbf{x} \in \mathbb{R}^k$. Here $ri(\mathcal{U})$ denotes the *relative interior* of $\mathcal{U}$. Recall that for any non-empty convex set $\mathcal{U}$, $ri(\mathcal{U}) \equiv \{\mathbf{u} \in \mathcal{U} : \forall \mathbf{z} \in \mathcal{U},\ \exists \lambda > 1 \text{ s.t. } \lambda \mathbf{u} + (1 - \lambda)\mathbf{z} \in \mathcal{U}\}$ (cf. [11]).

and (3) simplifies to

$$\delta^*\left(\mathbf{F}\mathbf{x}+\mathbf{f}_{\mathbf{u}}|\,\mathcal{U}\right)+\mathbf{f}_{\mathbf{x}}^T\mathbf{x}+f_0\leq 0. \tag{5}$$

In what follows, we concentrate on proving that we can represent $\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U})\leq t\}$ with a small number of convex inequalities suitable for off-the-shelf solvers for each of our sets $\mathcal{U}$. From (5), this representation will imply that (1) is theoretically and practically tractable for each of our sets whenever $f(\mathbf{u},\mathbf{x})$ is bi-affine.

On the other hand, Ben-Tal et al. [5] provide a number of other examples of $f(\mathbf{u},\mathbf{x})$ for which $f_*(\mathbf{v},\mathbf{x})$ is tractable, including:

Separable Concave: $f(\mathbf{u},\mathbf{x})=\sum_{i=1}^k f_i(\mathbf{u})x_i$, for $f_i(\mathbf{u})$ concave and $x_i\geq 0$.
Uncertain Exponentials: $f(\mathbf{u},\mathbf{x})=-\sum_{i=1}^k x_i^{u_i}$, for $x_i>1$ and $0<u_i\leq 1$.
Conic Quadratic Representable: $f(\mathbf{u},\mathbf{x})$ such that the set $\{(t,\mathbf{u})\in\mathbb{R}\times\mathbb{R}^d : f(\mathbf{u},\mathbf{x})\geq t\}$ is conic quadratic representable ( cf. [40]).

Consequently, by providing a representation of $\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U})\leq t\}$ for each of our sets, we will also have proven that (1) is tractable for each of these classes of functions via (3).

For some sets, our formulation of $\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U})\leq t\}$ will involve complex nonlinear constraints, such as exponential cone constraints (cf. Table 1). Although it is theoretically possible to optimize over such constraints in polynomial time, this approach may be numerically challenging. An alternative to solving (3) directly is to use cutting-plane, bundle, or online optimization methods (see [8,13,38] for details). While these methods differ in the specifics of how they address (1), the critical subroutine in each method is "solving the inner problem." Specifically, given a candidate solution $(\mathbf{v}_0,t_0)$, one must be able to easily compute $\mathbf{u}^*\in\arg\max_{\mathbf{u}\in\mathcal{U}}\mathbf{v}_0^T\mathbf{u}$ (notice $\mathbf{u}^*$ depends on $\mathbf{v}_0$). From the definitions of the support function and $\mathbf{u}^*$, we have $(\mathbf{v}_0,t_0)\in\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U})\leq t\}$ if and only if $\mathbf{v}_0^T\mathbf{u}^*\leq t_0$. In particular, if $(\mathbf{v}_0,t_0)\notin\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U})\leq t\}$, then the hyperplane $\{(\mathbf{v},t) : \mathbf{v}^T\mathbf{u}^*=t\}$ separates $(\mathbf{v}_0,t_0)$ from $\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U})\leq t\}$. Namely, any $(\mathbf{v},t)\in\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U})\leq t\}$ satisfies the inequality $\mathbf{v}^T\mathbf{u}^*\leq t$, but $(\mathbf{v}_0,t_0)$ does not. Such separating hyperplanes are used in cutting-plane and bundling methods to iteratively build up the constraint (1).

Although it is possible to use this idea to prove polynomial time tractability of robust constraints over our sets via the ellipsoid algorithm using separation oracles (see [30] for details), we do not pursue this idea. Rather, our primary motivation is in improving *practical* efficiency in the spirit of Bertsimas et al. [13] when the reformulation (3) may be challenging. To this end, when possible, we provide specialized algorithms for solving the inner problem and identifying a $\mathbf{u}^*$ through closed-form formulas or line searches. Practitioners can then employ these specialized algorithms within one of the above referenced cutting-plane, bundle, or online learning methods to yield practically efficient algorithms for large-scale instances.

### 2.2 Hypothesis Testing

We briefly review hypothesis testing. See [35] for a more complete treatment.

Given a null-hypothesis $H_0$ that makes a claim about an unknown distribution $\mathbb{P}^*$, a hypothesis test seeks to use data $\mathcal{S}$ drawn from $\mathbb{P}^*$ to either declare that $H_0$ is false, or, else, that there is insufficient evidence to determine its validity. For a given significance level $0 < \alpha < 1$, a typical test consists of a test statistic $T \equiv T(\mathcal{S}, H_0)$, depending on the data and $H_0$, and a threshold $\Gamma \equiv \Gamma(\alpha, \mathcal{S}, H_0)$, depending on $\alpha$, $\mathcal{S}$, and $H_0$. If $T > \Gamma$, we reject $H_0$. Since $T$ depends on $\mathcal{S}$, it is random. The threshold $\Gamma$ is chosen so that the probability with respect to $\mathbb{P}_{\mathcal{S}}$ of *incorrectly* rejecting $H_0$ is at most $\alpha$. The choice of $\alpha$ is often application specific, although values of $\alpha = 1, 5$ and $10\%$ are common (cf., [35, Chapt. 3.1].)

As an example, consider the two-sided Student's $t$ test [35, Chapt. 5].) Given $\mu_0 \in \mathbb{R}$, the $t$ test considers the null-hypothesis $H_0 : \mathbb{E}^{\mathbb{P}^*}[\tilde{u}] = \mu_0$ using the statistic $T = |(\hat{\mu} - \mu_0)/(\hat{\sigma}\sqrt{N})|$ and threshold $\Gamma = t_{N-1,1-\alpha/2}$. Here $\hat{\mu}, \hat{\sigma}$ are the sample mean and sample standard deviation, respectively, and $t_{N-1,1-\alpha}$ is the $1-\alpha$ quantile of the Student $t$ distribution with $N-1$ degrees of freedom. Under the a priori assumption that $\mathbb{P}^*$ is Gaussian, the test guarantees that we will incorrectly reject $H_0$ with probability at most $\alpha$.

Many of the tests we consider are common in applied statistics, and tables for their thresholds are widely available. Several of our tests, however, are novel (e.g., the deviations test in Sect. 5.2.) In these cases, we propose using the *bootstrap* to approximate a threshold (cf. Algorithm 1). $N_B$ should be chosen to be fairly large; we take $N_B = 10^4$ in our experiments. The bootstrap is a well-studied and widely-used technique in statistics [26,35]. Strictly speaking, hypothesis tests based on the bootstrap are only asymptotically valid for large $N$ (see the references for a precise statement). Nonetheless, they are routinely used in applied statistics, even with $N$ as small as 100, and a wealth of practical experience suggests they are extremely accurate. Consequently, we believe practitioners can safely use bootstrapped thresholds in the above tests.

**Algorithm 1** Bootstrapping a Threshold

**Input:** $\mathcal{S}, T, H_0, 0 < \alpha < 1, N_B \in \mathbb{Z}_+$
**Output:** Approximate Threshold $\Gamma$
  **for** $j = 1 \ldots N_B$ **do**
    $\mathcal{S}^j \leftarrow$ Resample $|\mathcal{S}|$ data points from $\mathcal{S}$ with replacement
    $T^j \leftarrow T(\mathcal{S}^j, H_0)$
  **end for**
  **return** $\lceil N_B(1-\alpha) \rceil$-largest value of $T^1, \ldots, T^{N_B}$.

Finally, we introduce the confidence region of a test, which will play a critical role in our construction. Given data $\mathcal{S}$, the $1-\alpha$ confidence region of a test is the set of null-hypotheses that would not be rejected for $\mathcal{S}$ at level $1-\alpha$. For example, the $1-\alpha$ confidence region of the $t$ test is $\left\{\mu_0 \in \mathbb{R} : \left|\frac{\hat{\mu}-\mu_0}{\hat{\sigma}\sqrt{N}}\right| \leq t_{N-1,1-\alpha/2}\right\}$. In what follows, however, we commit a slight abuse of nomenclature and instead use the term confidence region to refer to the set of all measures that are consistent with any a priori assumptions of the test and also satisfy a null-hypothesis that would not be rejected. In the case of the $t$ test, the confidence region in the context of this paper is

$$\mathcal{P}^t \equiv \left\{\mathbb{P} \in \Theta(-\infty, \infty) : \mathbb{P} \text{ is Gaussian with mean } \mu_0, \text{ and } \left|\frac{\hat{\mu} - \mu_0}{\hat{\sigma}\sqrt{N}}\right| \leq t_{N-1,1-\alpha/2}\right\}, \tag{6}$$

where $\Theta(-\infty, \infty)$ is the set of Borel probability measures on $\mathbb{R}$.

By construction, the probability (with respect to the sampling procedure $\mathbb{P}_S$) that $\mathbb{P}^*$ is a member of its confidence region is at least $1 - \alpha$ as long as all a priori assumptions are valid. This is a critical observation. Despite not knowing $\mathbb{P}^*$, we can use a hypothesis test to create a set of distributions from the data that contains $\mathbb{P}^*$ for any specified probability.

## 3 Designing Data-Driven Uncertainty Sets

### 3.1 Geometric Characterization of the Probabilistic Guarantee

As a first step towards our schema, we provide a geometric characterization of (P2). One might intuit that a set $\mathcal{U}$ implies a probabilistic guarantee at level $\epsilon$ only if $\mathbb{P}^*(\tilde{\mathbf{u}} \in \mathcal{U}) \geq 1 - \epsilon$. As noted by Ben-Tal et al. [6, pp. 32–33], however, this intuition is false. Often, sets that are much smaller than the $1 - \epsilon$ support will still imply a probabilistic guarantee at level $\epsilon$, and such sets should be preferred because they are less conservative.

The crux of the issue is that there may be many realizations $\tilde{\mathbf{u}} \notin \mathcal{U}$ where nonetheless $f(\tilde{\mathbf{u}}, \mathbf{x}^*) \leq 0$. Thus, $\mathbb{P}^*(\tilde{\mathbf{u}} \in \mathcal{U})$ is in general an underestimate of $\mathbb{P}^*(f(\tilde{\mathbf{u}}, \mathbf{x}^*) \leq 0)$. One needs to exploit the dependence of $f$ on $\mathbf{u}$ to refine the estimate. We note in passing that many existing data-driven approaches for robust optimization, e.g., [21], do not leverage this dependence. Consequently, although these approaches are general purpose, they may yield overly conservative uncertainty sets for (1).

In order to tightly characterize (P2), we introduce the Value at Risk. For any $\mathbf{v} \in \mathbb{R}^d$ and measure $\mathbb{P}$, the Value at Risk at level $\epsilon$ with respect to $\mathbf{v}$ is

$$\text{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) \equiv \inf\left\{t : \mathbb{P}\left(\mathbf{v}^T\tilde{\mathbf{u}} \leq t\right) \geq 1 - \epsilon\right\}. \tag{7}$$

Value at Risk is positively homogenous (in $\mathbf{v}$), but typically non-convex. (Recall a function $g(\mathbf{v})$ is positively homogenous if $g(\lambda\mathbf{v}) = \lambda g(\mathbf{v})$ for all $\lambda > 0$.) The critical result underlying our method is a relationship between Value at Risk and support functions of sets which satisfy (P2) (cf. (4)):

**Theorem 1** a) *Suppose $\mathcal{U}$ is non-empty, convex, and compact. Assume that for every $\mathbf{v} \in \mathbb{R}^d$, $\delta^*(\mathbf{v}|\,\mathcal{U}) \geq \text{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)$ $\forall \mathbf{v} \in \mathbb{R}^d$. Then, $\mathcal{U}$ implies a probabilistic guarantee at level $\epsilon$ for $\mathbb{P}$.*

b) *Suppose $\exists \mathbf{v} \in \mathbb{R}^d$ such that $\delta^*(\mathbf{v}|\,\mathcal{U}) < \text{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)$. Then, there exists bi-affine functions $f(\mathbf{u}, \mathbf{x})$ for which (2) does not hold.*

The first part generalizes a result implicitly used in [6,23] when designing uncertainty sets for the special case of bi-affine functions. To the best of our knowledge, the extension to general concave functions $f$ is new.

### 3.2 Our Schema

The principal challenge in applying Theorem 1 to designing uncertainty sets is that $\mathbb{P}^*$ is not known. Recall, however, that the confidence region $\mathcal{P}$ of a hypothesis test, will contain $\mathbb{P}^*$ with probability at least $1-\alpha$. This motivates the following schema: Fix $0 < \alpha < 1$ and $0 < \epsilon < 1$.

Data-Driven Uncertainty Set Schema:

1. Let $\mathcal{P}(\mathcal{S}, \alpha, \epsilon)$ be the confidence region of a hypothesis test at level $\alpha$.
2. Construct a closed, convex, finite-valued, positively homogenous (in $\mathbf{v}$) upper-bound $g(\mathbf{v}, \mathcal{S}, \epsilon, \alpha)$ to the worst-case Value at Risk over $\mathcal{P}(\mathcal{S}, \alpha, \epsilon)$:

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{S},\alpha,\epsilon)} \text{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) \leq g\left(\mathbf{v}, \mathcal{S}, \epsilon, \alpha\right) \quad \forall \mathbf{v} \in \mathbb{R}^d.$$

3. Identify the closed, convex set $\mathcal{U}(\mathcal{S}, \epsilon, \alpha)$ such that $g(\mathbf{v}, \mathcal{S}, \epsilon, \alpha) = \delta^*(\mathbf{v}|\, \mathcal{U}(\mathcal{S}, \epsilon, \alpha))$.

Note, the existence of the set in Step 3 is guaranteed by the bijection between closed, finite-valued, positively homogenous convex functions and convex, compact sets (see [11]).

**Theorem 2** *With probability at least $1-\alpha$ with respect to $\mathbb{P}_\mathcal{S}$, the resulting set $\mathcal{U}(\mathcal{S}, \epsilon, \alpha)$ implies a probabilistic guarantee at level $\epsilon$ for $\mathbb{P}^*$.*

*Remark 1* Note that $\delta^*(\mathbf{v}|\, \mathcal{U}(\mathcal{S}, \epsilon, \alpha)) \leq t$ is a safe-approximation to the ambiguous chance constraint $\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{S},\alpha,\epsilon)} \mathbb{P}(\mathbf{v}^T\tilde{\mathbf{u}} \leq t) \geq 1-\epsilon$ as defined in [6]. Ambiguous chance-constraints are closely related to sets which satisfy (P2). See [6] for more details. Practitioners who prefer to model with ambiguous chance constraints can directly use $\delta^*(\mathbf{v}|\, \mathcal{U}(\mathcal{S}, \epsilon, \alpha)) \leq t$ in their formulations as a data-driven approach. We provide explicit descriptions of $\delta^*(\mathbf{v}|\, \mathcal{U}(\mathcal{S}, \epsilon, \alpha)) \leq t$ below for each of our sets for this purpose.

Theorem 2 ensures that with probability at least $1-\alpha$ with respect to the sampling procedure $\mathbb{P}_\mathcal{S}$, a robust feasible solution $\mathbf{x}$ will satisfy a *single* uncertain constraint $f(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0$ with probability at least $1-\epsilon$. Often, however, we face $m > 1$ uncertain constraints $f_j(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0,\ \ j = 1, \ldots, m$, and seek $\mathbf{x}$ that will simultaneously satisfy these constraints, i.e.,

$$\mathbb{P}\left(\max_{j=1,\ldots,m} f_j\left(\tilde{\mathbf{u}}, \mathbf{x}\right) \leq 0\right) \geq 1-\overline{\epsilon}, \tag{8}$$

for some given $\overline{\epsilon}$. One approach is to replace each uncertain constraint with a corresponding robust constraint

$$f_j\left(\mathbf{u}, \mathbf{x}\right) \leq 0, \quad \forall \mathbf{u} \in \mathcal{U}\left(\mathcal{S}, \epsilon_j, \alpha\right), \tag{9}$$

where $\mathcal{U}(\mathcal{S}, \epsilon_j, \alpha)$ is constructed via our schema at level $\epsilon_j = \epsilon/m$. By the union bound and Theorem 2, with probability at least $1-\alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$, any $\mathbf{x}$ which satisfies (9) will satisfy (8).

The choice $\epsilon_j = \epsilon/m$ is somewhat arbitrary. We would prefer to treat the $\epsilon_j$ as decision variables and optimize over them, i.e., replace the $m$ uncertain constraints by

$$\min_{\epsilon_1+\ldots+\epsilon_m \leq \overline{\epsilon}, \boldsymbol{\epsilon} \geq \mathbf{0}} \left\{ \max_{j=1,\ldots,m} \left\{ \max_{\mathbf{u} \in \mathcal{U}(\mathcal{S}, \epsilon_j, \alpha)} f_j(\mathbf{u}, \mathbf{x}) \right\} \right\} \leq 0$$

or, equivalently,

$$\exists \epsilon_1 + \ldots + \epsilon_m \leq \overline{\epsilon},\ \boldsymbol{\epsilon} \geq \mathbf{0} \ :\ f_j(\mathbf{u}, \mathbf{x}) \leq 0 \ \ \forall \mathbf{u} \in \mathcal{U}\left(\mathcal{S}, \epsilon_j, \alpha\right), \quad j = 1, \ldots, m. \tag{10}$$

Unfortunately, Theorem 2 does not imply that with probability at least $1-\alpha$ any feasible solution to (10) will satisfy (8). The issue is that Theorem 2 requires selecting $\epsilon$ independently of $\mathcal{S}$, whereas the optimal $\epsilon_j$'s in (10) *will* depend on $\mathcal{S}$, creating an in-sample bias. We next introduce a stronger requirement on an uncertainty set than "implying a probabilistic guarantee," and adapt Theorem 2 to address (10).

Given a family of sets indexed by $\epsilon$, $\{\mathcal{U}(\epsilon) : 0 < \epsilon < 1\}$, we say this family *simultaneously* implies a probabilistic guarantee for $\mathbb{P}^*$ if, for all $0 < \epsilon < 1$, each $\mathcal{U}(\epsilon)$ implies a probabilistic guarantee for $\mathbb{P}^*$ at level $\epsilon$.

**Theorem 3** *Suppose $\mathcal{P}(\mathcal{S}, \alpha, \epsilon) \equiv \mathcal{P}(\mathcal{S}, \alpha)$ does not depend on $\epsilon$ in Step 1. above. Let $\{\mathcal{U}(\mathcal{S}, \epsilon, \alpha) : 0 < \epsilon < 1\}$ be the resulting family of sets obtained from our schema.*

*a) With probability at least $1-\alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$, $\{\mathcal{U}(\mathcal{S}, \epsilon, \alpha) : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee for $\mathbb{P}^*$.*
*b) With probability at least $1-\alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$, any $\mathbf{x}$ which satisfies (10) will satisfy (8).*

We provide numerical evidence in Sect. 9 that (10) offers significant benefit over (9). In some special cases, we can optimize the $\epsilon_j$'s in (10) exactly (see "Appendix 4). More generally, we must approximate this outer optimization numerically. We propose a specialized method leveraging the structure of our sets for this purpose in "Appendix 3".

Depending on the quality of bound $g(\cdot)$ in Step 2 of our schema, the resulting set $\mathcal{U}(\mathcal{S}, \epsilon, \alpha)$ may not be contained in the support of $\mathbb{P}^*$. When a priori information is available on this support, we can always improve our set by taking intersections:

**Theorem 4** *Suppose supp$(\mathbb{P}^*) \subseteq \mathcal{U}_0$ where $\mathcal{U}_0$ is closed and convex. Suppose further that $\mathcal{U}_\epsilon$ is convex, compact. Then,*

*a) If $\mathcal{U}_\epsilon$ implies a probabilistic guarantee for $\mathbb{P}^*$ at level $\epsilon$, then $\mathcal{U}_\epsilon \cap \mathcal{U}_0$ also implies a probabilistic guarantee for $\mathbb{P}^*$ at level $\epsilon$.*
*b) If $\{\mathcal{U}_\epsilon : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee for $\mathbb{P}^*$, then $\{\mathcal{U}_\epsilon \cap \mathcal{U}_0 : 0 < \epsilon < 1\}$ also simultaneously impies a probabilistic guarantee for $\mathbb{P}^*$.*

*Remark 2* The convexity condition on $\mathcal{U}_0$ is necessary. It is not difficult to construct examples where $\mathcal{U}_0$ is non-convex and $\mathcal{U}_0 \cap \mathcal{U}_\epsilon = \emptyset$, e.g., the example from [6, pp. 32–33] has this property.

*Remark 3* For many sets $\mathcal{U}_0$, such as boxes, polyhedrons or ellipsoids, robust constraints over $\mathcal{U}_\epsilon \cap \mathcal{U}_0$ are essentially as tractable as robust constraints over $\mathcal{U}_\epsilon$. Specifically, from [5, Lemma A.4],

$$\begin{aligned}&\left\{(\mathbf{v}, t) : \delta^*\left(\mathbf{v}|\mathcal{U}_\epsilon \cap \mathcal{U}_0\right) \le t\right\} \\ &= \Big\{ (\mathbf{v}, t) : \exists \mathbf{w} \in \mathbb{R}^d,\ t_1, t_2 \in \mathbb{R} \text{ s.t. } \delta^*\left(\mathbf{v} - \mathbf{w}|\mathcal{U}_\epsilon\right) \\ &\qquad \le t_1,\ \delta^*\left(\mathbf{w}|\mathcal{U}_0\right) \le t_2,\ t_1 + t_2 \le t \Big\}. \end{aligned} \tag{11}$$

Consequently, whenever the constraint $\delta^*(\mathbf{w}|\,\mathcal{U}_0) \le t_2$ is tractable, the constraint (11) will also be tractable.

The next four sections apply this schema to create uncertainty sets. Since, $\epsilon$, $\alpha$ and $\mathcal{S}$ are typically fixed, we suppress some or all of them in the notation.

## 4 Uncertainty Sets Built from Discrete Distributions

In this section, we assume $\mathbb{P}^*$ has known, finite support, i.e., $\text{supp}(\mathbb{P}^*) \subseteq \{\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}\}$. We consider two hypothesis tests: Pearson's $\chi^2$ test and the $G$ test [42]). Both tests consider the hypothesis $H_0 : \mathbb{P}^* = \mathbb{P}_0$ where $\mathbb{P}_0$ is some specified measure. Specifically, let $p_i = \mathbb{P}_0(\tilde{\mathbf{u}} = \mathbf{a}_i)$ be the specified null-hypothesis, and let $\hat{\mathbf{p}}$ denote the empirical probability distribution, i.e.,

$$\hat{p}_i \equiv \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left(\hat{\mathbf{u}}^j = \mathbf{a}_i\right) \quad i = 0, \ldots, n-1.$$

In words, $\hat{p}_i$ represents the proportion of the sample taking value $\mathbf{a}_i$. Pearson's $\chi^2$ test rejects $H_0$ at level $\alpha$ if $\sum_{i=0}^{n-1} \frac{(p_i - \hat{p}_i)^2}{2p_i} > \frac{1}{2N} \chi^2_{n-1,1-\alpha}$, where $\chi^2_{n-1,1-\alpha}$ is the $1-\alpha$ quantile of a $\chi^2$ distribution with $n-1$ degrees of freedom. Similarly, the $G$ test rejects the null hypothesis at level $\alpha$ if $D(\hat{\mathbf{p}}, \mathbf{p}) > \frac{1}{2N} \chi^2_{n-1,1-\alpha}$ where

$$D\left(\mathbf{p}, \mathbf{q}\right) \equiv \sum_{i=0}^{n-1} p_i \log\left(p_i / q_i\right) \tag{12}$$

is the relative entropy between $\mathbf{p}$ and $\mathbf{q}$.

The confidence regions for Pearson's $\chi^2$ test and the $G$ test are, respectively,

$$\mathcal{P}^{\chi^2} = \left\{ \mathbf{p} \in \Delta_n : \sum_{i=0}^{n-1} \frac{\left(p_i - \hat{p}_i\right)^2}{2p_i} \leq \frac{1}{2N} \chi^2_{n-1,1-\alpha} \right\},$$
$$\mathcal{P}^{G} = \left\{ \mathbf{p} \in \Delta_n : D\left(\hat{\mathbf{p}}, \mathbf{p}\right) \leq \frac{1}{2N} \chi^2_{n-1,1-\alpha} \right\}. \tag{13}$$

Here $\Delta_n = \left\{(p_0, \ldots, p_{n-1})^T : \mathbf{e}^T \mathbf{p} = 1, \quad p_i \geq 0 \;\; i = 0, \ldots, n-1\right\}$ denotes the probability simplex. We will use these two confidence regions in Step 1 of our schema.

For a fixed measure $\mathbb{P}$, and vector $\mathbf{v} \in \mathbb{R}^d$, recall the Conditional Value at Risk:

$$\text{CVaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T \tilde{\mathbf{u}}\right) \equiv \min_t \left\{ t + \frac{1}{\epsilon} \mathbb{E}^{\mathbb{P}} \left[ \left( \tilde{\mathbf{u}}^T \mathbf{v} - t \right)^+ \right] \right\}. \tag{14}$$

Conditional Value at Risk is well-known to be a convex upper bound to Value at Risk [1,43] for a fixed $\mathbb{P}$. We can compute a bound in Step 2 by considering the worst-case Conditional Value at Risk over the above confidence regions, yielding

**Theorem 5** *Suppose supp*$(\mathbb{P}^*) \subseteq \{\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}\}$. *With probability* $1 - \alpha$ *over the sample, the families* $\{\mathcal{U}_\epsilon^{\chi^2} : 0 < \epsilon < 1\}$ *and* $\{\mathcal{U}_\epsilon^{G} : 0 < \epsilon < 1\}$ *simultaneously imply a probabilistic guarantee for* $\mathbb{P}^*$, *where*

$$\mathcal{U}_\epsilon^{\chi^2} = \left\{ \mathbf{u} \in \mathbb{R}^d : \mathbf{u} = \sum_{j=0}^{n-1} q_j \mathbf{a}_j, \;\; \mathbf{q} \in \Delta_n, \;\; \mathbf{q} \leq \frac{1}{\epsilon} \mathbf{p}, \;\; \mathbf{p} \in \mathcal{P}^{\chi^2} \right\}, \tag{15}$$

$$\mathcal{U}_\epsilon^{G} = \left\{ \mathbf{u} \in \mathbb{R}^d : \mathbf{u} = \sum_{j=0}^{n-1} q_j \mathbf{a}_j, \;\; \mathbf{q} \in \Delta_n, \;\; \mathbf{q} \leq \frac{1}{\epsilon} \mathbf{p}, \;\; \mathbf{p} \in \mathcal{P}^{G} \right\}. \tag{16}$$

*Their support functions are given by*

$$\begin{aligned} \delta^*\left(\mathbf{v} \middle| \mathcal{U}_\epsilon^{\chi^2}\right) = \min_{\beta, \mathbf{w}, \eta, \lambda, \mathbf{t}, \mathbf{s}} \;\; & \beta + \frac{1}{\epsilon}\left(\eta + \frac{\lambda \chi^2_{n-1,1-\alpha}}{N} + 2\lambda - 2\sum_{i=0}^{n-1} \hat{p}_i s_i\right) \\ \text{s.t.} \;\; & \mathbf{0} \leq \mathbf{w} \leq (\lambda + \eta)\,\mathbf{e}, \;\; \lambda \geq 0, \;\; \mathbf{s} \geq \mathbf{0}, \\ & \left\| \begin{matrix} 2s_i \\ w_i - \eta \end{matrix} \right\| \leq 2\lambda - w_i + \eta, \;\; \mathbf{a}_i^T \mathbf{v} - w_i \leq \beta, \quad i = 0, \ldots, n-1, \end{aligned} \tag{17}$$

$$\begin{aligned} \delta^*\left(\mathbf{v} \middle| \mathcal{U}_\epsilon^{G}\right) = \min_{\beta, \mathbf{w}, \eta, \lambda} \;\; & \beta + \frac{1}{\epsilon}\left(\eta + \frac{\lambda \chi^2_{n-1,1-\alpha}}{2N} - \lambda \sum_{i=0}^{n-1} \hat{p}_i \log\left(1 - \frac{w_i - \eta}{\lambda}\right)\right) \\ \text{s.t} \;\; & \mathbf{0} \leq \mathbf{w} \leq (\lambda + \eta)\,\mathbf{e}, \;\; \lambda \geq 0, \\ & \mathbf{a}_i^T \mathbf{v} - w_i \leq \beta, \quad i = 0, \ldots, n-1. \end{aligned} \tag{18}$$

*Remark 4* The sets $\mathcal{U}_\epsilon^{\chi^2}, \mathcal{U}_\epsilon^{G}$ strongly resemble the uncertainty set for $\text{CVaR}_\epsilon^{\hat{\mathbb{P}}}$ in [12]. In fact, as $N \to \infty$, all three of these sets converge almost surely to the set $\mathcal{U}^{\text{CVaR}_\epsilon^{\mathbb{P}^*}}$

defined by $\delta^*(\mathbf{v}|\mathcal{U}^{\mathrm{CVaR}_\epsilon^{\mathbb{P}^*}}) = \mathrm{CVaR}_\epsilon^{\mathbb{P}^*}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)$. The key difference is that for finite $N$, $\mathcal{U}_\epsilon^{\chi^2}$ and $\mathcal{U}_\epsilon^G$ imply a probabilistic guarantee for $\mathbb{P}^*$ at level $\epsilon$, while $\mathcal{U}^{\mathrm{CVaR}_\epsilon^{\hat{\mathbb{P}}}}$ does not.

*Remark 5* Theorem 5 exemplifies the distinction drawn in the introduction between uncertainty sets for discrete probability distributions—such as $\mathcal{P}^{\chi^2}$ or $\mathcal{P}^G$ which have been proposed in [9]—and uncertainty sets for general uncertain parameters like $\mathcal{U}_\epsilon^{\chi^2}$ and $\mathcal{U}_\epsilon^G$. The relationship between these two types of sets is explicit in Eqs. (15) and (16) because we have known, finite support. For continuous support and our other sets, the relationship is implicit in the worst-case value-at-risk in Step 2 of our schema.

*Remark 6* When representing $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^{\chi^2}) \le t\}$ or $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^G) \le t\}$, it suffices to find auxiliary variables that are feasible in (17) or (18). Thus, these sets are second-order-cone and exponential-cone representable, respectively. Although theoretically tractable, the exponential cone can be numerically challenging.

Because of these numerical issues, modeling with $\mathcal{U}_\epsilon^{\chi^2}$ is perhaps preferable to modeling with $\mathcal{U}_\epsilon^G$. Fortunately, for large $N$, the difference between these two sets is negligible:

**Proposition 1** *With arbitrarily high probability, for any* $\mathbf{p} \in \mathcal{P}^G$, $|D(\hat{\mathbf{p}}, \mathbf{p}) - \sum_{j=0}^{n-1} \frac{(\hat{p}_j - p_j)^2}{2p_j}| = O(nN^{-3})$.

Thus, for large $N$, $\mathcal{P}^G$ is approximately equal to $\mathcal{P}^{\chi^2}$, whereby $\mathcal{U}_\epsilon^G$ is approximately equal to $\mathcal{U}_\epsilon^{\chi^2}$. For large $N$, then, $\mathcal{U}_\epsilon^{\chi^2}$ should be preferred for its computational tractability.

### 4.1 A Numerical Example of $\mathcal{U}_\epsilon^{\chi^2}$ and $\mathcal{U}_\epsilon^G$

Figure 1 illustrates the sets $\mathcal{U}_\epsilon^{\chi^2}$ and $\mathcal{U}_\epsilon^G$ with a particular numerical example. The true distribution is supported on the vertices of the given octagon. Each vertex is labeled with its true probability. In the absence of data when the support of $\mathbb{P}^*$ is known, the only uncertainty set $\mathcal{U}$ which implies a probabilistic guarantee for $\mathbb{P}^*$ is the convex hull of these points. We construct the sets $\mathcal{U}_\epsilon^{\chi^2}$ (grey line) and $\mathcal{U}_\epsilon^G$ (black line) for $\alpha = \epsilon = 10\%$ for various $N$. For reference, we also plot $\mathcal{U}^{\mathrm{CVaR}_\epsilon^{\mathbb{P}^*}}$ (shaded region) which is the limit of both sets as $N \to \infty$.

For small $N$, our data-driven sets are equivalent to the convex hull of $\mathrm{supp}(\mathbb{P}^*)$, however, as $N$ increases, our sets shrink considerably. For large $N$, as predicted by Propostion 1, $\mathcal{U}_\epsilon^G$ and $\mathcal{U}_\epsilon^{\chi^2}$ are very similarly shaped.

*Remark 7* Figure 1 also enables us to contrast our approach to that of Campi and Garatti [21]. Namely, suppose that $f(\mathbf{u}, \mathbf{x})$ is linear in $\mathbf{u}$. In this case, $\mathbf{x}$ satisfies $f(\hat{\mathbf{u}}^j, \mathbf{x}) \le 0$ for $j = 1, \ldots, N$, if and only if $f(\mathbf{u}, \mathbf{x}) \le 0$ for all $\mathbf{u} \in \mathrm{conv}(\mathcal{A})$ where $\mathcal{A} \equiv \{\mathbf{a} \in \mathrm{supp}(\mathbb{P}^*) : \exists 1 \le j \le N \text{ s.t. } \mathbf{a} = \hat{\mathbf{u}}^j\}$. As $N \to \infty$, $\mathcal{A} \to \mathrm{supp}(\mathbb{P}^*)$ almost

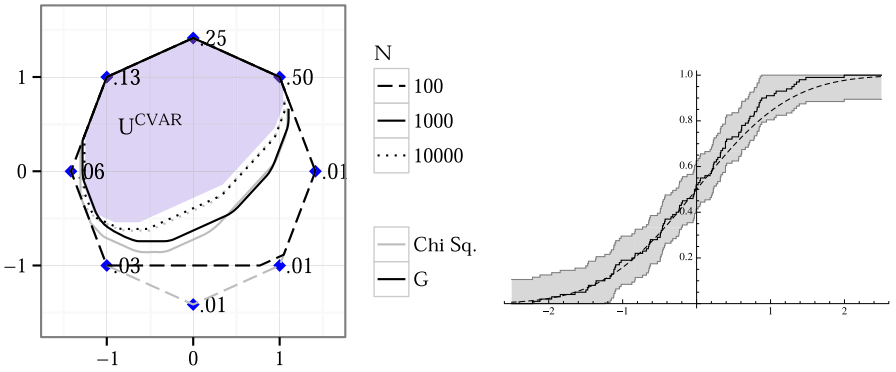


**Fig. 1** The *left panel* shows the sets $\mathcal{U}_\epsilon^{\chi^2}$ and $\mathcal{U}_\epsilon^{G}$, $\alpha = \epsilon = 10\%$. When $N = 0$, the smallest set which implies a probabilistic guarantee is $\text{supp}(\mathbb{P}^*)$, the given octagon. As $N$ increases, both sets shrink to the $\mathcal{U}^{\text{CVaR}_\epsilon^{\mathbb{P}^*}}$ given by the shaded region. The *right panel* shows the empirical distribution function and confidence region corresponding to the Kolmogorov–Smirnov test

surely. In other words, as $N \to \infty$, the method of Campi and Garatti [21] in this case is equivalent to using the entire support as an uncertainty set, which is much larger than $\mathcal{U}^{\text{CVaR}_\epsilon^{\mathbb{P}^*}}$ above. Similar examples can be constructed with continuous distributions or the method of Calafiore and Monastero [19]. In each case, the critical observation is that these methods do not explicitly leverage the concave (or, in this case, linear) structure of $f(\mathbf{u}, \mathbf{x})$.

## 5 Independent Marginal Distributions

We next assume $\mathbb{P}^*$ may have continuous support, but the marginal distributions $\mathbb{P}_i^*$ are independent. Our strategy is to build a multivariate test by combining univariate tests for each marginal distribution.

### 5.1 Uncertainty Sets Built from the Kolmogorov–Smirnov Test

For this section, we assume that $\text{supp}(\mathbb{P}^*)$ is contained in a known, finite box $[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}] \equiv \{\mathbf{u} \in \mathbb{R}^d : \hat{u}_i^{(0)} \le u_i \le \hat{u}_i^{(N+1)}, \quad i = 1, \ldots, d\}$.

Given a univariate measure $\mathbb{P}_{0,i}$, the Kolmogorov–Smirnov (KS) goodness-of fit test applied to marginal $i$ considers the null-hypothesis $H_0 : \mathbb{P}_i^* = \mathbb{P}_{0,i}$. It rejects this hypothesis if

$$\max_{j=1,\ldots,N} \max\left(\frac{j}{N} - \mathbb{P}_{0,i}\left(\tilde{u} \le \hat{u}_i^{(j)}\right), \mathbb{P}_{0,i}\left(\tilde{u} < \hat{u}_i^{(j)}\right) - \frac{j-1}{N}\right) > \Gamma^{KS}.$$

where $\hat{u}_i^{(j)}$ is the $j$th largest element among $\hat{u}_i^1, \ldots, \hat{u}_i^N$. Tables for $\Gamma^{KS}$ are widely available [47,48].

The confidence region of the above test for the $i$-th marginal distribution is

$$\mathcal{P}_i^{KS} = \left\{ \mathbb{P}_i \in \Theta\left[\hat{u}_i^{(0)}, \hat{u}_i^{(N+1)}\right] : \mathbb{P}_i\left(\tilde{u}_i \le \hat{u}_i^{(j)}\right) \right.$$
$$\left. \ge \frac{j}{N} - \Gamma^{KS},\ \mathbb{P}_i\left(\tilde{u}_i < \hat{u}_i^{(j)}\right) \le \frac{j-1}{N} + \Gamma^{KS},\ j = 1, \ldots, N \right\},$$

where $\Theta[\hat{u}_i^{(0)}, \hat{u}_i^{(N+1)}]$ is the set of all Borel probability measures on $[\hat{u}_i^{(0)}, \hat{u}_i^{(N+1)}]$. Unlike $\mathcal{P}^{\chi^2}$ and $\mathcal{P}^G$, this confidence region is infinite dimensional.

Figure 1 illustrates an example. The true distribution is a standard normal whose cumulative distribution function (cdf) is the dotted line. We draw $N = 100$ data points and form the empirical cdf (solid black line). The 80% confidence region of the KS test is the set of measures whose cdfs are more than $\Gamma^{KS}$ above or below this solid line, i.e. the grey region.

Now consider the multivariate null-hypothesis $H_0 : \mathbb{P}^* = \mathbb{P}_0$. Since $\mathbb{P}^*$ has independent components, the test which rejects if $\mathbb{P}_i$ fails the KS test at level $\alpha' = 1 - \sqrt[d]{1-\alpha}$ for any $i$ is a valid test. Namely, $\mathbb{P}^*_{\mathcal{S}}(\mathbb{P}_i^*$ is accepted by KS at level $\alpha'$ for all $i = 1, \ldots, d) = \prod_{i=1}^d (1-\alpha') = 1-\alpha$ by independence. The confidence region of this multivariate test is

$$\mathcal{P}^I = \left\{ \mathbb{P} \in \Theta\left[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}\right] : \mathbb{P} = \prod_{i=1}^d \mathbb{P}_i,\ \ \mathbb{P}_i \in \mathcal{P}_i^{KS}\ \ i = 1, \ldots, d \right\}.$$

("I" in $\mathcal{P}^I$ emphasizes independence). We use this confidence region in Step 1 of our schema.

When the marginals are independent, Nemirovski and Shapiro [41] proved

$$\mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) \le \inf_{\lambda \ge 0}\left(\lambda \log(1/\epsilon) + \lambda \sum_{i=1}^d \log \mathbb{E}^{\mathbb{P}_i}\left[e^{v_i \tilde{u}_i/\lambda}\right]\right).$$

This bound implies the worst-case bound

$$\sup_{\mathbb{P}\in\mathcal{P}^I} \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) \le \inf_{\lambda \ge 0}\left(\lambda \log(1/\epsilon) + \lambda \sum_{i=1}^d \log \sup_{\mathbb{P}_i \in \mathcal{P}_i^{KS}} \mathbb{E}^{\mathbb{P}_i}\left[e^{v_i \tilde{u}_i/\lambda}\right]\right), \quad (19)$$

which we use in Step 2 of our schema. We solve the inner-most supremum explicitly by leveraging the simple geometry of $\mathcal{P}_i^{KS}$. Intuitively, the worst-case distribution will either be the lefthand boundary or the righthand boundary of the region in Fig. 1

depending on the sign of $v_i$. These boundaries are defined by the discrete distributions $\mathbf{q}^L(\Gamma), \mathbf{q}^R(\Gamma) \in \Delta_{N+2}$ supported on $\hat{u}_i^{(0)}, \ldots, \hat{u}_i^{(N+1)}$ and defined by

$$q_j^L(\Gamma) = \begin{cases} \Gamma & \text{if } j = 0, \\ \frac{1}{N} & \text{if } 1 \le j \le \lfloor N(1-\Gamma) \rfloor, \\ 1 - \Gamma - \frac{\lfloor N(1-\Gamma) \rfloor}{N} & \text{if } j = \lfloor N(1-\Gamma) \rfloor + 1, \\ 0 & \text{otherwise,} \end{cases} \qquad q_j^R(\Gamma) = q_{N+1-j}^L(\Gamma), \quad j = 0, \ldots, N+1. \tag{20}$$

(Recall that $D(\cdot, \cdot)$ denotes the relative entropy, cf. (12).) Then, we have

**Theorem 6** *Suppose $\mathbb{P}^*$ has independent components, with supp$(\mathbb{P}^*) \subseteq [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$. With probability at least $1 - \alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$, $\{\mathcal{U}_\epsilon^I : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee for $\mathbb{P}^*$, where*

$$\begin{aligned} \mathcal{U}_\epsilon^I = \Bigg\{ & \mathbf{u} \in \mathbb{R}^d : \exists \theta_i \in [0,1],\ \mathbf{q}^i \in \Delta_{N+2},\ i = 1 \ldots, d, \\ & \sum_{j=0}^{N+1} \hat{u}_i^{(j)} q_j^i = u_i,\ i = 1, \ldots, d,\ \sum_{i=1}^{d} D\left(\mathbf{q}_i, \theta_i \mathbf{q}^L\left(\Gamma^{KS}\right)\right. \\ & \left. + (1 - \theta_i)\, \mathbf{q}^R\left(\Gamma^{KS}\right)\right) \le \log(1/\epsilon) \Bigg\}. \end{aligned} \tag{21}$$

*Moreover,*

$$\begin{aligned} \delta^*\left(\mathbf{v} \middle| \mathcal{U}_\epsilon^I\right) = \inf_{\lambda \ge 0} \Bigg\{ & \lambda \log(1/\epsilon) \\ & + \lambda \sum_{i=1}^{d} \log\left[ \max\left( \sum_{j=0}^{N+1} q_j^L(\Gamma^{KS}) e^{v_i \hat{u}_i^{(j)}/\lambda}, \sum_{j=0}^{N+1} q_j^R(\Gamma^{KS}) e^{v_i \hat{u}_i^{(j)}/\lambda} \right) \right] \Bigg\} \end{aligned} \tag{22}$$

*Remark 8* When representing $\{(\mathbf{v}, t) : \delta^*(\mathbf{v} | \mathcal{U}^I) \le t)\}$, we can drop the infimum over $\lambda \ge 0$ in (22). This set is exponential cone representable, which, again, may be numerically challenging.

*Remark 9* By contrast, because $\mathbf{q}^L(\Gamma)$ (resp. $\mathbf{q}^R(\Gamma)$) is decreasing (resp. increasing) in its components, the lefthand branch of the innermost maximum in (22) will be attained when $v_i \le 0$ and the righthand branch is attained otherwise. Thus, for fixed $\mathbf{v}$, the optimization problem in $\lambda$ is convex and differentiable and can be efficiently solved with a line search. We can use this line search to identify a worst-case realization of $\mathbf{u}$ for a fixed $\mathbf{v}$. Specifically, let $\lambda^*$ be an optimal solution. Define

$$\mathbf{p}^i = \begin{cases} \mathbf{q}^L & \text{if } v_i \le 0, \\ \mathbf{q}^R & \text{otherwise,} \end{cases} \qquad q_j^i = \frac{p_j^i e^{v_i \hat{u}_i^{(j)}/\lambda^*}}{\sum_{j=0}^{N+1} p_j^i e^{v_i \hat{u}_i^{(j)}/\lambda^*}}, \quad j = 0, \ldots, N+1, \quad i = 1, \ldots, d,$$

$$u_i^* = \sum_{j=0}^{N+1} q_j^i \hat{u}_i^{(j)}, \quad i = 1 \ldots, d.$$

Then $\mathbf{u}^* \in \arg\max_{\mathbf{u} \in \mathcal{U}_\epsilon^I} \mathbf{v}^T \mathbf{u}$. That this procedure is valid follows from the proof of Theorem 6.

*Remark 10* The KS test is one of many goodness-of-fit tests based on the empirical distribution function (EDF), including the Kuiper (K), Cramer von-Mises (CvM), Watson (W) and Andersen-Darling (AD) tests [48, Chapt. 5]. We can define analogues of $\mathcal{U}_\epsilon^I$ for each of these tests, each having slightly different shape. Separating over $\{(\mathbf{v}, t) : \delta^*(\mathbf{v} | \mathcal{U}) \le t\}$ is polynomial time tractable for each of these sets, but we no longer have a simple algorithm for generating violated cuts. Thus, these sets are considerably less attractive from a computational point of view. Fortunately, through simulation studies with a variety of different distributions, we have found that the version of $\mathcal{U}_\epsilon^I$ based on the KS test generally performs as well as or better than the other EDF tests. Consequently, we recommend using the sets $\mathcal{U}_\epsilon^I$ as described. For completeness, we present the constructions for the analogous tests in "Appendix 5".

### 5.2 Uncertainty Sets Motivated by Forward and Backward Deviations

In [23], the authors propose an uncertainty set based on the forward and backward deviations of a distribution. Recall, for a univariate distribution $\mathbb{P}_i$, its forward and backward deviations are defined by

$$\begin{aligned} \sigma_{fi}(\mathbb{P}_i) &= \sup_{x>0} \sqrt{-\frac{2\mu_i}{x} + \frac{2}{x^2} \log\left(\mathbb{E}^{\mathbb{P}_i}\left[e^{x\tilde{u}_i}\right]\right)}, \\ \sigma_{bi}(\mathbb{P}_i) &= \sup_{x>0} \sqrt{\frac{2\mu_i}{x} + \frac{2}{x^2} \log\left(\mathbb{E}^{\mathbb{P}_i}\left[e^{-x\tilde{u}_i}\right]\right)}, \end{aligned} \tag{23}$$

where $\mathbb{E}^{\mathbb{P}_i}[\tilde{u}_i] = \mu_i$. The optimizations defining $\sigma_{fi}(\mathbb{P}_i)$, $\sigma_{bi}(\mathbb{P}_i)$ are one dimensional, convex problems which can be solved by a line search.

Chen et al. [23] focus on a non-data-driven setting, where the mean and support of $\mathbb{P}^*$ are known a priori, and show how to upper bound these deviations to calibrate their set. In a setting where one has data *and a priori knows the mean of* $\mathbb{P}^*$ *precisely*, they propose a method based on sample average approximation to estimate these deviations. Unfortunately, the precise statistical behavior of these estimators is not known, so it is not clear that this set calibrated from data implies a probabilistic guarantee with high probability with respect to $\mathbb{P}_\mathcal{S}$.

In this section, we use our schema to generalize the set of Chen et al. [23] to a data-driven setting where *neither the mean of the distribution nor its support are*

*known*. Our set differs in shape and size from their proposal, and, unlike their original proposal, will simultaneously imply a probabilistic guarantee for $\mathbb{P}^*$.

We begin by creating an appropriate multivariate hypothesis test. To streamline the exposition, we assume throughout this section $\mathbb{P}^*$ has bounded (but potentially unknown) support. This assumption ensures both $\sigma_{fi}(\mathbb{P}_i)$, $\sigma_{bi}(\mathbb{P}_i)$ are finite [23].

Let $\alpha' = 1 - \sqrt[d]{1-\alpha}$. For a given $\mu_{0,i}, \sigma_{0,fi}, \sigma_{0,bi} \in \mathbb{R}$, consider the following null-hypotheses

$$H_0^1 : \mathbb{E}^{\mathbb{P}_i^*}\left[\tilde{u}\right] = \mu_{0,i}, \quad H_0^2 : \sigma_{fi}\left(\mathbb{P}_i^*\right) \leq \sigma_{0,fi}, \quad H_0^3 : \sigma_{bi}\left(\mathbb{P}_i^*\right) \leq \sigma_{0,bi} \tag{24}$$

and the three tests that rejects if $|\hat{\mu}_i - \mu_{0,i}| > t_i$, $\sigma_{fi}(\hat{\mathbb{P}}_i) > \overline{\sigma}_{fi}$ and $\sigma_{bi}(\hat{\mathbb{P}}_i) > \overline{\sigma}_{bi}$, respectively. Pick the thresholds $t_i$, $\overline{\sigma}_{fi}$ and $\overline{\sigma}_{bi}$ so that these tests are valid at levels $\alpha'/2$, $\alpha'/4$, and $\alpha'/4$, respectively. Since these three tests are not common in applied statistics, there are no tables for their thresholds. In practice, however, we will compute approximate thresholds for each test using the bootstrap (Algorithm 1). By the union bound, the test that rejects if any of these three tests rejects is valid at level $\alpha'$ for the null-hypothesis that $H_0^1$, $H_0^2$ and $H_0^3$ are all true. The confidence region of this test is

$$\mathcal{P}_i^{FB} = \{\mathbb{P}_i \in \Theta\,(-\infty, \infty) : m_{bi} \leq \mathbb{E}_i^{\mathbb{P}}\left[\tilde{u}_i\right] \leq m_{fi}, \quad \sigma_{fi}\,(\mathbb{P}_i) \leq \overline{\sigma}_{fi}, \quad \sigma_{bi}\,(\mathbb{P}_i) \leq \overline{\sigma}_{bi}\},$$

where $m_{bi} = \hat{\mu}_i - t_i$ and $m_{fi} = \hat{\mu}_i + t_i$.

Now consider the multivariate null-hypothesis and test

$$\begin{aligned} &H_0 : \mathbb{E}^{\mathbb{P}_i^*}\left[\tilde{u}\right] = \mu_{0,i},\ \sigma_{fi}\left(\mathbb{P}_i^*\right) \leq \sigma_{0,fi},\ \sigma_{bi}\left(\mathbb{P}_i^*\right) \leq \sigma_{0,bi} \quad \forall i = 1, \ldots, d, \\ &\text{Reject if } |\hat{\mu}_i - \mu_{0,i}| > t_i \text{ or } \sigma_{fi}\left(\hat{\mathbb{P}}_i\right) > \overline{\sigma}_{fi} \text{ or } \sigma_{bi}\left(\hat{\mathbb{P}}_i\right) > \overline{\sigma_{bi}} \text{ for any } i = 1, \ldots, d \end{aligned} \tag{25}$$

where $t_i, \overline{\sigma}_{fi}, \overline{\sigma_{bi}}$ are valid thresholds for the previous univariate test (24) at levels $\alpha'/2$, $\alpha'/4$ and $\alpha'/4$, respectively. As in Sect. 5, this is a valid test at level $\alpha$. Its confidence region is $\mathcal{P}^{FB} = \{\mathbb{P} : \mathbb{P}_i \in \mathcal{P}_i^{FB}\ i = 1, \ldots, d\}$. We use this confidence region in Step 1 of our schema.

When the mean and deviations for $\mathbb{P}$ are known and the marginals are independent, Chen et al. [23] prove

$$\mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) \leq \sum_{i=1}^{d} \mathbb{E}^{\mathbb{P}}\left[\tilde{u}_i\right] v_i + \sqrt{2\log(1/\epsilon)\left(\sum_{i:v_i<0} \sigma_{bi}^2(\mathbb{P}) v_i^2 + \sum_{i:v_i\geq 0} \sigma_{fi}^2(\mathbb{P}) v_i^2\right)}. \tag{26}$$

Computing the worst-case value of this bound over the above confidence region in Step 2 of our schema yields:

**Theorem 7** *Suppose $\mathbb{P}^*$ has independent components and bounded support. Let $t_i$, $\overline{\sigma}_{fi}$ and $\overline{\sigma}_{bi}$ be thresholds such that (25) is a valid test at level $\alpha$. With probability*

*$1-\alpha$ with respect to the sample, the family $\{\mathcal{U}_\epsilon^{FB} : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee for $\mathbb{P}^*$, where*

$$\mathcal{U}_\epsilon^{FB} = \left\{ \mathbf{y}_1 + \mathbf{y}_2 - \mathbf{y}_3 : \mathbf{y}_2, \mathbf{y}_2 \in \mathbb{R}_+^d, \ \sum_{i=1}^{d} \frac{y_{2i}^2}{2\overline{\sigma}_{fi}^2} + \frac{y_{3i}^2}{2\overline{\sigma}_{bi}^2} \right.$$
$$\left. \leq \log(1/\epsilon), \ \ m_{bi} \leq y_{1i} \leq m_{fi}, \ \ i = 1, \ldots, d \right\}. \tag{27}$$

*Moreover,*

$$\delta^*\left(\mathbf{v} | \, \mathcal{U}_\epsilon^{FB}\right) = \sum_{i:v_i \geq 0} m_{fi} v_i + \sum_{i:v_i<0} m_{bi} v_i$$
$$+ \sqrt{2\log(1/\epsilon)\left(\sum_{i:v_i \geq 0} \overline{\sigma}_{fi}^2 v_i^2 + \sum_{i:v_i<0} \overline{\sigma}_{bi}^2 v_i^2\right)} \tag{28}$$

*Remark 11* From (28), $\{(\mathbf{v}, t) : \delta^*(\mathbf{v} | \, \mathcal{U}_\epsilon^{FB}) \leq t\}$ is second order cone representable. We can identify a worst-case realization of $\mathbf{u}$ in closed-form. Given $\mathbf{v}$, let

$$\lambda = \sqrt{\frac{\sum_{i:v_i>0} v_i^2 \overline{\sigma}_{fi}^2 + \sum_{i:v_i \leq 0} v_i^2 \overline{\sigma}_{bi}^2}{2\log(1/\epsilon)}}, \quad u_i^* = \begin{cases} m_{fi} + \frac{v_i \overline{\sigma}_{fi}^2}{\lambda} & \text{if } v_i > 0 \\ m_{bi} + \frac{v_i \overline{\sigma}_{bi}^2}{\lambda} & \text{otherwise.} \end{cases}$$

Then $\mathbf{u}^* \in \arg\max_{\mathbf{u} \in \mathcal{U}_\epsilon^{FB}} \mathbf{v}^T \mathbf{u}$. The correctness of this procedure follows from the proof of Theorem 7.

*Remark 12* $\mathcal{U}_\epsilon^{FB}$ need not be contained within $\text{supp}(\mathbb{P}^*)$. If a priori information about $\text{supp}(\mathbb{P}^*)$ is known, we should apply Theorem 4 to refine $\mathcal{U}_\epsilon^{FB}$ to the smaller intersection $\mathcal{U}_\epsilon^{FB} \cap \text{conv}(\text{supp}(\mathbb{P}^*))$

### 5.3 Comparing $\mathcal{U}_\epsilon^I$ and $\mathcal{U}_\epsilon^{FB}$

Figure 2 illustrates the sets $\mathcal{U}_\epsilon^I$ and $\mathcal{U}_\epsilon^{FB}$ numerically. The marginal distributions of $\mathbb{P}^*$ are independent and their densities are given in the *left panel*. Notice that the first marginal is symmetric while the second is highly skewed.

In the absence of data, knowing only $\text{supp}(\mathbb{P}^*)$ and that $\mathbb{P}^*$ has independent components, the smallest uncertainty which implies a probabilistic guarantee is the unit square (dotted line). With $N = 100$ data points from this distribution (blue circles), however, we can construct both $\mathcal{U}_\epsilon^I$ (dashed black line) and $\mathcal{U}_\epsilon^{FB}$ (solid black line) with $\epsilon = \alpha = 10\%$, as shown. We also plot the limiting shape of these two sets as $N \to \infty$ (corresponding grey lines).

Several features are evident from the plots. First, both sets are able to *learn* from the data that $\mathbb{P}^*$ is symmetric in its first coordinate (the sets display vertical symmetry)

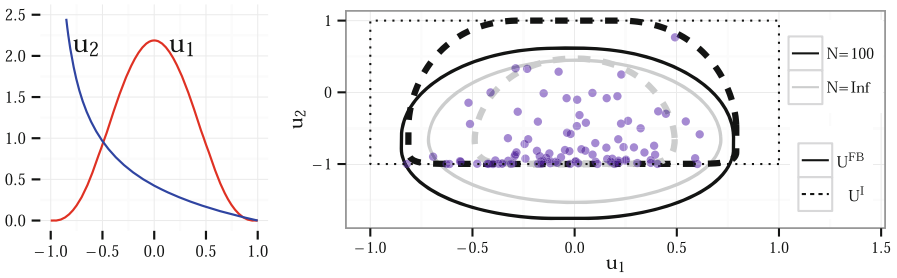


**Fig. 2** The *left panel* shows the marginal densities. The *right panel* shows $\mathcal{U}_\epsilon^I$ (*dashed black line*) and $\mathcal{U}_\epsilon^{FB}$ (*solid black line*) built from $N = 100$ data points (*blue circles*) and in the limit as $N \to \infty$ (corresponding *grey lines*) (color figure online)

and that $\mathbb{P}^*$ is skewed downwards in its second coordinate (the sets taper more sharply towards the top). Second, although $\mathcal{U}_\epsilon^I$ is a strict subset of $\text{supp}(\mathbb{P}^*)$, $\mathcal{U}_\epsilon^{FB}$ is not. Finally, neither set is a subset of the other, and, although for $N = 100$, $\mathcal{U}_\epsilon^{FB} \cap \text{supp}(\mathbb{P}^*)$ has smaller volume than $\mathcal{U}_\epsilon^I$, the reverse holds for larger $N$. Consequently, the best choice of set likely depends on $N$.

## 6 Uncertainty Sets Built from Marginal Samples

In this section, we observe samples from the marginal distributions of $\mathbb{P}^*$ separately, but do not assume these marginals are independent. This happens, e.g., when samples are drawn asynchronously, or when there are many missing values. In these cases, it is impossible to learn the joint distribution of $\mathbb{P}^*$ from the data. To streamline the exposition, we assume that we observe exactly $N$ samples of each marginal distribution. The results generalize to the case of different numbers of samples at the expense of more notation.

In the univariate case, David and Nagaraja [24] develop a hypothesis test for the $1-\epsilon/d$ quantile, or equivalently $\text{VaR}_{\epsilon/d}^{\mathbb{P}_i}(\tilde{u}_i)$ of a distribution $\mathbb{P}$. Namely, given $\overline{q}_{i,0} \in \mathbb{R}$, consider the hypothesis $H_{0,i} : \text{VaR}_{\epsilon/d}^{\mathbb{P}^*}(\tilde{u}_i) \geq \overline{q}_{i,0}$. Define the index $s$ by

$$s = \min \left\{ k \in \mathbb{N} : \sum_{j=k}^{N} \binom{N}{j} (\epsilon/d)^{N-j} (1-\epsilon/d)^j \leq \frac{\alpha}{2d} \right\}, \tag{29}$$

and let $s = N + 1$ if the corresponding set is empty. Then, the test which rejects if $q_{i,0} > \hat{u}_i^{(s)}$ is valid at level $\alpha/2d$ [24, Sect.7.1]. David and Nagaraja [24] also prove that $\frac{s}{N} \downarrow (1-\epsilon/d)$.

The above argument applies symmetrically to the hypothesis $H_{0,i} : \text{VaR}_{\epsilon/d}^{\mathbb{P}^*}(-\tilde{u}_i) \geq \underline{q}_{i,0}$ where the rejection threshold now becomes $\hat{u}_i^{(N-s+1)}$. In the typical case when $\epsilon/d$ is small, $N - s + 1 < s$ so that $\hat{u}_i^{(N-s+1)} \leq \hat{u}_i^{(s)}$.

Next given $\overline{q}_{i,0}, \underline{q}_{i,0} \in \mathbb{R}$ for $i = 1, \ldots, d$, consider the multivariate hypothesis:

$$H_0 : \text{VaR}^{\mathbb{P}^*}_{\epsilon/d}(\tilde{u}_i) \geq \overline{q}_{i,0} \quad \text{and} \quad \text{VaR}^{\mathbb{P}^*}_{\epsilon/d}(-\tilde{u}_i) \geq \underline{q}_{i,0} \quad \text{for all } i = 1, \ldots, d.$$

By the union bound, the test which rejects if $\hat{u}_i^{(s)} < \overline{q}_i$ or $-\hat{u}_i^{(N-s+1)} < \underline{q}_i$, i.e., the above tests fail for the $i$-th component, is valid at level $\alpha$. Its confidence region is

$$\mathcal{P}^M = \left\{ \mathbb{P} \in \mathcal{O}\left[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}\right] : \ \text{VaR}^{\mathbb{P}_i}_{\epsilon/d}(\tilde{u}_i) \leq \hat{u}_i^{(s)}, \right.$$
$$\left. \text{VaR}^{\mathbb{P}_i}_{\epsilon/d}(-\tilde{u}_i) \geq \hat{u}_i^{(N-s+1)}, \ \ i = 1, \ldots, d \right\}.$$

Here "M" is to emphasize "marginals." We use this confidence region in Step 1 of our schema.

When the marginals of $\mathbb{P}$ are known, Embrechts et al. [27] proves

$$\text{VaR}^{\mathbb{P}}_{\epsilon}\left(\mathbf{v}^T \tilde{\mathbf{u}}\right) \leq \min_{\boldsymbol{\lambda}: \mathbf{e}^T \boldsymbol{\lambda} = \epsilon} \sum_{i=1}^{d} \text{VaR}^{\mathbb{P}}_{\lambda_i}(v_i \tilde{u}_i) \leq \sum_{i=1}^{d} \text{VaR}^{\mathbb{P}}_{\epsilon/d}(v_i \tilde{u}_i) \tag{30}$$

where the last inequality is obtained by letting $\lambda_i = \epsilon/d$ for all $i$. From our schema,

**Theorem 8** *If $s$ defined by Eq. (29) satisfies $N - s + 1 < s$, then, with probability at least $1 - \alpha$ over the sample, the set*

$$\mathcal{U}^M_{\epsilon} = \left\{ \mathbf{u} \in \mathbb{R}^d : \hat{u}_i^{(N-s+1)} \leq u_i \leq \hat{u}_i^{(s)} \ \ i = 1, \ldots, d \right\}. \tag{31}$$

*implies a probabilistic guarantee for $\mathbb{P}^*$ at level $\epsilon$. Moreover,*

$$\delta^*(\mathbf{v} | \ \mathcal{U}^M_{\epsilon}) = \sum_{i=1}^{d} \max\left(v_i \hat{u}_i^{(N-s+1)}, v_i \hat{u}_i^{(s)}\right). \tag{32}$$

*Remark 13* Notice that the family $\{\mathcal{U}^M_{\epsilon} : 0 < \epsilon < 1\}$, may *not* simultaneously imply a probabilistic guarantee for $\mathbb{P}^*$ because the confidence region $\mathcal{P}^M$ depends on $\epsilon$.

*Remark 14* The set $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\mathcal{U}^M) \leq t\}$ is a simple box, representable by linear inequalities. From (32), a worst-case realization is given by $u_i^* = \hat{u}_i^{(s)} \mathbb{I}(v_i > 0) + \hat{u}_i^{(N-s+1)} \mathbb{I}(v_i < 0)$.

## 7 Uncertainty Sets for Potentially Non-independent Components

In this section, we assume we observe samples drawn from the joint distribution of $\mathbb{P}^*$ which may have unbounded support. We consider a goodness-of-fit hypothesis test based on linear-convex ordering proposed in [15]. Specifically, given some multivariate $\mathbb{P}_0$, consider the null-hypothesis $H_0 : \mathbb{P}^* = \mathbb{P}_0$. Bertsimas et al. [15] prove that the

test which rejects $H_0$ if there exists $(\mathbf{a}, b) \in \mathcal{B} \equiv \{\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R} : \|\mathbf{a}\|_1 + |b| \leq 1\}$ such that

$$\mathbb{E}^{\mathbb{P}_0}\left[\left(\mathbf{a}^T\tilde{\mathbf{u}} - b\right)^+\right] - \frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{a}^T\hat{\mathbf{u}}^j - b\right)^+$$
$$> \Gamma_{LCX} \quad \text{or} \quad \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\mathbf{u}}^j\right)^T\hat{\mathbf{u}}^j - \mathbb{E}^{\mathbb{P}_0}\left[\tilde{\mathbf{u}}^T\tilde{\mathbf{u}}\right] > \Gamma_\sigma$$

for appropriate thresholds $\Gamma_{LCX}, \Gamma_\sigma$ is a valid test at level $\alpha$. The authors provide an explicit bootstrap algorithm to compute $\Gamma_{LCX}, \Gamma_\sigma$ as well as exact formulae for upper-bounding these quantities.

The confidence region of this test is

$$\mathcal{P}^{LCX} = \Bigg\{\mathbb{P} \in \Theta(\mathbb{R}^d) : \mathbb{E}^{\mathbb{P}}\left[\left(\mathbf{a}^T\tilde{\mathbf{u}} - b\right)^+\right] \leq \frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{a}^T\hat{\mathbf{u}}_j - b\right)^+ + \Gamma_{LCX} \ \ \forall(\mathbf{a}, b) \in \mathcal{B},$$
$$\mathbb{E}^{\mathbb{P}}\left[\|\tilde{\mathbf{u}}\|^2\right] \geq \frac{1}{N}\sum_{j=1}^{N}\|\hat{\mathbf{u}}_j\|^2\Bigg] - \Gamma_\sigma\Bigg\}. \tag{33}$$

We use this confidence region in Step 1 of our schema. By explicitly computing the worst-case Value-at-Risk and applying our schema,

**Theorem 9** *The family* $\{\mathcal{U}_\epsilon^{LCX} : 0 < \epsilon < 1\}$ *simultaneously implies a probabilistic guarantee for* $\mathbb{P}^*$ *where*

$$\mathcal{U}_\epsilon^{LCX} = \Bigg\{\mathbf{u} \in \mathbb{R}^d : \exists \mathbf{r} \in \mathbb{R}^d,\ 1 \leq z \leq 1/\epsilon,\ \mathbf{s}^1, \mathbf{s}^2, \mathbf{s}^3 \in \mathbb{R}^N \ \ s.t.$$
$$\mathbf{0} \leq \mathbf{s}^k \leq \frac{z}{N}\mathbf{e}, \quad k = 1, 2, 3,$$
$$|z - \mathbf{e}^T\mathbf{s}^1| \leq \Gamma_{LCX}, \quad |(z-1) - \mathbf{e}^T\mathbf{s}^2| \leq \Gamma_{LCX}, \quad |1 - \mathbf{e}^T\mathbf{s}^3| \leq \Gamma_{LCX},$$
$$\|\mathbf{r} + \mathbf{u} - \sum_{j=1}^{n} s_j^1\hat{\mathbf{u}}_j\|_\infty \leq \Gamma_{LCX}, \quad \|\mathbf{r} - \sum_{j=1}^{n} s_j^2\hat{\mathbf{u}}_j\|_\infty \leq \Gamma_{LCX},$$
$$\|\mathbf{u} - \sum_{j=1}^{n} s_j^3\hat{\mathbf{u}}_j\|_\infty \leq \Gamma_{LCX}\Bigg\}. \tag{34}$$

*Moreover,* $\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^{LCX}) = \sup_{\mathbb{P}\in\mathcal{P}^{LCX}} VaR_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)$ *where*

$$\delta^*\left(\mathbf{v}|\,\mathcal{U}_\epsilon^{LCX}\right) = \min_{\tau,\theta,\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{y}^1,\mathbf{y}^2,\mathbf{y}^3} \ \frac{1}{\epsilon}\tau - \theta + \Gamma_{LCX}\|\boldsymbol{\alpha}\|_1 + 2\Gamma_{LCX}\|\boldsymbol{\beta}\|_1 + \Gamma_{LCX}\|\mathbf{v}+\boldsymbol{\beta}\|_1$$

$$
\begin{aligned}
s.t. \quad & -\theta + \tau + \alpha_1 + \alpha_2 = \frac{1}{N}\sum_{j=1}^{n} y_j^1 + y_j^2 + y_j^3 \\
& \alpha_1 - \beta^T \hat{u}_j \leq y_j^1, \;\; \alpha_2 + \beta^T \hat{u}_j \leq y_j^2, \;\; \alpha_3 + \beta^T \hat{u}_j + v^T \hat{u}_j \leq y_j^3, \\
& j = 1, \ldots, N, \\
& \tau, \theta \geq 0, \;\; \mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3 \geq \mathbf{0}
\end{aligned} \tag{35}
$$

*Remark 15* By adding auxiliary variables, we can represent $\mathcal{U}_\epsilon^{LCX}$ as the intersection of linear inequalities. Robust constraints over $\mathcal{U}_\epsilon^{LCX}$ are thus tractable.

*Remark 16* We stress that the robust constraint $\max_{\mathbf{u}\in\mathcal{U}_\epsilon^{LCX}} \mathbf{v}^T\mathbf{u}$ is exactly equivalent to the ambiguous chance-constraint $\sup_{\mathbb{P}\in\mathcal{P}^{LCX}} \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)$ above.

## 8 Hypothesis Testing: A Unifying Perspective

Several data-driven methods in the literature create families of measures $\mathcal{P}(\mathcal{S})$ that contain $\mathbb{P}^*$ with high probability. These methods do not explicitly reference hypothesis testing. In this section, we provide a hypothesis testing interpretation of two such methods [25,46]. Leveraging this new perspective, we show how standard techniques for hypothesis testing, such as the bootstrap, can be used to improve upon these methods. Finally, we illustrate how our schema can be applied to these improved family of measures to generate new uncertainty sets. To the best of our knowledge, generating uncertainty sets for (1) is a new application of both [25,46].

The key idea in both cases is to recast $\mathcal{P}(\mathcal{S})$ as the confidence region of a hypothesis test. This correspondence is not unique to these methods. There is a one-to-one correspondence between families of measures which contain $\mathbb{P}^*$ with probability at least $1-\alpha$ with respect to $\mathbb{P}_\mathcal{S}$ and the confidence regions of hypothesis tests. This correspondence is sometimes called the "duality between confidence regions and hypothesis testing" in the statistical literature [42]. It implies that any data-driven method predicated on a family of measures that contain $\mathbb{P}^*$ with probability $1-\alpha$ can be interpreted in the light of hypothesis testing.

This observation is interesting for two reasons. First, it provides a unified framework to compare distinct methods in the literature and ties them to the well-established theory of hypothesis testing in statistics. Secondly, there is a wealth of practical experience with hypothesis testing. In particular, we know empirically which tests are best suited to various applications and which tests perform well even when the underlying assumptions on $\mathbb{P}^*$ that motivated the test may be violated. In the next section, we leverage some of this practical experience with hypothesis testing to strengthen these methods, and then derive uncertainty sets corresponding to these hypothesis tests to facilitate comparison between the approaches.

### 8.1 Uncertainty Set Motivated by Cristianini and Shawe-Taylor 2003

Let $\|\cdot\|_F$ denote the Frobenius norm of matrices. In a particular machine learning context, Shawe-Taylor and Cristianini [46] prove

**Theorem 10** (Cristianini and Shawe-Taylor, 2003) *Suppose that supp*$(\mathbb{P}^*)$ *is contained within the ball of radius R and that* $N > (2 + 2\log(2/\alpha))^2$. *Then, with probability at least* $1 - \alpha$ *with respect to* $\mathbb{P}_{\mathcal{S}}$, $\mathbb{P}^* \in \mathcal{P}^{CS}(\Gamma_1(\alpha/2, N), \Gamma_2(\alpha/2, N))$, *where*

$$\mathcal{P}^{CS}(\Gamma_1, \Gamma_2) = \Big\{\mathbb{P} \in \Theta(R) : \|\mathbb{E}^{\mathbb{P}}\left[\tilde{\mathbf{u}}\right] - \hat{\boldsymbol{\mu}}\|_2 \leq \Gamma_1 \quad \text{and} \\ \|\mathbb{E}^{\mathbb{P}}\left[\tilde{\mathbf{u}}\tilde{\mathbf{u}}^T\right] - \mathbb{E}^{\mathbb{P}}\left[\tilde{\mathbf{u}}\right]\mathbb{E}^{\mathbb{P}}\left[\tilde{\mathbf{u}}^T\right] - \hat{\boldsymbol{\Sigma}}\|_F \leq \Gamma_2, \Big\}$$

*where* $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ *denote the sample mean and covariance,*

$$\Gamma_1(\alpha, N) = \frac{R}{\sqrt{N}}\left(2 + \sqrt{2\log 1/\alpha}\right), \quad \Gamma_2(\alpha, N) = \frac{2R^2}{\sqrt{N}}\left(2 + \sqrt{2\log 2/\alpha}\right),$$

*and* $\Theta(R)$ *denotes the set of Borel probability measures supported on the ball of radius R.*

We note that the key step in their proof utilizes a general purpose concentration inequality to compute $\Gamma_1(\alpha, N)$, $\Gamma_2(\alpha, N)$. (cf. [46, Theorem 1])

On the other hand, $\mathcal{P}^{CS}(\Gamma_1(\alpha/2, N), \Gamma_2(\alpha/2, N))$ is also the $1 - \alpha$ confidence region of a hypothesis test for the mean and covariance of $\mathbb{P}^*$. Namely, consider the null-hypothesis and test

$$H_0 : \mathbb{E}^{\mathbb{P}^*}[\tilde{\mathbf{u}}] = \boldsymbol{\mu}_0 \text{ and } \mathbb{E}^{\mathbb{P}^*}\left[\tilde{\mathbf{u}}\tilde{\mathbf{u}}^T\right] - \mathbb{E}^{\mathbb{P}^*}\left[\tilde{\mathbf{u}}\right]\mathbb{E}^{\mathbb{P}^*}\left[\tilde{\mathbf{u}}^T\right] = \boldsymbol{\Sigma}_0, \tag{36}$$

$$\text{Reject if } \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\| > \Gamma_1 \text{ or } \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0\| > \Gamma_2. \tag{37}$$

Theorem 10 proves that for $\Gamma_1 \to \Gamma_1(\alpha/2, N)$ and $\Gamma_2 \to \Gamma_2(\alpha/2, N)$, this is a valid test at level $\alpha$ and $\mathcal{P}^{CS}(\Gamma_1, \Gamma_2)$ is its confidence region.

Practical experience in applied statistics suggests, however, that tests whose thresholds are computed as above using general purpose concentration inequalities, while valid, are typically very conservative for reasonable values of $\alpha$, $N$. They reject $H_0$ when it is false only when $N$ is very large. The standard remedy is to use the bootstrap (Algorithm 1) to approximate thresholds $\Gamma_1^B$, $\Gamma_2^B$. These bootstrapped thresholds are typically much smaller than thresholds based on concentration inequalities, but are still (approximately) valid at level $1 - \alpha$. The first five columns of Table 2 illustrates the magnitude of the difference with a particular example. Entries of $\infty$ indicate that the threshold as derived in [46] does not apply for this value of $N$. The data are drawn from a standard normal distribution with $d = 2$ truncated to live in a ball of radius 9.2. We take $\alpha = 10\%$, $N_B = 10{,}000$. We can see that the reduction can be a full-order of magnitude, or more.

**Table 2** Comparing Thresholds with and without bootstrap using $N_B = 10{,}000$ replications, $\alpha = 10\%$

| | Shawe-Taylor and Cristianini [46] | | | | Delage and Ye [25] | | | |
|---|---|---|---|---|---|---|---|---|
| N | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_1^B$ | $\Gamma_2^B$ | $\gamma_1$ | $\gamma_2$ | $\gamma_1^B$ | $\gamma_2^B$ |
| 10 | $\infty$ | $\infty$ | 0.805 | 1.161 | $\infty$ | $\infty$ | 0.526 | 5.372 |
| 50 | $\infty$ | $\infty$ | 0.382 | 0.585 | $\infty$ | $\infty$ | 0.118 | 1.684 |
| 100 | 3.814 | 75.291 | 0.262 | 0.427 | $\infty$ | $\infty$ | 0.061 | 1.452 |
| 500 | 1.706 | 33.671 | 0.105 | 0.157 | $\infty$ | $\infty$ | 0.012 | 1.154 |
| 50,000 | 0.171 | 3.367 | 0.011 | 0.018 | $\infty$ | $\infty$ | 1e−4 | 1.015 |
| 100,000 | 0.121 | 2.381 | 0.008 | 0.013 | 0.083 | 5.044 | 6e−5 | 1.010 |

Reducing the thresholds $\Gamma_1, \Gamma_2$ shrinks $\mathcal{P}^{CS}(\Gamma_1, \Gamma_2)$. Thus, replacing $\Gamma_1(\alpha/2, N)$, $\Gamma_2(\alpha/2, N)$ by $\Gamma_1^B, \Gamma_2^B$ reduces the conservativeness of any method using $\mathcal{P}^{CS}$ (including the original machine learning application of Shawe-Taylor and Cristianini [46]) while retaining its robustness to ambiguity in $\mathcal{P}^*$ since $\Gamma_1^B, \Gamma_2^B$ are *approximately* valid thresholds which become exact as $N \to \infty$. Thus in applications where having a precise $1-\alpha$ guarantee is not necessary, or $N$ is very large, bootstrapped thresholds should be preferred.

We use $\mathcal{P}^{CS}(\Gamma_1, \Gamma_2)$ in Step 1 of our schema. In [18], the authors prove that for any $\Gamma_1, \Gamma_2$,

$$\sup_{\mathbb{P} \in \mathcal{P}^{CS}(\Gamma_1, \Gamma_2)} \text{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T \tilde{\mathbf{u}}\right) = \hat{\boldsymbol{\mu}}^T \mathbf{v} + \Gamma_1 \|\mathbf{v}\|_2 + \sqrt{\frac{1-\epsilon}{\epsilon}} \sqrt{\mathbf{v}^T \left(\hat{\boldsymbol{\Sigma}} + \Gamma_2 \mathbf{I}\right) \mathbf{v}}. \quad (38)$$

We translate this bound into an uncertainty set.

**Theorem 11** *Suppose $\Gamma_1, \Gamma_2$ are such that the test (37) is valid at level $\alpha$. With probability at least $1-\alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$, the family $\{\mathcal{U}_\epsilon^{CS} : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee for $\mathbb{P}^*$, where*

$$\mathcal{U}_\epsilon^{CS} = \left\{ \hat{\boldsymbol{\mu}} + \mathbf{y} + \mathbf{C}^T \mathbf{w} : \exists \mathbf{y}, \mathbf{w} \in \mathbb{R}^d \text{ s.t. } \|\mathbf{y}\| \le \Gamma_1, \quad \|\mathbf{w}\| \le \sqrt{\frac{1}{\epsilon} - 1} \right\}, \quad (39)$$

*where $\mathbf{C}^T\mathbf{C} = \hat{\boldsymbol{\Sigma}} + \Gamma_2 \mathbf{I}$ is a Cholesky decomposition. Moreover, $\delta^*(\mathbf{v} | \mathcal{U}_\epsilon^{CS})$ is given explicitly by the right-hand side of Eq. (38).*

*Remark 17* Notice that (38) is written with an *equality*. Thus, the robust constraint $\max_{\mathbf{u} \in \mathcal{U}_\epsilon^{CS}} \mathbf{v}^T \mathbf{x} \le 0$ is exactly equivalent to the ambiguous chance-constraint $\sup_{\mathbb{P} \in \mathcal{P}^{CS}(\Gamma_1, \Gamma_2)} \text{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T \tilde{\mathbf{u}}\right) \le 0$.

*Remark 18* From (38), $\{(\mathbf{v}, t) : \delta^*(\mathbf{v} | \mathcal{U}_\epsilon^{CS}) \le t\}$ is second order cone representable. Moreover, we can identify a worst-case realization in closed-form. Given $\mathbf{v}$, let $\mathbf{u}^* = \boldsymbol{\mu} + \frac{\Gamma_1}{\|\mathbf{v}\|}\mathbf{v} + \sqrt{\frac{1}{\epsilon} - 1}\frac{\mathbf{C}\mathbf{v}}{\|\mathbf{C}\mathbf{v}\|}$. Then $\mathbf{u}^* \in \arg\max_{\mathbf{u} \in \mathcal{U}_\epsilon^{CS}} \mathbf{v}^T \mathbf{u}$ (cf. Proof of Theorem 11).

*Remark 19* $\mathcal{U}_\epsilon^{CS}$ need not be a subset of supp($\mathbb{P}^*$). Consequently, when a priori knowledge of the support is available, we can refine this set as in Theorem 4.

To emphasize the benefits of bootstrapping when constructing uncertainty sets, Fig. 5 in the electronic companion illustrates the set $\mathcal{U}_\epsilon^{CS}$ for the example considered in Fig. 2 with thresholds computed with and without the bootstrap.

### 8.2 Uncertainty Set Motivated by Delage and Ye 2010

Delage and Ye [25] propose a data-driven approach for solving distributionally robust optimization problems. Their method relies on a slightly more general version of the following:[3]

**Theorem 12** (Delage and Ye [25]) *Let $R$ be such that $\mathbb{P}^*((\tilde{\mathbf{u}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\tilde{\mathbf{u}}-\boldsymbol{\mu}) \leq R^2) = 1$ where $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ are the true mean and covariance of $\tilde{\mathbf{u}}$ under $\mathbb{P}^*$. Let,*

$$\beta_2 \equiv \frac{R^2}{N}\left(2+\sqrt{2\log(2/\alpha)}\right)^2, \quad \beta_1 \equiv \frac{R^2}{\sqrt{N}}\left(\sqrt{1-\frac{d}{R^4}}+\sqrt{\log(4/\alpha)}\right),$$

*and suppose $N$ is large enough so that $1-\beta_1-\beta_2 > 0$. Finally suppose supp($\mathbb{P}^*$) $\subseteq [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$. Then with probability at least $1-\alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$, $\mathbb{P}^* \in \mathcal{P}^{DY}(\frac{\beta_2}{1-\beta_1-\beta_2}, \frac{1+\beta_2}{1-\beta_1-\beta_2})$ where*

$$\mathcal{P}^{DY}(\gamma_1,\gamma_2) \equiv \left\{\mathbb{P} \in \Theta\left[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}\right] : \left(\mathbb{E}^{\mathbb{P}}\left[\tilde{\mathbf{u}}\right]-\hat{\boldsymbol{\mu}}\right)^T\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbb{E}^{\mathbb{P}}\left[\tilde{\mathbf{u}}\right]-\hat{\boldsymbol{\mu}}\right) \leq \gamma_1, \right.$$
$$\left. \mathbb{E}^{\mathbb{P}}\left[\left(\tilde{\mathbf{u}}-\hat{\boldsymbol{\mu}}\right)\left(\tilde{\mathbf{u}}-\hat{\boldsymbol{\mu}}\right)^T\right] \preceq \gamma_2\hat{\boldsymbol{\Sigma}}\right\}.$$

The key idea is again to compute the thresholds using a general purpose concentration inequality. The condition on $N$ is required for the confidence region to be well-defined.

We again observe that $\mathcal{P}^{DY}(\gamma_1,\gamma_2)$ is the $1-\alpha$ confidence region of a hypothesis test. Again, consider the null-hypothesis (36) and the test

$$\text{Reject if } (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_0)^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_0) > \gamma_1 \text{ or } \max_{\boldsymbol{\lambda}} \frac{\boldsymbol{\lambda}^T\left(\boldsymbol{\Sigma}_0+(\boldsymbol{\mu}_0-\hat{\boldsymbol{\mu}})\left(\boldsymbol{\mu}_0-\hat{\boldsymbol{\mu}}\right)^T\right)\boldsymbol{\lambda}}{\boldsymbol{\lambda}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\lambda}} > \gamma_2. \tag{40}$$

Then, Theorem 12 proves that replacing $\gamma_1 \to \frac{\beta_2}{1-\beta_1-\beta_2}$ and $\gamma_2 \to \frac{1+\beta_2}{1-\beta_1-\beta_2}$ yields a test valid at level $\alpha$ whose confidence region is $\mathcal{P}^{DY}(\gamma_1,\gamma_2)$.

Again, these thresholds are calculated via a general purpose inequality. Instead, we can approximate new thresholds using the bootstrap. Table 2 shows the reduction in magnitude. Observe that the bootstrap thresholds exist for all $N$, not just $N$

[3] Specifically, since $R$ is typically unknown, the authors describe an estimation procedure for $R$ and prove a modified version of the Theorem 12 using this estimate and different constants. We treat the simpler case where $R$ is known here. Extensions to the other case are straightforward.

sufficiently large. Moreover, they are significantly smaller, so that $\mathcal{P}^{DY}(\gamma_1^B, \gamma_2^B)$ is significantly smaller than $\mathcal{P}^{DY}(\frac{\beta_2}{1-\beta_1-\beta_2}, \frac{1+\beta_2}{1-\beta_1-\beta_2})$, while retaining (approximately) the same probabilistic guarantee. Therefore, in applications where having a precise $1-\alpha$ guarantee is not necessary or $N$ is very large, they may be preferred. We use $\mathcal{P}^{DY}(\gamma_1^B, \gamma_2^B)$ in Step 1 of our schema.

**Theorem 13** *Let $\gamma_1, \gamma_2$ be such that the test (40) is valid at level $\alpha$. Suppose $supp(\mathbb{P}^*) \subset [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$. Then, with probability at least $1-\alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$, the family $\{\mathcal{U}_\epsilon^{DY} : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee for $\mathbb{P}^*$, where*

$$\begin{aligned}
\mathcal{U}_\epsilon^{DY} = \Big\{ \mathbf{u} \in \Big[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}\Big] : \exists \lambda \in \mathbb{R},\ \mathbf{w}, \mathbf{m} \in \mathbb{R}^d,\ \mathbf{A}, \hat{\mathbf{A}} \succeq \mathbf{0}\ s.t. \\
\lambda \leq \frac{1}{\epsilon},\ (\lambda-1)\hat{\mathbf{u}}^{(0)} \leq \mathbf{m} \leq (\lambda-1)\hat{\mathbf{u}}^{(N+1)},\ \lambda\hat{\boldsymbol{\mu}} \\
= \mathbf{m} + \mathbf{u} + \mathbf{w},\ \|\mathbf{C}\mathbf{w}\| \leq \lambda\sqrt{\gamma_1^B}, \\
\begin{pmatrix} \lambda - 1 & \mathbf{m}^T \\ \mathbf{m} & \mathbf{A} \end{pmatrix} \succeq \mathbf{0},\ \begin{pmatrix} 1 & \mathbf{u}^T \\ \mathbf{u} & \hat{\mathbf{A}} \end{pmatrix} \succeq \mathbf{0},\ \lambda\left(\gamma_2^B \hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T\right) \\
- \mathbf{A} - \hat{\mathbf{A}} - \mathbf{w}\hat{\boldsymbol{\mu}}^T - \hat{\boldsymbol{\mu}}\mathbf{w}^T \succeq \mathbf{0} \Big\},
\end{aligned} \tag{41}$$

*$C^T C = \hat{\boldsymbol{\Sigma}}^{-1}$ is a Cholesky-decomposition, and $\gamma_1^B, \gamma_2^B$ are computed by bootstrap. Moreover,*

$$\begin{aligned}
\delta^*\left(\mathbf{v} | \mathcal{U}_\epsilon^{DY}\right) = \sup_{\mathbb{P} \in \mathcal{P}^{DY}(\gamma_1^B, \gamma_2^B)} VaR_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) = \inf \quad t \\
s.t.\ \ r + s \leq \theta\epsilon, \\
\begin{pmatrix} r + \mathbf{y}_1^{+T}\hat{\mathbf{u}}^{(0)} - \mathbf{y}_1^{-T}\hat{\mathbf{u}}^{(N+1)} & \frac{1}{2}(\mathbf{q} - \mathbf{y}_1)^T, \\ \frac{1}{2}(\mathbf{q} - \mathbf{y}_1) & \mathbf{Z} \end{pmatrix} \succeq \mathbf{0}, \\
\begin{pmatrix} r + \mathbf{y}_2^{+T}\hat{\mathbf{u}}^{(0)} - \mathbf{y}_2^{-T}\hat{\mathbf{u}}^{(N+1)} + t - \theta & \frac{1}{2}(\mathbf{q} - \mathbf{y}_2 - \mathbf{v})^T, \\ \frac{1}{2}(\mathbf{q} - \mathbf{y}_2 - \mathbf{v}) & \mathbf{Z} \end{pmatrix} \succeq \mathbf{0}, \\
s \geq \left(\gamma_2^B \hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T\right) \circ \mathbf{Z} + \hat{\boldsymbol{\mu}}^T\mathbf{q} + \sqrt{\gamma_1^B}\|\mathbf{q} + 2\mathbf{Z}\hat{\boldsymbol{\mu}}\|_{\hat{\boldsymbol{\Sigma}}^{-1}}, \\
\mathbf{y}_1 = \mathbf{y}_1^+ - \mathbf{y}_1^-,\ \ \mathbf{y}_2 = \mathbf{y}_2^+ - \mathbf{y}_2^-,\ \ \mathbf{y}_1^+, \mathbf{y}_1^-, \mathbf{y}_2^+, \mathbf{y}_2^-, \theta \geq \mathbf{0}.
\end{aligned}$$

*Remark 20* Similar to $\mathcal{U}_\epsilon^{CS}$, the robust constraint $\max_{\mathbf{u} \in \mathcal{U}_\epsilon^{DY}} \mathbf{v}^T\mathbf{u} \leq 0$ is equivalent to the ambiguous chance constraint $\sup_{\mathbb{P} \in \mathcal{P}^{DY}(\gamma_1^B, \gamma_2^B)} \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) \leq 0$.

*Remark 21* The set $\{(\mathbf{v}, t) : \delta^*(\mathbf{v} | \mathcal{U}^{DY}) \leq t\}$ is representable as a linear matrix inequality. At time of writing, solvers for linear matrix inequalities are not as developed as those for second order cone programs. Consequently, one may prefer $\mathcal{U}_\epsilon^{CS}$ to $\mathcal{U}_\epsilon^{DY}$ in practice for its simplicity.

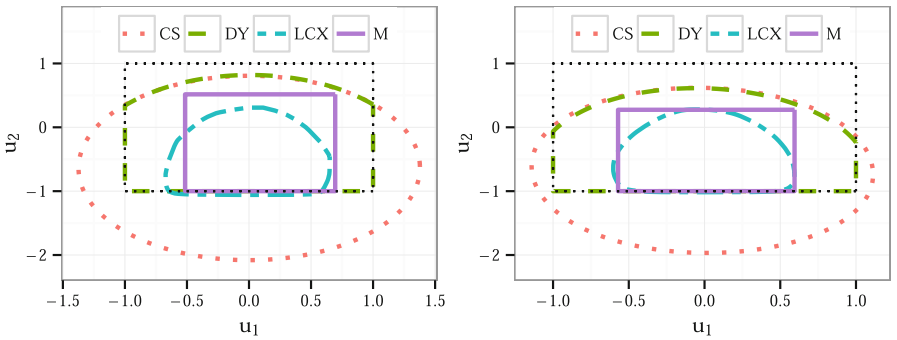


**Fig. 3** Comparing $\mathcal{U}_\epsilon^M, \mathcal{U}_\epsilon^{LCX}, \mathcal{U}_\epsilon^{CS}$ and $\mathcal{U}_\epsilon^{DY}$ for the example from Fig. 2, $\epsilon = 10\%$, $\alpha = 20\%$. The *black dotted line* represents supp($\mathbb{P}^*$). The *left panel* uses $N = 100$ data points, while the *right panel* uses $N = 1000$ data points

### 8.3 Comparing $\mathcal{U}_\epsilon^M, \mathcal{U}_\epsilon^{LCX}, \mathcal{U}_\epsilon^{CS}$ and $\mathcal{U}_\epsilon^{DY}$

One of the benefits of deriving uncertainty sets corresponding to the methods of Delage and Ye [25] and Shawe-Taylor and Cristianini [46] is that it facilitates comparisons between these methods and our own proposals. In particular, we can make visual, qualitative assessments of the conservatism (in terms of size) and modeling power (in terms of shape). In Fig. 3, we illustrate the sets $\mathcal{U}_\epsilon^M, \mathcal{U}_\epsilon^{LCX}, \mathcal{U}_\epsilon^{CS}$ and $\mathcal{U}_\epsilon^{DY}$ for the same numerical example from Fig. 2. Note that each of these sets implies a probabilistic guarantee when data are drawn i.i.d. from a general joint distribution. Because $\mathcal{U}^M$ does not leverage the joint distribution $\mathbb{P}^*$, it does not learn that its marginals are independent. Consequently, $\mathcal{U}^M$ has pointed corners permitting extreme values of both coordinates simultaneously. The remaining sets do learn the marginal independence from the data and, hence, have rounded corners.

Interestingly, $\mathcal{U}_\epsilon^{CS} \cap \text{supp}(\mathbb{P}^*)$ is very similar to $\mathcal{U}_\epsilon^{DY}$ for this example (indistinguishable in picture). Since $\mathcal{U}^{CS}$ and $\mathcal{U}^{DY}$ only depend on the first two moments of $\mathbb{P}^*$, neither is able to capture the skewness in the second coordinate. Finally, $\mathcal{U}^{LCX}$ is contained within supp($\mathbb{P}^*$) and displays symmetry in the first coordinate and skewness in the second. In this example it is also the smallest set (in terms of volume). All sets shrink as $N$ increases.

### 8.4 Refining $\mathcal{U}_\epsilon^{FB}$

Another common approach to hypothesis testing in applied statistics is to use tests designed for Gaussian data that are "robust to departures from normality." The best known example of this approach is the $t$ test from Sect. 2.2, for which there is a great deal of experimental evidence to suggest that the test is still approximately valid when the underlying data are non-Gaussian [35, Chapt. 11.3]. Moreover, certain nonparametric tests of the mean for non-Gaussian data are asymptotically equivalent to the $t$ test, so that the $t$ test, itself, is asymptotically valid for non-Gaussian data [35, p.180]. Consequently, the $t$ test is routinely used in practice, even when the Gaussian assumption may be invalid.

We use the $t$ test in combination with bootstrapping to refine $\mathcal{U}_\epsilon^{FB}$. We replace $m_{fi}, m_{bi}$ in Eq. (27), with the upper and lower thresholds of a $t$ test at level $\alpha'/2$. We expect these new thresholds to correctly bound the true mean $\mu_i$ with probability approximately $1-\alpha'/2$ with respect to the data. We then use the bootstrap to calculate bounds on the forward and backward deviations $\overline{\sigma}_{fi}, \overline{\sigma}_{bi}$.

We stress not all tests designed for Gaussian data are robust to departures from normality. Applying Gaussian tests that lack this robustness will likely yield poor performance. Consequently, some care must be taken when choosing an appropriate test.

## 9 Implementation Details and Applications

### 9.1 Choosing the "Right" Set and Tuning $\alpha$, $\epsilon$

Choosing an appropriate set from amongst those consistent with the a priori knowledge of $\mathbb{P}^*$ is a non-trivial task that depends on the application, data and $N$. In what follows, we adapt classical model selection procedures from machine learning by viewing a robust optimal solution $\mathbf{x}^*$ as analogous to a fitted parameter in a statistical model. There are, of course, a wide-variety of common model selection procedures (see [2,31]), some of which may be more appropriate to the specific application than others. Perhaps the simplest approach is to split the data into two parts, a training set and a hold-out set. Use the training set to construct each potential uncertainty set, in turn, and solve the robust optimization problem. Evaluate each of the corresponding solutions out-of-sample on the hold-out set, and select the best solution. ("Best" may be interpreted in an application specific way.) When choosing among $k$ sets that each imply a probabilistic guarantee at level $\epsilon$ with probability $1-\alpha$, this procedure will yield a set that satisfies a probabilistic guarantee at level $\epsilon$ with probability at least $1-k\alpha$ by the union bound.

In situations where $N$ is only moderately large and using only half the data to calibrate an uncertainty set is impractical, we suggest using $k$-fold cross-validation to select a set (see [31] for a review of cross-validation). Unlike the above procedure, we cannot prove that the set chosen by $k$-fold cross-validation implies a probabilistic guarantee. Nevertheless, experience in machine learning suggests cross-validation is extremely effective. In what follows, we use fivefold cross-validation to select our sets.

As an aside, we point out that in applications where there is no natural choice for $\alpha$ or $\epsilon$, similar techniques can also be used to tune these parameters. Namely, solve the model over a grid of potential values for $\alpha$ and/or $\epsilon$ and then select the best value either using a hold-out set or cross-validation. Since the optimal value likely depends on the choice of uncertainty set, we suggest choosing the set and these parameters jointly.

### 9.2 Applications

We demonstrate how our new sets may be used in two applications: portfolio management and queueing theory. Our goals are to, first, illustrate their application and, second, to compare them to one another. We summarize our major insights:

– In these two applications, our data-driven sets outperform traditional, non-data driven uncertainty sets, and, moreover, robust models built with our sets perform as well or better than other data-driven approaches.
– Although our data-driven sets all shrink as $N \to \infty$, they learn different features of $\mathbb{P}^*$, such as correlation structure and skewness. Consequently, different sets may be better suited to different applications, and the right choice of set may depend on $N$. Cross-validation effectively identifies the best set.
– Optimizing the $\epsilon_j$'s in the case of multiple constraints can significantly improve performance.

Because of space considerations, we treat only the portfolio management application in the main text. The queueing application can be found in "Appendix 4".

### 9.3 Portfolio Management

Portfolio management has been well-studied in the robust optimization literature [19, 29,39]. For simplicity, we will consider the one period allocation problem:

$$\max_{\mathbf{x}} \left\{ \min_{\mathbf{r} \in \mathcal{U}} \ \mathbf{r}^T \mathbf{x} : \ \mathbf{e}^T \mathbf{x} = 1, \ \ \mathbf{x} \geq \mathbf{0} \right\}, \tag{42}$$

which seeks the portfolio $\mathbf{x}$ with maximal worst-case return over the set $\mathcal{U}$. If $\mathcal{U}$ implies a probabilistic guarantee for $\mathbb{P}^*$ at level $\epsilon$, then the optimal value $z^*$ of this optimization is a conservative bound on the $\epsilon$-worst case return for the optimal solution $\mathbf{x}^*$.

We consider a synthetic market with $d = 10$ assets. Returns are generated according to the following model from [39]:

$$\tilde{r}_i = \begin{cases} \frac{\sqrt{(1-\beta_i)\beta_i}}{\beta_i} & \text{with probability } \beta_i \\ -\frac{\sqrt{(1-\beta_i)\beta_i}}{1-\beta_i} & \text{with probability } 1-\beta_i \end{cases}, \quad \beta_i = \frac{1}{2}\left(1 + \frac{i}{11}\right), \ \ i = 1, \ldots, 10. \tag{43}$$

In this model, all assets have the same mean return (0%), the same standard deviation (1.00%), but have different skew and support. Higher indexed assets are highly skewed; they have a small probability of achieving a very negative return. Returns for different assets are independent. We simulate $N = 500$ returns as data.

We will utilize our sets $\mathcal{U}_\epsilon^M$ and $\mathcal{U}_\epsilon^{LCX}$ in this application. We do not consider the sets $\mathcal{U}_\epsilon^I$ or $\mathcal{U}_\epsilon^{FB}$ since we do not know a priori that the returns are independent. To contrast to the methods of Delage and Ye [25] and Shawe-Taylor and Cristianini [46] we also construct the sets $\mathcal{U}_\epsilon^{CS}$ and $\mathcal{U}_\epsilon^{DY}$. Recall from Remarks 17 and 20 that robust linear constraints over these sets are equivalent to ambiguous chance-constraints in the original methods, but with improved thresholds. As discussed in Remark 19, we also construct $\mathcal{U}_\epsilon^{CS} \cap \text{supp}(\mathbb{P}^*)$ for comparison. We use $\alpha = \epsilon = 10\%$ in all of our sets. Finally, we will also compare to the method of Calafiore and Monastero [19] (denoted "CM" in our plots), which is not an uncertainty set based method. We calibrate this

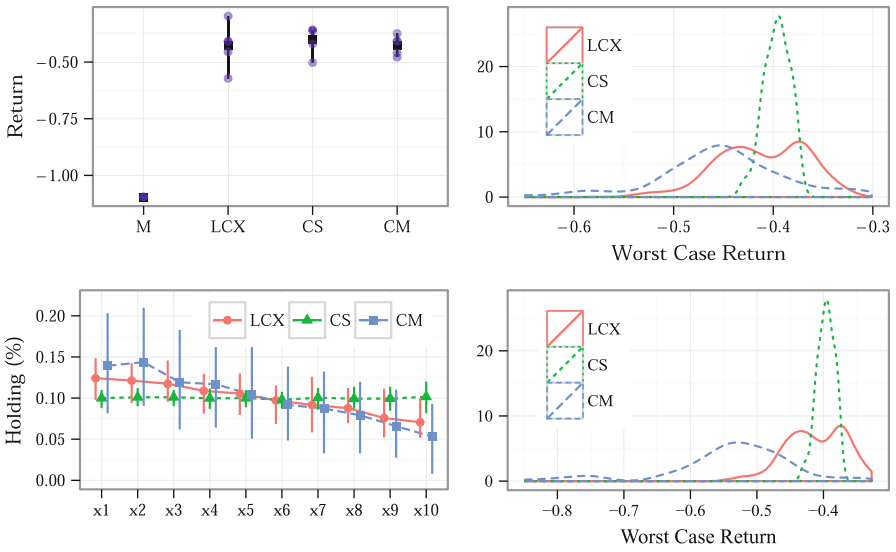


**Fig. 4** Portolio performance by method: $\alpha = \epsilon = 10\%$. *Top left* Cross-validation results. *Top right* Out-of-sample distribution of the 10% worst-case return over 100 runs. *Bottom left* Average portfolio holdings by method. *Bottom right* Out-of-sample distribution of the 10% worst-case return over 100 runs. The *bottom right panel* uses $N = 2000$. The remainder use $N = 500$

**Table 3** Portfolio statistics for each of our methods

| | $N = 500$ | | | | $N = 2000$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $z_{In}$ | CV | $z_{Out}$ | $z_{Avg}$ | $z_{In}$ | CV | $z_{Out}$ | $z_{Avg}$ |
| M | −1.095 | −1.095 | −1.095 | −1.095 | −1.095 | −1.095 | −1.095 | −1.095 |
| LCX | −0.699 | −0.373 | −0.373 | −0.411 | −0.89 | −0.428 | −0.395 | −0.411 |
| CS | −1.125 | −0.403 | −0.416 | −0.397 | −1.306 | −0.400 | −0.417 | −0.396 |
| CM | −0.653 | −0.495 | −0.425 | −0.539 | −0.739 | −0.426 | −0.549 | −0.451 |

$\mathcal{U}_\epsilon^{DY}$ and $\mathcal{U}_\epsilon^{CS} \cap \text{supp}(\mathbb{P}^*)$ perform identically to $\mathcal{U}_\epsilon^{M}$. "CM" refers to the method of [19]

method to also provide a bound on the 10% worst-case return that holds with at least 90% with respect to $\mathbb{P}_S$ so as to provide a fair comparison.

We first consider the problem of selecting an appropriate set via 5-fold cross-validation. The top left panel in Fig. 4 shows the out-of-sample 10% worst-case return for each of the 5 runs (blue dots), as well as the average performance on the 5 runs for each set (black square). Sets $\mathcal{U}_\epsilon^M$, $\mathcal{U}_\epsilon^{CS} \cap \text{supp}(\mathbb{P}^*)$ and $\mathcal{U}_\epsilon^{DY}$ yield identical portfolios (investing everything in the first asset) so we only include $\mathcal{U}^M$ in our graphs. The average performance is also shown in Table 3 under column CV (for "cross-validation.") The optimal objective value of (42) for each of our sets (trained with the entire data set) is shown in column $z_{In}$.

Based on the top left panel of Fig. 4, it is clear that $\mathcal{U}_\epsilon^{LCX}$ and $\mathcal{U}_\epsilon^{CS}$ significantly outperform the remaining sets. They seem to perform similarly to the CM method. Consequently, we would choose one of these two sets in practice.

We can assess the quality of this choice by using the ground-truth model (43) to calculate the true 10% worst-case return for each of the portfolios. These are shown in Table 3 under column $z_{Out}$. Indeed, these sets perform better than the alternatives, and, as expected, the cross-validation estimates are reasonably close to the true out-of-sample performance. By contrast, the in-sample objective value $z_{In}$ is a loose bound. We caution against using this in-sample value to select the best set.

Interestingly, we point out that while $\mathcal{U}_\epsilon^{CS} \cap \text{supp}(\mathbb{P}^*)$ is potentially smaller (with respect to subset containment) than $\mathcal{U}_\epsilon^{CS}$, it performs much worse out-of-sample (it performs identically to $\mathcal{U}_\epsilon^{M}$). This experiment highlights the fact that size calculations alone cannot predict performance. Cross-validation or similar techniques are required.

One might ask if these results are specific to the particular draw of 500 data points we use. We repeat the above procedure 100 times. The resulting distribution of 10% worst-case return is shown in the top right panel of Fig. 4 and the average of these runs is shown Table 3 under column $z_{Avg}$. As might have been guessed from the cross-validation results, $\mathcal{U}_\epsilon^{CS}$ delivers more stable and better performance than either $\mathcal{U}_\epsilon^{LCX}$ or CM. $\mathcal{U}_\epsilon^{LCX}$ slightly outperforms CM, and its distribution is shifted right.

We next look at the distribution of actual holdings between these methods. We show the average holding across these 100 runs as well as 10% and 90% quantiles for each asset in the bottom left panel of Fig. 4. Since $\mathcal{U}_\epsilon^{M}$ does not use the joint distribution, it sees no benefit to diversification. Portfolios built from $\mathcal{U}_\epsilon^{M}$ consistently hold all their wealth in the first asset over all the runs, hence, they are omitted from graphs. The set $\mathcal{U}_\epsilon^{CS}$ depends only on the first two moments of the data, and, consequently, cannot distinguish between the assets. It holds a very stable portfolio of approximately the same amount in each asset. By contrast, $\mathcal{U}^{LCX}$ is able to learn the asymmetry in the distributions, and holds slightly less of the higher indexed (toxic) assets. CM is similar to $\mathcal{U}^{LCX}$, but demonstrates more variability in the holdings.

We point out that the performance of each method depends slightly on $N$. We repeat the above experiments with $N = 2000$. Results are summarized in Table 3. The bottom right panel of Fig. 4 shows the distribution of the 10% worst-case return. (Additional plots are also available in "Additional Portfolio Results" in Appendix.) Both $\mathcal{U}^{LCX}$ and CM perform noticeably better with the extra data, but $\mathcal{U}^{LCX}$ now noticeably outperforms CM and its distribution is shifted significantly to the right.

## 10 Conclusions

The prevalence of high quality data is reshaping operations research. Indeed, a new data-centered paradigm is emerging. In this work, we took a step towards adapting traditional robust optimization techniques to this new paradigm. Specifically, we proposed a novel schema for designing uncertainty sets for robust optimization from data using hypothesis tests. Sets designed using our schema imply a probabilistic guarantee and are typically much smaller than corresponding data poor variants. Models built from these sets are thus less conservative than conventional robust approaches, yet retain the same robustness guarantees.

**Acknowledgements** We would like to thank the area editor, associate editor and two anonymous reviewers for their helpful comments on an earlier draft of this manuscript. Part of this work was supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374.

## Appendix 1: Omitted Proofs

### Proof of Theorem 1

*Proof* For the first part, let $\mathbf{x}^*$ be such that $f(\mathbf{u}, \mathbf{x}^*) \leq 0$ for all $\mathbf{u} \in \mathcal{U}_\epsilon$, and consider the closed, convex set $\{\mathbf{u} \in \mathbb{R}^d : f(\mathbf{u}, \mathbf{x}^*) \geq t\}$ where $t > 0$. That $\mathbf{x}^*$ is robust feasible implies $\max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{u}, \mathbf{x}^*) \leq 0$ which implies that $\mathcal{U}$ and $\{\mathbf{u} \in \mathbb{R}^d : f(\mathbf{u}, \mathbf{x}^*) \geq t\}$ are disjoint. From the separating hyperplane theorem, there exists a strict separating hyperplane $\mathbf{v}^T\mathbf{u} = v_0$ such that $v_0 > \mathbf{v}^T\mathbf{u}$ for all $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v}^T\mathbf{u} > v_0$ for all $\mathbf{u} \in \{\mathbf{u} \in \mathbb{R}^d : f(\mathbf{u}, \mathbf{x}^*) \geq t\}$. Observe

$$v_0 > \max_{\mathbf{u}\in\mathcal{U}} \mathbf{v}^T\mathbf{u} = \delta^*(\mathbf{v}|\,\mathcal{U}) \geq \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right),$$

and

$$\mathbb{P}\left(f\left(\tilde{\mathbf{u}}, \mathbf{x}^*\right) \geq t\right) \leq \mathbb{P}\left(\mathbf{v}^T\tilde{\mathbf{u}} > v_0\right) \leq \mathbb{P}\left(\mathbf{v}^T\tilde{\mathbf{u}} > \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)\right) \leq \epsilon.$$

Taking the limit as $t \downarrow 0$ and using the continuity of probability proves $\mathbb{P}(f(\tilde{\mathbf{u}}, \mathbf{x}^*) > 0) \leq \epsilon$ and that (2) is satisfied.

For the second part of the theorem, let $t > 0$ be such that $\delta^*(\mathbf{v}|\,\mathcal{U}) \leq \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) - t$. Define $f(\mathbf{u}, x) \equiv \mathbf{v}^T\mathbf{u} - x$. Then $x^* = \delta^*(\mathbf{v}|\,\mathcal{U})$ satisfies $f(\mathbf{u}, \mathbf{x}^*) \leq 0$ for all $\mathbf{u} \in \mathcal{U}$, but

$$\mathbb{P}\left(f(\tilde{\mathbf{u}}, \mathbf{x}) > 0\right) = \mathbb{P}\left(\mathbf{v}^T\tilde{\mathbf{u}} > \delta^*\left(\mathbf{v}|\,\mathcal{U}\right)\right) \geq \mathbb{P}\left(\mathbf{v}^T\tilde{\mathbf{u}} \geq \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) - t\right) > \epsilon$$

by (7). Thus, (2) does not hold. □

### Proofs of Theorems 2–4

*Proof of Theorem 2*

$$\begin{aligned}
&\mathbb{P}^*_{\mathcal{S}}\Big(\mathcal{U}(\mathcal{S}, \epsilon, \alpha) \text{ implies a probabilistic guarantee at level} \epsilon \text{ for } \mathbb{P}^*\Big) \\
&= \mathbb{P}^*_{\mathcal{S}}\left(\delta^*\left(\mathbf{v}|\,\mathcal{U}(\mathcal{S}, \epsilon, \alpha)\right) \geq \mathrm{VaR}_\epsilon^{\mathbb{P}^*}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)\ \forall \mathbf{v} \in \mathbb{R}^d\right) && \text{(Theorem 1)} \\
&\geq \mathbb{P}^*_{\mathcal{S}}\left(\mathbb{P}^* \in \mathcal{P}(\mathcal{S}, \epsilon, \alpha)\right) && \text{(Step 2 of schema)} \\
&\geq 1 - \alpha && \text{(Confidence region).}
\end{aligned}$$

□

*Proof* For the first part,

$$
\begin{aligned}
&\mathbb{P}^*_{\mathcal{S}}(\{\mathcal{U}(\mathcal{S},\epsilon,\alpha):\ 0<\epsilon<1\} \text{ simultaneously implies a probabilistic guarantee})\\
&\quad=\mathbb{P}^*_{\mathcal{S}}(\delta^*(\mathbf{v}|\,\mathcal{U}(\mathcal{S},\epsilon,\alpha))\geq \mathrm{VaR}^{\mathbb{P}^*}_{\epsilon}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)\ \forall \mathbf{v}\in\mathbb{R}^d,\ 0<\epsilon<1) && \text{(Theorem 1)}\\
&\quad\geq \mathbb{P}^*_{\mathcal{S}}(\mathbb{P}^*\in\bigcap_{\epsilon:0<\epsilon<1}\mathcal{P}(\mathcal{S},\epsilon,\alpha)) && \text{(Step 2 of schema)}\\
&\quad=\mathbb{P}^*_{\mathcal{S}}(\mathbb{P}^*\in\mathcal{P}(\mathcal{S},\alpha)) && ((\mathcal{P}(\mathcal{S},\alpha)) \text{ is independent of } \epsilon)\\
&\quad\geq 1-\alpha && \text{(Confidence region).}
\end{aligned}
$$

For the second part, let $\epsilon_1,\ldots,\epsilon_m$ denote any feasible $\epsilon_j$'s in (10).

$$
\begin{aligned}
1-\alpha &\leq \mathbb{P}^*_{\mathcal{S}}\left(\{\mathcal{U}\left(\mathcal{S},\epsilon,\alpha\right):\ 0<\epsilon<1\} \text{ simultaneously implies a probabilistic guarantee}\right)\\
&\leq \mathbb{P}^*_{\mathcal{S}}\left(\mathcal{U}\left(\mathcal{S},\epsilon_j,\alpha\right) \text{ implies a probabilistic guarantee at level } \epsilon_j,\ j=1,\ldots,m\right).
\end{aligned}
$$

Applying the union-bound and Theorem 2 yields the result. □

*Proof of Theorem 4* Consider the first statement. By Theorem 1, $\mathrm{VaR}^{\mathbb{P}}_{\epsilon}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)\leq \delta^*(\mathbf{v}|\mathcal{U}_\epsilon)$. Moreover, since $\mathrm{supp}(\mathbb{P}^*)\subseteq\mathcal{U}_0$, $0=\mathbb{P}(\mathbf{v}^T\tilde{\mathbf{u}}>\max_{\mathbf{u}\in\mathrm{supp}(\mathbb{P}^*)}\mathbf{v}^T\mathbf{u})\geq \mathbb{P}(\mathbf{v}^T\tilde{\mathbf{u}}>\max_{\mathbf{u}\in\mathcal{U}_0}\mathbf{v}^T\mathbf{u})=\mathbb{P}(\mathbf{v}^T\tilde{\mathbf{u}}>\delta^*(\mathbf{v}|\,\mathcal{U}_0))$ This, in turn, implies $\mathrm{VaR}^{\mathbb{P}}_{\epsilon}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)\leq \delta^*(\mathbf{v}|\mathcal{U}_0)$. Combining, we have $\mathrm{VaR}^{\mathbb{P}}_{\epsilon}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)\leq \min\left(\delta^*(\mathbf{v}|\mathcal{U}_\epsilon),\delta^*(\mathbf{v}|\mathcal{U}_0)\right)=\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon\cap\mathcal{U}_0)$ where the last equality follows because both $\mathcal{U}_0$ and $\mathcal{U}_\epsilon$ are convex. Thus, $\mathcal{U}_\epsilon\cap\mathcal{U}_0$ implies a probabilistic guarantee by Theorem 1. The second statement is entirely similar. □

## Proof of Theorem 5 and Proposition 1

We require the following well-known result.

**Theorem 14** (Rockafellar and Ursayev [43]) *Suppose* $supp(\mathbb{P})\subseteq\{\mathbf{a}_0,\ldots,\mathbf{a}_{n-1}\}$ *and let* $\mathbb{P}(\tilde{\mathbf{u}}=\mathbf{a}_j)=p_j$. *Let*

$$
\mathcal{U}^{CVaR^{\mathbb{P}}_{\epsilon}}=\left\{\mathbf{u}\in\mathbb{R}^d:\mathbf{u}=\sum_{j=0}^{n-1}q_j\mathbf{a}_j,\ \mathbf{q}\in\Delta_n,\ \mathbf{q}\leq\frac{1}{\epsilon}\mathbf{p}\right\}. \tag{44}
$$

*Then,* $\delta^*(\mathbf{v}|\,\mathcal{U}^{CVaR^{\mathbb{P}}_{\epsilon}})=CVaR^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)$.

We now prove the theorem.

*Proof of Theorem 5* We prove the theorem for $\mathcal{U}^{\chi^2}_{\epsilon}$. The proof for $\mathcal{U}^{G}_{\epsilon}$ is similar. From Theorem 2, it suffices to show that $\delta^*(\mathbf{v}|\,\mathcal{U}^{\chi^2}_{\epsilon})$ is an upper bound to $\sup_{\mathbb{P}\in\mathcal{P}^{\chi^2}}\mathrm{VaR}^{\mathbb{P}}_{\epsilon}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)$:

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}^{\chi^2}} \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) &\le \sup_{\mathbb{P}\in\mathcal{P}^{\chi^2}} \mathrm{CVaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) && \left(\mathrm{CVaR}_\epsilon^{\mathbb{P}} \text{ is an upper bound to } \mathrm{VaR}_\epsilon^{\mathbb{P}}\right)\\ &= \sup_{\mathbb{P}\in\mathcal{P}^{\chi^2}} \max_{\mathbf{u}\in\mathcal{U}^{\mathrm{CVaR}_\epsilon^{\mathbb{P}}}} \mathbf{v}^T\mathbf{u} && \text{(Theorem 14)}\\ &= \max_{\mathbf{u}\in\mathcal{U}_\epsilon^{\chi^2}} \mathbf{v}^T\mathbf{u} && \text{(Combining Eqs. (15) and (13)).}\end{aligned}$$

To obtain the expression for $\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^{\chi^2})$ observe,

$$\delta^*\left(\mathbf{v}|\,\mathcal{U}_\epsilon^{\chi^2}\right) = \inf_{\mathbf{w}\ge 0}\left\{\max_{\mathbf{q}\in\Delta_n}\sum_{i=0}^{n-1} q_i\left(\mathbf{a}_i^T\mathbf{v}-w_i\right)+\frac{1}{\epsilon}\max_{\mathbf{p}\in\mathcal{P}^{\chi^2}}\mathbf{w}^T\mathbf{p}\right\},$$

from Lagrangian duality. The optimal value of the first maximization is $\beta = \max_i \mathbf{a}_i^T\mathbf{v} - w_i$. The second maximization is of the form studied in [9, Corollary 1] and has optimal value

$$\inf_{\lambda\ge 0,(\lambda+\eta)\mathbf{e}\ge\mathbf{w},\eta}\eta+\frac{\lambda\chi^2_{n-1,1-\alpha}}{N}+2\lambda-2\sum_{i=0}^{n-1}\hat{p}_i\sqrt{\lambda}\sqrt{\lambda+\eta-w_i}.$$

Introduce the auxiliary variables $s_i$, such that $s_i^2 \le \lambda\cdot(\lambda+\eta-w_i)$. This last constraint is of hyperbolic type. Using [37], we can rewrite

$$\begin{aligned}&s_i^2 \le \lambda\cdot(\lambda+\eta-w_i), \quad i=0,\ldots,n-1,\\ &\lambda+\eta\ge w_i, \quad i=0,\ldots,n-1,\\ &\lambda\ge 0\end{aligned} \quad\Leftrightarrow\quad \left\|\begin{matrix}2s_i\\ \eta-w_i\end{matrix}\right\| \le 2\lambda+\eta-w_i, \quad i=0,\ldots,n-1.$$

Substituting these constraints (and the auxiliary variable) above yields the given formulation. □

*Proof of Proposition 1* Let $\Delta_j \equiv \frac{\hat{p}_j-p_j}{p_j}$. Then, $D(\hat{\mathbf{p}},\mathbf{p}) = \sum_{j=0}^{n-1}\hat{p}_j\log(\hat{p}_j/p_j) = \sum_{j=0}^{n-1} p_j(\Delta_j+1)\log(\Delta_j+1)$. Using a Taylor expansion of $x\log x$ around $x=1$ yields,

$$D\left(\hat{\mathbf{p}},\mathbf{p}\right) = \sum_{j=0}^{n-1} p_j\left(\Delta_j+\frac{\Delta_j^2}{2}+O\left(\Delta_j^3\right)\right) = \sum_{j=0}^{n-1}\frac{\left(\hat{p}_j-p_j\right)^2}{2p_j}+\sum_{j=0}^{n-1}O\left(\Delta_j^3\right), \tag{45}$$

where the last equality follows by expanding out terms and observing that $\sum_{j=0}^{n-1}\hat{p}_j = \sum_{j=0}^{n-1} p_j = 1$. Next, note $\mathbf{p}\in\mathcal{P}^G \implies \hat{p}_j/p_j \le \exp(\frac{\chi^2_{n-1,1-\alpha}}{2N\hat{p}_j})$. From the Strong Law of Large Numbers, for any $0<\alpha'<1$, there exists $M$ such that for all $N>M$, $\hat{p}_j \ge p_j^*/2$ with probability at least $1-\alpha'$ for all $j=0,\ldots,n-1$, simultaneously. It follows that for $N$ sufficiently large, with probability $1-\alpha'$, $\mathbf{p}\in\mathcal{P}^G \implies \hat{p}_j/p_j \le$

$\exp(\frac{\chi^2_{n-1,1-\alpha}}{Np_j^*})$ which implies that $|\Delta_j| \leq \exp(\frac{\chi^2_{n-1,1-\alpha}}{Np_j^*}) - 1 = O(N^{-1})$. Substituting into (45) completes the proof. □

**Proof of Theorems 6 and 7**

We first prove the following auxiliary result that will allow us to evaluate the inner supremum in (19).

**Theorem 15** *Suppose $g(u)$ is monotonic. Then,*

$$\sup_{\mathbb{P}_i \in \mathcal{P}_i^{KS}} \mathbb{E}^{\mathbb{P}_i}\left[g\left(\tilde{u}_i\right)\right] = \max\left(\sum_{j=0}^{N+1} q_j^L\left(\Gamma^{KS}\right) g\left(\hat{u}_i^{(j)}\right), \sum_{j=0}^{N+1} q_j^R\left(\Gamma^{KS}\right) g\left(\hat{u}_i^{(j)}\right)\right) \tag{46}$$

*Proof* Observe that the discrete distribution which assigns mass $q_j^L\left(\Gamma^{KS}\right)$ (resp. $q_j^R\left(\Gamma^{KS}\right)$) to the point $\hat{u}^{(j)}$ for $j = 0, \ldots, N+1$ is an element of $\mathcal{P}_i^{KS}$. Thus, Eq. (46) holds with "=" replaced by "≥".

For the reverse inequality, we have two cases. Suppose first that $g(u_i)$ is non-decreasing. Given $\mathbb{P}_i \in \mathcal{P}_i^{KS}$, consider the measure $\mathbb{Q}$ defined by

$$\begin{aligned}
&\mathbb{Q}\left(\tilde{u}_i = \hat{u}_i^{(0)}\right) \equiv 0, \quad \mathbb{Q}\left(\tilde{u}_i = \hat{u}_i^{(1)}\right) \equiv \mathbb{P}_i\left(\hat{u}_i^{(0)} \leq \tilde{u}_i \leq \hat{u}_i^{(1)}\right), \\
&\mathbb{Q}\left(\tilde{u}_i = \hat{u}_i^{(j)}\right) \equiv \mathbb{P}_i\left(\hat{u}_i^{(j-1)} < \tilde{u}_i \leq \hat{u}_i^{(j)}\right), \quad j = 2, \ldots, N+1.
\end{aligned} \tag{47}$$

Then, $\mathbb{Q} \in \mathcal{P}^{KS}$, and since $g(u_i)$ is non-decreasing, $\mathbb{E}^{\mathbb{P}_i}[g(\tilde{u}_i)] \leq \mathbb{E}^{\mathbb{Q}}[g(\tilde{u}_i)]$. Thus, the measure attaining the supremum on the left-hand side of Eq. (46) has discrete support $\{\hat{u}_i^{(0)}, \ldots, \hat{u}_i^{(N+1)}\}$, and the supremum is equivalent to the linear optimization problem:

$$\begin{aligned}
\max_{\mathbf{p}} \quad & \sum_{j=0}^{N+1} p_j g\left(\hat{u}^{(j)}\right) \\
\text{s.t.} \quad & \mathbf{p} \geq \mathbf{0}, \quad \mathbf{e}^T\mathbf{p} = 1, \\
& \sum_{k=0}^{j} p_k \geq \frac{j}{N} - \Gamma^{KS}, \quad \sum_{k=j}^{N+1} p_k \geq \frac{N-j+1}{N} - \Gamma^{KS}, \quad j = 1, \ldots, N,
\end{aligned} \tag{48}$$

(We have used the fact that $\mathbb{P}_i(\tilde{u}_i < \hat{u}_i^{(j)}) = 1 - \mathbb{P}_i(\tilde{u} \geq \hat{u}_i^{(j)})$.) Its dual is:

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{y},t} \quad & \sum_{j=1}^{N} x_j\left(\Gamma^{KS} - \frac{j}{N}\right) + \sum_{j=1}^{N} y_j\left(\Gamma^{KS} - \frac{N-j+1}{N}\right) + t \\
\text{s.t.} \quad & t - \sum_{k \leq j \leq N} x_j - \sum_{1 \leq j \leq k} y_j \geq g\left(\hat{u}^{(k)}\right), \quad k = 0, \ldots, N+1, \\
& \mathbf{x}, \mathbf{y} \geq \mathbf{0}.
\end{aligned}$$

Observe that the primal solution $\mathbf{q}^R\left(\Gamma^{KS}\right)$ and dual solution $\mathbf{y} = \mathbf{0}$, $t = g(\hat{u}_i^{(N+1)})$ and

$$x_j = \begin{cases} g\left(\hat{u}_i^{(j+1)}\right) - g\left(\hat{u}_i^{(j)}\right) & \text{for } N - j^* \le j \le N, \\ 0 & \text{otherwise,} \end{cases}$$

constitute a primal-dual optimal pair. This proves (46) when $g$ is non-decreasing. The case of $g(u_i)$ non-increasing is similar. □

*Proof of of Theorem 6* Notice by Theorem 15, Eq. (19) is equivalent to the given expression for $\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^I)$. By our schema, it suffices to show then that this expression is truly the support function of $\mathcal{U}_\epsilon^I$. By Lagrangian duality,

$$\delta^*\left(\mathbf{v}|\,\mathcal{U}_\epsilon^I\right) = \inf_{\lambda \ge 0}\left(\begin{array}{c} \lambda \log(1/\epsilon) + \max\limits_{\mathbf{q},\boldsymbol{\theta}} \sum\limits_{i=1}^{d} v_i \sum\limits_{j=0}^{N+1} \hat{u}_i^{(j)} q_j^i - \lambda \sum\limits_{i=1}^{d} D\left(\mathbf{q}^i, \theta_i \mathbf{q}^L + (1-\theta_i)\mathbf{q}^R\right) \\ \text{s.t.} \quad \mathbf{q}^i \in \Delta_{N+2}, 0 \le \theta_i \le 1, \quad i = 1, \ldots, d. \end{array}\right)$$

The inner maximization decouples in the variables indexed by $i$. The $i^{\text{th}}$ subproblem is

$$\max_{\theta_i \in [0,1]} \lambda \left\{ \max_{\mathbf{q}_i \in \Delta_{N+2}} \left\{ \sum_{j=0}^{N+1} \frac{v_i \hat{u}_i^{(j)}}{\lambda} q_{ij} - D\left(\mathbf{q}^i, \theta_i \mathbf{q}^L + (1-\theta_i)\mathbf{q}^R\right) \right\} \right\}.$$

The inner maximization can be solved analytically [17, p. 93], yielding:

$$q_j^i = \frac{p_j^i e^{v_i \hat{u}_i^{(j)}/\lambda}}{\sum_{j=0}^{N+1} p_j^i e^{v_i \hat{u}_i^{(j)}/\lambda}}, \quad p_j^i = \theta_i q_j^L\left(\Gamma^{KS}\right) + (1-\theta_i)\, q_j^R\left(\Gamma^{KS}\right). \tag{49}$$

Substituting in this solution and recombining subproblems yields

$$\lambda \log(1/\epsilon) + \lambda \sum_{i=1}^{d} \log\left( \max_{\theta_i \in [0,1]} \sum_{j=0}^{N+1} \left(\theta_i q_j^L(\Gamma^{KS}) + (1-\theta_i) q_j^R(\Gamma^{KS})\right) e^{v_i \hat{u}_i^{(j)}/\lambda}. \right) \tag{50}$$

The inner optimizations over $\theta_i$ are all linear, and hence achieve an optimal solution at one of the end points, i.e., either $\theta_i = 0$ or $\theta_i = 1$. This yields the given expression for $\delta^*(\mathbf{v}|\,\mathcal{U})$.

Following this proof backwards to identify the optimal $\mathbf{q}^i$, and, thus, $\mathbf{u} \in \mathcal{U}^I$ also proves the validity of the procedure given in Remark 8 □

*Proof of Theorem 7* By inspection, (28) is the worst-case value of (26) over $\mathcal{P}^{FB}$. By Theorem 3, it suffices to show that this expression truly is the support function of $\mathcal{U}_\epsilon^{FB}$. First observe

$$\max_{\mathbf{u}\in\mathcal{U}_\epsilon^{FB}} \mathbf{u}^T\mathbf{v} = \min_{\lambda\geq 0}\left\{\lambda\log(1/\epsilon) + \max_{\mathbf{m}_b\leq\mathbf{y}_1\leq\mathbf{m}_b,\mathbf{y}_2\geq,\mathbf{y}_3\geq\mathbf{0}} \sum_{i=1}^{d} v_i\,(y_{1i}+y_{2i}-y_{3i}) \right.$$
$$\left. -\lambda\sum_{i=1}^{d}\frac{y_{2i}^2}{2\overline{\sigma}_{fi}^2}+\frac{y_{3i}^2}{2\overline{\sigma}_{bi}^2}\right\}$$

by Lagrangian strong duality. The inner maximization decouples by $i$. The $i^{\text{th}}$ subproblem further decouples into three sub-subproblems. The first is $\max_{m_{bi}\leq y_{i1}\leq m_{fi}} v_i y_{1i}$ with optimal solution

$$y_{1i} = \begin{cases} m_{fi} & \text{if } v_i \geq 0, \\ m_{bi} & \text{if } v_i < 0. \end{cases}$$

The second sub-subproblem is $\max_{y_{2i}\geq 0} v_i y_{2i} - \lambda\frac{y_{2i}^2}{2\overline{\sigma}_{fi}^2}$. This is maximizing a concave quadratic function of one variable. Neglecting the non-negativity constraint, the optimum occurs at $y_{2i}^* = \frac{v_i\sigma_{fi}^2}{\lambda}$. If this value is negative, the optimum occurs at $y_{2i}^* = 0$. Consequently,

$$\max_{y_{2i}\geq 0}\; v_i y_{2i} - \lambda\frac{y_{2i}^2}{2\overline{\sigma}_{fi}^2} = \begin{cases} \frac{v_i\sigma_{fi}^2}{2\lambda} & \text{if } v_i \geq 0, \\ 0 & \text{if } v_i < 0. \end{cases}$$

Similarly, we can show that the third subproblem has the following optimum value

$$\max_{y_{3i}\geq 0}\; -v_i y_{3i} - \lambda\frac{y_{3i}^2}{2\overline{\sigma}_{bi}^2} = \begin{cases} \frac{v_i\sigma_{bi}^2}{2\lambda} & \text{if } v_i \leq 0, \\ 0 & \text{if } v_i > 0. \end{cases}$$

Combining the three sub-subproblems yields

$$\delta^*\left(\mathbf{v}|\mathcal{U}_\epsilon^{FB}\right) = \sum_{i:v_i>0} v_i m_{fi} + \sum_{i:v_i<0} v_i m_{bi} + \min_{\lambda\geq 0}\lambda\log(1/\epsilon)$$
$$+\frac{1}{2\lambda}\left(\sum_{i:v_i>0} v_i^2\overline{\sigma}_{fi}^2 + \sum_{i:v_i<0} v_i^2\overline{\sigma}_{bi}^2\right).$$

This optimization can be solved closed-form, yielding

$$\lambda^* = \sqrt{\frac{\sum_{i:v_i>0} v_i^2\overline{\sigma}_{fi}^2 + \sum_{i:v_i\leq 0} v_i^2\overline{\sigma}_{bi}^2}{2\log(1/\epsilon)}}.$$

Simplifying yields the right hand side of (28). Moreover, following the proof backwards to identify the maximizing $\mathbf{u} \in \mathcal{U}_\epsilon^{FB}$ proves the validity of the procedure given in Remark 11. □

**Proof of Theorem 8**

*Proof* Observe,

$$\sup_{\mathbb{P}\in\mathcal{P}^M} \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) \le \sup_{\mathbb{P}\in\mathcal{P}^M} \sum_{i=1}^{d} \mathrm{VaR}_{\epsilon/d}^{\mathbb{P}}\left(v_i\tilde{u}_i\right) = \sum_{i:v_i>0} v_i\hat{u}_i^{(s)} + \sum_{i:v_i\le 0} v_i\hat{u}_i^{(N-s+1)}, \tag{51}$$

where the equality follows from the positive homogeneity of $\mathrm{VaR}_\epsilon^{\mathbb{P}}$, and this last expression is equivalent to (32) because $\hat{u}_i^{(N-s+1)} \le \hat{u}_i^{(s)}$. By Theorem 2, it suffices to show that $\delta^*(\mathbf{v}|\,\mathcal{U}^M)$ is the support function of $\mathcal{U}_\epsilon^M$, and this is immediate. □

**Proof of Theorem 9**

*Proof* We first compute $\sup_{\mathbb{P}\in\mathcal{P}^{LCX}} \mathbb{P}(\mathbf{v}^T\tilde{\mathbf{u}} > t)$ for fixed $\mathbf{v}, t$. In this spirit of Bertsimas et al. [45] and Shapiro [15] this optimization admits the following strong dual:

$$\begin{aligned}
\inf_{\theta, w_\sigma, \lambda(\mathbf{a},b)} \quad & \theta - \left(\frac{1}{N}\sum_{j=1}^{N} \|\hat{\mathbf{u}}_j\|^2 - \Gamma_\sigma\right) w_\sigma + \int_{\mathcal{B}} \Gamma(\mathbf{a},b) d\lambda(\mathbf{a},b) \\
\text{s.t.} \quad & \theta - w_\sigma\|\mathbf{u}\|^2 + \int_{\mathcal{B}} \left(\mathbf{a}^T\mathbf{u} - b\right)^+ d\lambda\left(\mathbf{a},b\right) \ge \mathbb{I}\left(\mathbf{u}^T\mathbf{v} > t\right) \quad \forall \mathbf{u}\in\mathbb{R}^d, \\
& w_\sigma \ge 0, \quad d\lambda\left(\mathbf{a},b\right) \ge 0,
\end{aligned} \tag{52}$$

where $\Gamma(\mathbf{a},b) \equiv \frac{1}{N}\sum_{j=1}^{N}(\mathbf{a}^T\hat{\mathbf{u}}_j - b)^+ + \Gamma_{LCX}$. We claim that $w_\sigma = 0$ in any feasible solution. Indeed, suppose $w_\sigma > 0$ in some feasible solution. Note $(\mathbf{a},b) \in \mathcal{B}$ implies that $(\mathbf{a}^T\mathbf{u} - b)^+ = O(\|\mathbf{u}\|)$ as $\|\mathbf{u}\| \to \infty$. Thus, the left-hand side of Eq. (52) tends to $-\infty$ as $\|\mathbf{u}\| \to \infty$ while the right-hand side is bounded below by zero. This contradicts the feasibility of the solution.

Since $w_\sigma = 0$ in any feasible solution, rewrite the above as

$$\begin{aligned}
\inf_{\theta,\lambda(\mathbf{a},b)} \quad & \theta + \int_{\mathcal{B}} \Gamma\left(\mathbf{a},b\right) d\lambda\left(\mathbf{a},b\right) \\
\text{s.t.} \quad & \theta + \int_{\mathcal{B}} \left(\mathbf{a}^T\mathbf{u} - b\right)^+ d\lambda\left(\mathbf{a},b\right) \ge 0 \quad \forall \mathbf{u}\in\mathbb{R}^d, && \text{(53a)} \\
& \theta + \int_{\mathcal{B}} \left(\mathbf{a}^T\mathbf{u} - b\right)^+ d\lambda\left(\mathbf{a},b\right) \ge 1 \quad \forall \mathbf{u}\in\{\mathbf{u}\in\mathbb{R}^d : \mathbf{u}^T\mathbf{v} > t\}, \\
& d\lambda\left(\mathbf{a},b\right) \ge 0. && \text{(53b)}
\end{aligned}$$

The two infinite constraints can be rewritten using duality. Specifically, Eq. (53a) is

$$\begin{aligned} -\theta \leq \quad & \min_{s(\mathbf{a},b)\geq 0, \tilde{\mathbf{u}}\in\mathbb{R}^d} \quad \int_{\mathcal{B}} s\,(\mathbf{a},b)\,d\lambda\,(\mathbf{a},b) \\ & \text{s.t.} \quad s\,(\mathbf{a},b) \geq \left(\mathbf{a}^T\tilde{\mathbf{u}} - b\right) \quad \forall\,(\mathbf{a},b)\in\mathcal{B}, \end{aligned}$$

which admits the dual:

$$\begin{aligned} -\theta \leq \max_{y_1(\mathbf{a},b)} \quad & -\int_{\mathcal{B}} b\,dy_1\,(\mathbf{a},b) \\ \text{s.t.} \quad & 0 \leq dy_1\,(\mathbf{a},b) \leq d\lambda\,(\mathbf{a},b) \quad \forall\,(\mathbf{a},b)\in\mathcal{B}, \\ & \int_{\mathcal{B}} \mathbf{a}\,dy_1\,(\mathbf{a},b) = 0. \end{aligned}$$

Equation (53b) can be treated similarly using continuity to take the closure of $\{\mathbf{u}\in\mathbb{R}^d : \mathbf{u}^T\mathbf{v} > t\}$. Combining both constraints yields the equivalent representation of (53)

$$\begin{aligned} \inf_{\substack{\theta,\tau,\lambda(\mathbf{a},b),\\ y_1(\mathbf{a},b),y_2(\mathbf{a},b)}} \quad & \theta + \int_{\mathcal{B}} \Gamma\,(\mathbf{a},b)\,d\lambda\,(\mathbf{a},b) \\ \text{s.t.} \quad & \theta - \int_{\mathcal{B}} b\,dy_1\,(\mathbf{a},b) \geq 0, \quad \theta + t\tau - \int_{\mathcal{B}} b\,dy_2\,(\mathbf{a},b) \geq 1, \\ & 0 \leq dy_1\,(\mathbf{a},b) \leq d\lambda\,(\mathbf{a},b) \quad \forall\,(\mathbf{a},b)\in\mathcal{B}, \\ & 0 \leq dy_2\,(\mathbf{a},b) \leq d\lambda\,(\mathbf{a},b) \quad \forall\,(\mathbf{a},b)\in\mathcal{B}, \\ & \int_{\mathcal{B}} \mathbf{a}\,dy_1\,(\mathbf{a},b) = 0, \\ & \tau\mathbf{v} = \int_{\mathcal{B}} \mathbf{a}\,dy_2\,(\mathbf{a},b)\,, \\ & \tau \geq 0. \end{aligned} \tag{54}$$

Now the worst-case Value at Risk can be written as

$$\begin{aligned} \sup_{\mathbb{P}\in\mathcal{P}^{LCX}} \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) = \quad & \inf_{\substack{\theta,\tau,t,\lambda(\mathbf{a},b),\\ y_1(\mathbf{a},b),y_2(\mathbf{a},b)}} \quad t \\ & \text{s.t.} \quad \theta + \int_{\mathcal{B}} \Gamma(\mathbf{a},b)d\lambda(\mathbf{a},b) \leq \epsilon, \\ & \quad (\theta,\tau,\lambda(\mathbf{a},b),y_1(\mathbf{a},b),y_2(\mathbf{a},b),t) \text{ feasible in (54).} \end{aligned}$$

We claim that $\tau > 0$ in an optimal solution. Suppose to the contrary that $\tau = 0$ in a candidate solution to this optimization problem. As $t \to -\infty$, this candidate solution remains feasible, which implies that for all $t$ arbitrarily small $\mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{v} > t) \leq \epsilon$ for all

$\mathbb{P} \in \mathcal{P}^{LCX}$. However, the empirical distribution $\hat{\mathbb{P}} \in \mathcal{P}^{LCX}$, and for this distribution, we can find a finite $t'$ such that $\hat{\mathbb{P}}(\tilde{\mathbf{u}}^T \mathbf{v} > t') > \epsilon$. This is a contradiction.

Since $\tau > 0$, apply the transformation $(\theta/\tau, 1/\tau, \lambda(\mathbf{a}, b)/\tau, \mathbf{y}(\mathbf{a}, b)/\tau) \rightarrow (\theta, \tau, \lambda(\mathbf{a}, b), \mathbf{y}(\mathbf{a}, b))$ yielding

$$
\begin{aligned}
\inf_{\substack{\theta,\tau,t,\lambda(\mathbf{a},b),\\ y_1(\mathbf{a},b),y_2(\mathbf{a},b)}} \quad & t \\
\text{s.t.} \quad & \theta + \int_{\mathcal{B}} \Gamma(\mathbf{a}, b) d\lambda(\mathbf{a}, b) \leq \epsilon\tau \\
& \theta - \int_{\mathcal{B}} b \, dy_1(\mathbf{a}, b) \geq 0, \quad \theta + t - \int_{\mathcal{B}} b \, dy_2(\mathbf{a}, b) \geq \tau, \\
& 0 \leq dy_1(\mathbf{a}, b) \leq d\lambda(\mathbf{a}, b) \quad \forall (\mathbf{a}, b) \in \mathcal{B}, \\
& 0 \leq dy_2(\mathbf{a}, b) \leq d\lambda(\mathbf{a}, b) \quad \forall (\mathbf{a}, b) \in \mathcal{B}, \\
& \int_{\mathcal{B}} \mathbf{a} \, dy_1(\mathbf{a}, b) = 0, \quad \mathbf{v} = \int_{\mathcal{B}} \mathbf{a} \, dy_2(\mathbf{a}, b), \\
& \tau \geq 0.
\end{aligned}
$$

Eliminate the variable $t$, and make the transformation $(\tau\epsilon, \theta - \int_{\mathcal{B}} b dy_1(\mathbf{a}, b)) \rightarrow (\tau, \theta)$ to yield

$$
\begin{aligned}
\sup_{\mathbb{P} \in \mathcal{P}^{LCX}} \text{VaR}_{\epsilon}^{\mathbb{P}}\left(\mathbf{v}^T \tilde{\mathbf{u}}\right) = \min_{\tau,\theta,y_1,y_2,\lambda} \quad & \frac{1}{\epsilon}\tau - \theta - \int_{\mathcal{B}} b dy_1(\mathbf{a}, b) + \int_{\mathcal{B}} b dy_2(\mathbf{a}, b) \\
\text{s.t.} \quad & \theta + \int_{\mathcal{B}} b dy_1(\mathbf{a}, b) + \int_{\mathcal{B}} \Gamma(\mathbf{a}, b) \, d\lambda(\mathbf{a}, b) \leq \tau \\
& 0 \leq dy_1(\mathbf{a}, b) \leq d\lambda(\mathbf{a}, b) \quad \forall (\mathbf{a}, b) \in \mathcal{B}, \\
& 0 \leq dy_2(\mathbf{a}, b) \leq d\lambda(\mathbf{a}, b) \quad \forall (\mathbf{a}, b) \in \mathcal{B}, \\
& \int_{\mathcal{B}} \mathbf{a} \, dy_1(\mathbf{a}, b) = 0, \quad \mathbf{v} = \int_{\mathcal{B}} \mathbf{a} \, dy_2(\mathbf{a}, b), \\
& \theta, \tau \geq 0.
\end{aligned} \tag{55}
$$

Taking the dual of this last optimization problem and simplifying yields $\sup_{\mathbb{P} \in \mathcal{P}^{LCX}} \text{VaR}_{\epsilon}^{\mathbb{P}}\left(\mathbf{v}^T \tilde{\mathbf{u}}\right) = \max_{\mathbf{u} \in \mathcal{U}^{LCX}} \mathbf{v}^T \mathbf{u}$ where

$$
\begin{aligned}
\mathcal{U}_{\epsilon}^{LCX} = \Big\{ & \mathbf{u} \in \mathbb{R}^d : \exists \mathbf{r} \in \mathbb{R}^d, \ 1 \leq z \leq 1/\epsilon, \ \text{s.t.} \\
& \left(\mathbf{a}^T \mathbf{r} - b\,(z-1)\right)^+ + \left(\mathbf{a}^T \mathbf{u} - b\right)^+ \leq \frac{z}{N} \sum_{j=1}^{N} \left(\mathbf{a}^T \hat{\mathbf{u}}_j - b\right)^+ \\
& + \Gamma_{LCX}, \ \forall (\mathbf{a}, b) \in \mathcal{B} \Big\}.
\end{aligned} \tag{56}
$$

We next seek to remove the semi-infinite constraint in the definition of $\mathcal{U}^{LCX}$. Note that by considering the four possible signs of the two left hand side terms,

$$\begin{aligned}&\left(\mathbf{a}^T\mathbf{r} - b\,(z-1)\right)^+ + \left(\mathbf{a}^T\mathbf{u} - b\right)^+ \\ &= \max\left(\mathbf{a}^T\,(\mathbf{r}+\mathbf{u}) - bz, \mathbf{a}^T\mathbf{r} - b\,(z-1)\,, \mathbf{a}^T\mathbf{u} - b, 0\right)\end{aligned}$$

Thus, we can replace the original semi-infinite constraint with the following three semi-infinite constraints

$$\mathbf{a}^T\,(\mathbf{r}+\mathbf{u}) - bz - \frac{z}{N}\sum_{j=1}^{N}\left(\mathbf{a}^T\hat{\mathbf{u}}_j - b\right)^+ \leq \Gamma_{LCX}\ \forall\,(\mathbf{a},b)\in\mathcal{B}$$

$$\mathbf{a}^T\mathbf{r} - b\,(z-1) - \frac{z}{N}\sum_{j=1}^{N}\left(\mathbf{a}^T\hat{\mathbf{u}}_j - b\right)^+ \leq \Gamma_{LCX}\ \forall\,(\mathbf{a},b)\in\mathcal{B}$$

$$\mathbf{a}^T\mathbf{u} - b - \frac{z}{N}\sum_{j=1}^{N}\left(\mathbf{a}^T\hat{\mathbf{u}}_j - b\right)^+ \leq \Gamma_{LCX}\ \forall\,(\mathbf{a},b)\in\mathcal{B},$$

where the last case (corresponding to the fourth assignment of signs) is trivial since $\Gamma_{LCX} + \frac{z}{N}\sum_{j=1}^{N}(\mathbf{a}^T\hat{\mathbf{u}}_j - b)^+ \geq 0$. In contrast to the constraint (56), each of these constraints is concave in $(\mathbf{a}, b)$. We can find the robust counterpart of each constraint using [5] by computing the concave conjugate of the term functions on the left. The resulting representation is given in (34). Finally, to complete the theorem, we use linear programming duality to rewrite $\max_{\mathbf{u}\in\mathcal{U}^{LCX}}\mathbf{v}^T\mathbf{u}$ as a minimization, obtaining the representation of the support function and worst-case VaR. Some rearrangement yields the representation (35). □

### Proofs of Theorems 11 and 13

*Proof of Theorem 11* By Theorem 3, it suffices to show that $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{CS})$ is given by (38), which follows immediately from two applications of the Cauchy-Schwartz inequality. □

To prove Theorem 13 we require the following proposition:

**Proposition 2**

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}^{DY}\left(\gamma_1^B,\gamma_2^B\right)}\mathbb{P}\left(\tilde{\mathbf{u}}^T\mathbf{v} > t\right) = \min_{r,s,\theta,\mathbf{y}_1,\mathbf{y}_2,\mathbf{Z}}\quad & r+s \\ s.t.\quad & \begin{pmatrix} r + \mathbf{y}_1^{+T}\hat{\mathbf{u}}^{(0)} - \mathbf{y}_1^{-T}\hat{\mathbf{u}}^{(N+1)} & \frac{1}{2}\,(\mathbf{q}-\mathbf{y}_1)^T \\ \frac{1}{2}\,(\mathbf{q}-\mathbf{y}_1) & \mathbf{Z}\end{pmatrix} \succeq \mathbf{0}, \\ & \begin{pmatrix} r + \mathbf{y}_2^{+T}\hat{\mathbf{u}}^{(0)} - \mathbf{y}_2^{-T}\hat{\mathbf{u}}^{(N+1)} + \theta t - 1 & \frac{1}{2}\,(\mathbf{q}-\mathbf{y}_2-\theta\mathbf{v})^T \\ \frac{1}{2}\,(\mathbf{q}-\mathbf{y}_2-\theta\mathbf{v}) & \mathbf{Z}\end{pmatrix} \succeq \mathbf{0},\end{aligned}$$

$$s \geq \left(\gamma_2^B \hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T\right) \circ \mathbf{Z} + \hat{\boldsymbol{\mu}}^T \mathbf{q} + \sqrt{\gamma_1^B}\|\mathbf{q} + 2\mathbf{Z}\hat{\boldsymbol{\mu}}\|_{\hat{\boldsymbol{\Sigma}}^{-1}},$$
$$\mathbf{y}_1 = \mathbf{y}_1^+ - \mathbf{y}_1^-, \quad \mathbf{y}_2 = \mathbf{y}_2^+ - \mathbf{y}_2^-, \quad \mathbf{y}_1^+, \mathbf{y}_1^-, \mathbf{y}_2^+, \mathbf{y}_2^- \theta \geq \mathbf{0}. \quad (57)$$

*Proof* We suppress the dependence on $\gamma_1^B, \gamma_2^B$ in the notation. We claim that $\sup_{\mathbb{P}\in\mathcal{P}^{DY}} \mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{v} > t)$ has the following dual representation:

$$\begin{aligned}
\min_{r,s,\mathbf{q},\mathbf{Z},\mathbf{y}_1,\mathbf{y}_2,\theta} \quad & r + s \\
\text{s.t.} \quad & r + \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} \geq 0 \quad \forall \mathbf{u} \in \left[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}\right], \\
& r + \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} \geq 1 \quad \forall \mathbf{u} \in \left[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}\right] \cap \left\{\mathbf{u} : \mathbf{u}^T\mathbf{v} > t\right\}, \\
& s \geq \left(\gamma_2^B \hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T\right) \circ \mathbf{Z} + \hat{\boldsymbol{\mu}}^T\mathbf{q} + \sqrt{\gamma_1^B}\|\mathbf{q} + 2\mathbf{Z}\hat{\boldsymbol{\mu}}\|_{\hat{\boldsymbol{\Sigma}}^{-1}}, \\
& \mathbf{Z} \succeq \mathbf{0}. \qquad (58)
\end{aligned}$$

See the proof of Lemma 1 in [25] for details. Since $\mathbf{Z}$ is positive semidefinite, we can use strong duality to rewrite the two semi-infinite constraints:

$$\begin{aligned}
&\min_{\mathbf{u}} \ \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} \\
&\text{s.t.} \ \ \hat{\mathbf{u}}^{(0)} \leq \mathbf{u} \leq \hat{\mathbf{u}}^{(N+1)},
\end{aligned} \iff
\begin{aligned}
&\max_{\mathbf{y}_1,\mathbf{y}_1^+,\mathbf{y}_1^-} \ -\frac{1}{4}(\mathbf{q} - \mathbf{y}_1)^T\mathbf{Z}^{-1}(\mathbf{q} - \mathbf{y}_1) + \mathbf{y}_1^+\hat{\mathbf{u}}^{(0)} - \mathbf{y}_1^-\hat{\mathbf{u}}^{(N+1)} \\
&\text{s.t.} \ \ \mathbf{y}_1 = \mathbf{y}_1^+ - \mathbf{y}_1^-, \ \ \mathbf{y}_1^+, \mathbf{y}_1^- \geq \mathbf{0},
\end{aligned}$$

$$\begin{aligned}
&\min_{\mathbf{u}} \ \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} \\
&\text{s.t.} \ \ \hat{\mathbf{u}}^{(0)} \leq \mathbf{u} \leq \hat{\mathbf{u}}^{(N+1)}, \\
&\quad \mathbf{u}^T\mathbf{v} \geq t,
\end{aligned} \iff
\begin{aligned}
&\max_{\mathbf{y}_2,\mathbf{y}_2^+,\mathbf{y}_2^-} \ -\frac{1}{4}(\mathbf{q} - \mathbf{y}_2 - \theta\mathbf{v})^T\mathbf{Z}^{-1}(\mathbf{q} - \mathbf{y}_2 - \theta\mathbf{v}) + \mathbf{y}_2^+\hat{\mathbf{u}}^{(0)} - \mathbf{y}_2^-\hat{\mathbf{u}}^{(N+1)} + \theta t \\
&\text{s.t.} \ \ \mathbf{y}_2 = \mathbf{y}_2^+ - \mathbf{y}_2^-, \ \ \mathbf{y}_2^+, \mathbf{y}_2^- \geq \mathbf{0}, \ \theta \geq 0.
\end{aligned}$$

Then, by using Schur-Complements, we can rewrite Problem (58) as in the proposition. □

We can now prove the theorem.

*Proof of Theorem 13* Using Proposition 2, we can characterize the worst-case VaR by

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathcal{P}^{DY}} \mathrm{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right) \\
&= \inf\left\{t : r + s \leq \epsilon, (r, s, t, \theta, \mathbf{y}_1, \mathbf{y}_2, \mathbf{Z}) \text{ are feasible in problem (57)}\right\}. \qquad (59)
\end{aligned}$$

We claim that $\theta > 0$ in any feasible solution to the infimum in Eq. (59). Suppose to the contrary that $\theta = 0$. Then this solution is also feasible as $t \downarrow \infty$, which implies that $\mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{v} > -\infty) \leq \epsilon$ for all $\mathbb{P} \in \mathcal{P}^{DY}$. On the other hand, the empirical distribution $\hat{\mathbb{P}} \in \mathcal{P}^{DY}$, a contradiction.

Since $\theta > 0$, we can rescale all of the above optimization variables in problem (57) by $\theta$. Substituting this into Eq. (59) yields the given expression for

$\sup_{\mathbb{P}\in\mathcal{P}^{DY}} \text{VaR}_\epsilon^{\mathbb{P}}\left(\mathbf{v}^T\tilde{\mathbf{u}}\right)$. Rewriting this optimization problem as a semidefinite optimization problem and taking its dual yields $\mathcal{U}_\epsilon^{DY}$ in the theorem. By Theorem 3, this set simultaneously implies a probabilistic guarantee. □

**Proof of Theorem 16**

*Proof* For each part, the convexity in $(\mathbf{v}, t)$ is immediate since $\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon)$ is a support function of a convex set. For the first part, note that from the second part of Theorem 11, $\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^{CS}) \leq t$ will be convex in $\epsilon$ for a fixed $(\mathbf{v}, t)$ whenever $\sqrt{1/\epsilon - 1}$ is convex. Examining the second derivative of this function, this occurs on the interval $0 < \epsilon < .75$. Similarly, for the second part, note that from the second part of Theorem 7, $\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^{FB}) \leq t$ will be convex in $\epsilon$ for a fixed $(\mathbf{v}, t)$ whenever $\sqrt{2\log(1/\epsilon)}$ is convex. Examining the second derivative of this function, this occurs on the interval $0 < \epsilon < 1\sqrt{e}$.

From the representations of $\delta^*(\mathbf{v}|\mathcal{U}_\epsilon^{\chi^2})$ and $\delta^*(\mathbf{v}|\mathcal{U}_\epsilon^{G})$ in Theorem 5, we can see they will be convex in $\epsilon$ whenever $1/\epsilon$ is convex, i.e., $0 < \epsilon < 1$. From the representation of $\delta^*(\mathbf{v}|\mathcal{U}_\epsilon^{I})$ in Theorem 6 and since $\lambda \geq 0$, we see this function will be convex in $\epsilon$ whenever $\log(1/\epsilon)$ is convex, i.e., $0 < \epsilon < 1$.

Finally, examining the support functions of $\mathcal{U}_\epsilon^{LCX}$ and $\mathcal{U}_\epsilon^{DY}$ shows that $\epsilon$ occurs linearly in each of these functions. □

## Appendix 2: Omitted Figures

This section contains additional figures omitted from the main text.

### Additional Bootstrapping Results

Figure 5 illustrates the set $\mathcal{U}_\epsilon^{CS}$ for the example considered in Fig. 2 with thresholds computed with and without the bootstrap. Notice that for $N = 1000$, the non-bootstrapped set is almost as big as the full support and shrinks slowly to its infinite limit. Furthermore, the bootstrapped set with $N = 100$ points is smaller than the non-bootstrapped version with 50 times as many points.

### Additional Portfolio Results

Figure 6 summarizes the case $N = 2000$ for the experiment outlined in Sect. 9.3.

## Appendix 3: Optimizing $\epsilon_j$'s for Multiple Constraints

In this section, we propose an approach for solving (10). We say that a constraint $f(\mathbf{x}, \mathbf{y}) \leq 0$ is bi-convex in $\mathbf{x}$ and $\mathbf{y}$ if for every $\mathbf{y}$, the set $\{\mathbf{x} : f(\mathbf{x}, \mathbf{y}) \leq 0\}$ is convex and for every $\mathbf{x}$, the set $\{\mathbf{y} : f(\mathbf{x}, \mathbf{y}) \leq 0\}$ is convex. The key observation is then

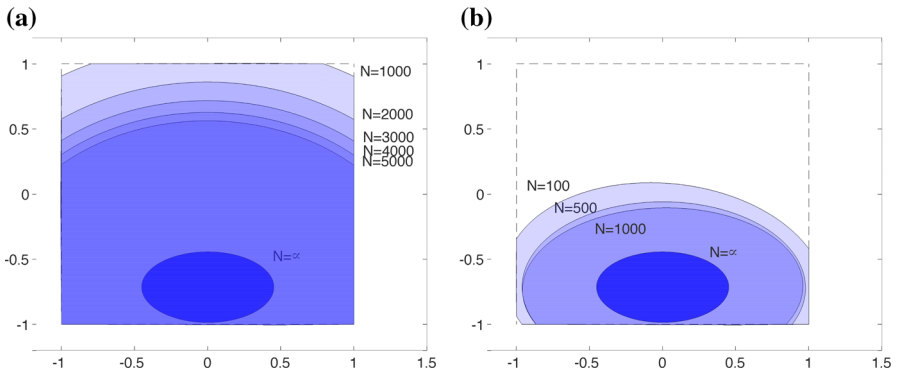


**Fig. 5** $\mathcal{U}_\epsilon^{CS}$ with and without bootstrapping for the example from Fig. 2. $N_B = 10{,}000$, $\alpha = 10\%$, $\epsilon = 10\%$

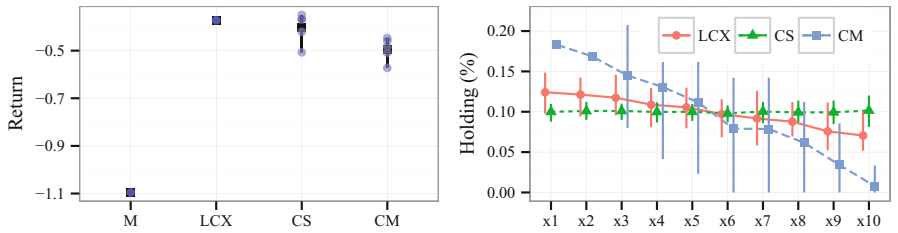


**Fig. 6** The case $N = 2000$ for the experiment outlined in Sect. 9.3. The *left panel* shows the cross-validation results. The *right panel* shows the average holdings by method. $\alpha = \epsilon = 10\%$

**Theorem 16** *a) The constraint* $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{CS}) \leq t$ *is bi-convex in* $(\mathbf{v}, t)$ *and* $\epsilon$*, for* $0 < \epsilon < .75$.

*b) The constraint* $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{FB}) \leq t$ *is bi-convex in* $(\mathbf{v}, t)$ *and* $\epsilon$*, for* $0 < \epsilon < 1/\sqrt{e}$.

*c) The constraint* $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon) \leq t$ *is bi-convex in* $(\mathbf{v}, t)$ *and* $\epsilon$*, for* $0 < \epsilon < 1$*, and* $\mathcal{U}_\epsilon \in \{\mathcal{U}_\epsilon^{\chi^2}, \mathcal{U}_\epsilon^{G}, \mathcal{U}_\epsilon^{I}, \mathcal{U}_\epsilon^{LCX}, \mathcal{U}_\epsilon^{DY}\}$.

This observations suggests a heuristic: Fix the values of $\epsilon_j$, and solve the robust optimization problem in the original decision variables. Then fix this solution and optimize over the $\epsilon_j$. Repeat until some stopping criteria is met or no further improvement occurs. Chen et al. [22] suggested a similar heuristic for multiple chance-constraints in a different context. We propose a refinement of this approach that solves a linear optimization problem to obtain the next iterates for $\epsilon_j$, incorporating dual information from the overall optimization and other constraints. Our proposal ensures the optimization value is non-increasing between iterations and that the procedure is finitely convergent.

For simplicity, we present our approach using $m$ constraints of the form $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{CS}) \leq t$. Without loss of generality, assume the overall optimization problem is a minimization. Consider the $j$th constraint, and let $(\mathbf{v}', t')$ denote the subset of the solution to the original optimization problem at the current iterate pertaining to the $j$th constraint. Let $\epsilon_j'$, $j = 1, \ldots, m$ denote the current iterate in $\epsilon$. Finally, let $\lambda_j$ denote the shadow price of the $j$th constraint in the overall optimization problem.

Notice from the second part of Theorem 11 that $\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^{CS})$ is decreasing in $\epsilon$. Thus, for all $\epsilon_j \geq \underline{\epsilon}_j$, $\delta^*(\mathbf{v}'|\,\mathcal{U}_{\epsilon_j}^{CS}) \leq t'$, where,

$$\underline{\epsilon}_j \equiv \left[ \frac{\left(t' - \hat{\boldsymbol{\mu}}^T \mathbf{v}' - \Gamma_1 \|\mathbf{v}'\|_2\right)^2}{\mathbf{v}'^T \left(\boldsymbol{\Sigma} + \Gamma_2 \mathbf{I}\right) \mathbf{v}'} + 1 \right]^{-1}.$$

Motivated by the shadow-price $\lambda_j$, we define the next iterates of $\epsilon_j$, $j = 1, \ldots, m$ to be the solution of the linear optimization problem

$$\begin{aligned} \min_{\boldsymbol{\epsilon}} \quad & -\sum_{j=1}^{d} \left( \frac{\sqrt{\mathbf{v}'^T (\boldsymbol{\Sigma} + \Gamma_2 \mathbf{I}) \mathbf{v}'}}{2\epsilon'^2 \sqrt{\frac{1}{\epsilon'} - 1}} \right) \lambda_j \cdot \epsilon_j \\ \text{s.t.} \quad & \underline{\epsilon}_j \leq \epsilon_j \leq .75, \quad j = 1, \ldots, m, \\ & \sum_{j=1}^{m} \epsilon_j \leq \overline{\epsilon}, \quad \|\boldsymbol{\epsilon}' - \boldsymbol{\epsilon}\|_1 \leq \kappa. \end{aligned} \tag{60}$$

The coefficient of $\epsilon_j$ in the objective function is $\lambda_j \cdot \partial_{\epsilon_j} \delta^*(\mathbf{v}'|\,\mathcal{U}_{\epsilon_j}^{CS})$ which is intuitively a first-order approximation to the improvement in the overall optimization problem for a small change in $\epsilon_j$. The norm constraint on $\boldsymbol{\epsilon}$ ensures that the next iterate is not too far away from the current iterate, so that the shadow-price $\lambda_j$ remains a good approximation. (We use $\kappa = .05$ in our experiments.) The upper bound ensures that we remain in a region where $\delta^*(\mathbf{v}|\,\mathcal{U}_{\epsilon_j}^{CS})$ is convex in $\epsilon_j$. Finally, the lower bounds on $\epsilon_j$ ensure that the previous iterate of the original optimization problem $(\mathbf{v}', t')$ will still be feasible for the new values of $\epsilon_j$. Consequently, the objective value of the original optimization problem is non-increasing. We terminate the procedure when the objective value no longer makes significant progress.

We can follow an entirely analogous procedure for each of our other sets, simply adjusting the formulas for $\underline{\epsilon}_j$, the upper bounds, and the objective coefficient appropriately. We omit the details.

## Appendix 4: Queueing Analysis

One of the strengths of our approach is the ability to retrofit existing robust optimization models by replacing their uncertainty sets with our proposed sets, thereby creating new data-driven models that satisfy strong guarantees. In this section, we illustrate this idea with a robust queueing model as in [4,14]. Bandi and Bertsimas [4] use robust optimization to generate *approximations* to a performance metric of a queueing network. We will combine their method with our new sets to generate *probabilistic upper bounds* to these metrics. For concreteness, we focus on the waiting time in a G/G/1 queue. Extending our analysis to more complex queueing networks can likely be accomplished similarly. We stress that we do not claim that our new bounds are the best possible – indeed there exist extremely accurate, specialized techniques for

the G/G/1 queue – but, rather, that the retrofitting procedure is general purpose and yields reasonably good results. These features suggest that a host of other robust optimization applications in information theory [3], supply-chain management [7] and revenue management [44] might benefit from this retrofitting.

Let $\tilde{\mathbf{u}}_i = (\tilde{x}_i, \tilde{t}_i)$ for $i = 1, \ldots, n$ denote the uncertain service times and interarrival times of the first $n$ customers in a queue. We assume that $\tilde{\mathbf{u}}_i$ is i.i.d. for all $i$ and has independent components, and that there exists $\hat{\mathbf{u}}^{(N+1)} \equiv (\overline{x}, \overline{t})$ such that $0 \leq \tilde{x}_i \leq \overline{x}$ and $0 \leq \tilde{t}_i \leq \overline{t}$ almost surely.

From Lindley's recursion [36], the waiting time of the $n$th customer is

$$\tilde{W}_n = \max_{1 \leq j \leq n} \left( \max \left( \sum_{l=j}^{n-1} \tilde{x}_l - \sum_{l=j+1}^{n} \tilde{t}_l, 0 \right) \right) = \max \left( 0, \max_{1 \leq j \leq n} \left( \sum_{l=j}^{n-1} \tilde{x}_l - \sum_{l=j+1}^{n} \tilde{t}_l \right) \right). \tag{61}$$

Motivated by Bandi and Bertsimas [4], we consider a worst-case realization of a Lindley recursion

$$\max \left( 0, \max_{1 \leq j \leq n} \max_{(\mathbf{x},\mathbf{t}) \in \mathcal{U}} \left( \sum_{l=j}^{n-1} \tilde{x}_l - \sum_{l=j+1}^{n} \tilde{t}_l \right) \right). \tag{62}$$

Taking $\mathcal{U} = \mathcal{U}_{\overline{\epsilon}/n}^{FB}$ and applying Theorem 7 to the inner-most optimization yields

$$\max_{1 \leq j \leq n} \left( m_{f1} - m_{b2} \right) (n - j) + \sqrt{2 \log \left( n/\overline{\epsilon} \right) \left( \sigma_{f1}^2 + \sigma_{b2}^2 \right)} \sqrt{n - j} \tag{63}$$

Relaxing the integrality on $j$, this optimization can be solved closed-form yielding

$$W_n^{1,FB} \equiv \begin{cases} \left( m_{f1} - m_{b2} \right) n + \sqrt{2 \log \left( \frac{n}{\overline{\epsilon}} \right) \left( \sigma_{f1}^2 + \sigma_{b2}^2 \right)} \sqrt{n} & \text{if } n < \frac{\log\left(\frac{n}{\overline{\epsilon}}\right)\left(\sigma_{f1}^2 + \sigma_{b2}^2\right)}{2\left(m_{b2} - m_{f1}\right)^2} \text{ or } m_{f1} > m_{b2}, \\ \frac{\log\left(\frac{n}{\overline{\epsilon}}\right)\left(\sigma_{f1}^2 + \sigma_{b2}^2\right)}{2\left(m_{b2} - m_{f1}\right)} & \text{otherwise.} \end{cases} \tag{64}$$

From (62), with probability at least $1 - \alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$, each of the inner-most optimizations upper bound their corresponding random quantity with probability $1 - \overline{\epsilon}/n$ with respect to $\mathbb{P}^*$. Thus, by union bound, $\mathbb{P}^*(\tilde{W}_n \leq W_n^{1,FB}) \geq 1 - \overline{\epsilon}$.

On the other hand, since $\{\mathcal{U}_{\epsilon}^{FB} : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee, we can also optimize the choice of $\epsilon_j$ in (63), yielding

$$
\begin{aligned}
W_n^{2,FB} \equiv \min_{w,\boldsymbol{\epsilon}} \quad & w \\
\text{s.t.} \quad & w \geq \left(m_{f1} - m_{b2}\right)(n-j) \\
& \qquad + \sqrt{2\log\left(1/\epsilon_j\right)\left(\sigma_{f1}^2 + \sigma_{b2}^2\right)}\sqrt{n-j}, \quad j = 1, \ldots, n-1, \\
& w \geq 0, \quad \boldsymbol{\epsilon} \geq \mathbf{0}, \quad \sum_{j=1}^{n-1} \epsilon_j \leq \overline{\epsilon}.
\end{aligned} \tag{65}
$$

From the KKT conditions, the constraint (65) will be tight for all $j$, so that $W_n^{2,FB}$ satisfies

$$
\sum_{j=1}^{n-1} \exp\left(-\frac{\left(W_n^{2,FB} - \left(m_{f1} - m_{b2}\right)\right)^2}{2(n-j)\left(\sigma_{f1}^2 + \sigma_{b2}^2\right)^2}\right) = \overline{\epsilon}, \tag{66}
$$

which can be solved by line search. Again, with probability $1-\alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$, $\mathbb{P}^*(\tilde{W}_n \leq W_n^{2,FB}) \geq 1-\overline{\epsilon}$, and $W_n^{2,FB} \leq W_n^{1,FB}$ by construction.

We can further refine our bound by truncating the recursion (61) at customer $\min(n, n^{(k)})$ where, with high probability, $\tilde{n} \leq n^{(k)}$. We next provide a formal derivation of this bound, which we denote $W_n^{3,FB}$. Notice that in (61), the optimizing index $j$ represents the most recent customer to arrive when the queue was empty. Let $\tilde{n}$ denote the number of customers served in a typical busy period. Intuitively, it suffices to truncate the recursion (61) at customer $\min(n, n^{(k)})$ where, with high probability, $\tilde{n} \leq n^{(k)}$. More formally, considering only the first half of the data $\hat{x}^1, \ldots, \hat{x}^{\lceil N/2 \rceil}$ and $\hat{t}^1, \ldots, \hat{t}^{\lceil N/2 \rceil}$, we compute the number of customers served in each busy period of the queue, denoted $\hat{n}^1, \ldots, \hat{n}^K$, which are i.i.d. realizations of $\tilde{n}$. Using the KS test at level $\alpha_1$, we observe that with probability at least $1-\alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$,

$$
\mathbb{P}\left(\tilde{n} > \hat{n}^{(k)}\right) \leq 1 - \frac{k}{K} + \Gamma^{KS}(\alpha), \quad \forall k = 1, \ldots, K. \tag{67}
$$

In other words, the queue empties every $\hat{n}^{(k)}$ customers with at least this probability.

Next, calculate the constants $\mathbf{m}_f, \mathbf{m}_b, \boldsymbol{\sigma}_f, \boldsymbol{\sigma}_b$ using only the second half of the data. Then, truncate the sum in (66) at $\min(n, n^{(k)})$ and replace the righthand side by $\overline{\epsilon} - 1 + \frac{k}{K} - \Gamma^{KS}(\alpha/2)$. Denote the solution of this equation by $W_n^{2,FB}(k)$. Finally, let $W_n^{3,FB} \equiv \min_{1 \leq k < K} W_n^{2,FB}(k)$, obtained by grid-search.

We claim that with probability at least $1-2\alpha$ with respect to $\mathbb{P}_{\mathcal{S}}$, $\mathbb{P}(\tilde{W}_n > W_n^{3,FB}) \leq \overline{\epsilon}$. Namely, from our choice of parameters, Eqs. (66) and (67) hold simultaneously with probability at least $1-2\alpha$. Restrict attention to a sample path where these equations hold. Since (67) holds for the optimal index $k^*$, recursion (61) truncated at $n^{(k^*)}$ is valid with probability at least $1 - \frac{k^*}{K} + \Gamma^{KS}(\alpha)$. Finally, $\mathbb{P}(\tilde{W}_n > W_n^{3,FB}) \leq \mathbb{P}((61) is invalid) + \mathbb{P}((\tilde{W}_n > W_n^{2,FB}(k^*)$ and (61) is valid) $\leq \overline{\epsilon}$. This proves the claim.

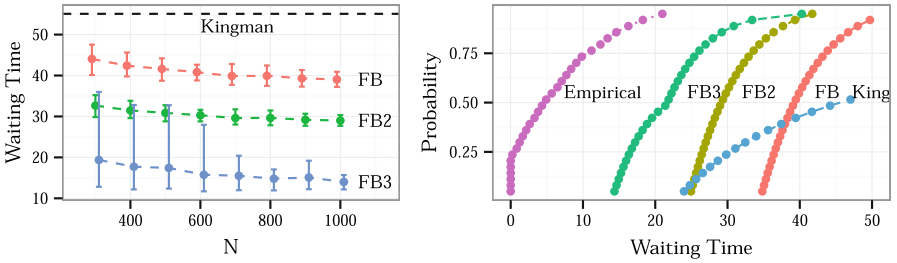


**Fig. 7** The *left panel* shows various bounds on the median waiting time ($\epsilon = .5$) for $n = 10$ and various values of $N$. The *right panel* bounds the entire cumulative distribution of the waiting time for $n = 10$ and $N = 1000$. using $W_n^{FB,3}$. In both cases, $\alpha = 20\%$

We observe in passing that since the constants $\mathbf{m}_f, \mathbf{m}_b, \boldsymbol{\sigma}_f, \boldsymbol{\sigma}_b$ are computed using only half the data, it may not be the case that $W_n^{3,FB} < W_n^{2,FB}$, particularly for small $N$, but that typically $W_n^{3,FB}$ is a much stronger bound than $W_n^{2,FB}$ (see also Fig. 7).

Finally, our choice of $\mathcal{U}_\epsilon^{FB}$ was somewhat arbitrary. Similar analysis can be performed for many of our sets. To illustrate, we next derive corresponding bounds for the set $\mathcal{U}^{CS}$. Following essentially the same argument yields:

$$W_n^{1,CS} \leq \begin{cases} \left(\hat{\mu}_1 - \hat{\mu}_2\right) n + \left(\Gamma_1 + \sqrt{\left(\frac{n}{\epsilon} - 1\right)\left(\sigma_1^2 + \sigma_2^2 + 2\Gamma_2\right)}\right)\sqrt{n} & \text{if } n < \frac{\left(\Gamma_1 + \sqrt{\left(\frac{n}{\epsilon}-1\right)\left(\sigma_1^2+\sigma_2^2+2\Gamma_2\right)}\right)^2}{4\left(\hat{\mu}_1 - \hat{\mu}_2\right)^2} \\ & \text{or } \hat{\mu}_1 > \hat{\mu}_2, \\ \frac{\left(\Gamma_1 + \sqrt{\left(\frac{n}{\epsilon}-1\right)\left(\sigma_1^2+\sigma_2^2+2\Gamma_2\right)}\right)^2}{4\left(\hat{\mu}_2 - \hat{\mu}_1\right)} & \text{otherwise} \end{cases}$$

$W_n^{2,CS}$ is the solution to

$$\sum_{j=1}^{n-1}\left[\left(\frac{W_n^{2,CS} - \left(\hat{\mu}_1 - \hat{\mu}_2\right)(n-j)}{\sqrt{n-j}\sqrt{\sigma_1^2 + \sigma_2^2 + 2\Gamma_2}} - \frac{\Gamma_1}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\Gamma_2}}\right)^2 + 1\right]^{-1} = \overline{\epsilon}, \quad (68)$$

and $W_n^{3,CS}$ defined analogously to $W_n^{3,FB}$ but using (68) in lieu of (66).

We illustrate these ideas numerically. Let service times follow a Pareto distribution with parameter 1.1 truncated at 15, and the interarrival times follow an exponential distribution with rate 3.05 truncated at 15.25. The resulting truncated distributions have means of approximately 3.029 and 3.372, respectively, yielding an approximate 90% utilization.

As a first experiment, we bound the median waiting time ($\epsilon = 50\%$) for the $n = 10$ customer, using each of our bounds with differing amounts of data. We repeat this procedure 100 times to study the variability of our bounds with respect to the data. The left panel of Fig. 7 shows the average value of the bound and error bars for the 10% and 90% quantiles. As can be seen, all of the bounds improve as we add more data. Moreover, optimizing the $\epsilon_j$'s (the difference between $W_n^{FB,1}$ and $W_n^{FB,2}$) is significant.

**Table 4** Summary statistics for various bounds on median waiting time

$N = 10{,}000$, $n = 10$, $\alpha = 10\%$. The last two columns refer to upper and lower quantiles over the simulation

| | Mean | SD | 10% | 90% |
|---|---|---|---|---|
| $W_n^{FB,1}$ | 34.6 | 0.4 | 34.0 | 35.2 |
| $W_n^{FB,2}$ | 25.8 | 0.3 | 25.4 | 26.2 |
| $W_n^{FB,3}$ | 14.4 | 1.2 | 13.5 | 15.5 |
| $W^{King}$ | 55.1 | 8.7 | 46.0 | 67.4 |

For comparison purposes, we include a sample analogue of Kingman's bound [33] on the $1-\epsilon$ quantile of the waiting time, namely,

$$W^{King} \equiv \frac{\hat{\mu}_x \left(\hat{\sigma}_a^2 \hat{\mu}_x^2 + \hat{\sigma}_x^2 \hat{\mu}_t^2\right)}{2\overline{\epsilon} \hat{\mu}_t^2 \left(\hat{\mu}_t - \hat{\mu}_x\right)},$$

where $\hat{\mu}_t, \hat{\sigma}_t^2$ are the sample mean and sample variance of the arrivals, $\hat{\mu}_x, \hat{\sigma}_x^2$ are the sample mean and sample variance of the service times, and we have applied Markov's inequality. Unfortunately, this bound is extremely unstable, even for large $N$. The dotted line in the left-panel of Fig. 7 is the average value over the 100 runs of this bound for $N = 10{,}000$ data points (the error-bars do not fit on graph.) Sample statistics for this bound and our bounds can also be seen in Table 4. As shown, our bounds are both significantly better (with less data), and exhibit less variability.

As a second experiment, we use our bounds to calculate a probabilistic upper bound on the entire CDF of $\tilde{W}_n$ for $n = 10$ with $N = 1000$, $\alpha = 20\%$. Results can be seen in the right panel of Fig. 7. We have included the empirical CDF of the waiting time and the sampled version of the Kingman bound comparison. As seen, our bounds significantly improve upon the sampled Kingman bound, and the benefit of optimizing the $\epsilon_j$'s is again, significant. We remark that the ability to simultaneously bound the entire CDF for any $n$, whether transient or steady-state, is an important strength of this type of analysis.

## Appendix 5: Constructing $\mathcal{U}_\epsilon^I$ from Other EDF Tests

In this section we show how to extend our constructions for $\mathcal{U}_\epsilon^I$ to other EDF tests. We consider several of the most popular, univariate goodness-of-fit, empirical distribution function test. Each test below considers the null-hypothesis $H_0 : \mathbb{P}_i^* = \mathbb{P}_{0,i}$.

Kuiper (K) Test: The K test rejects the null hypothesis at level $\alpha$ if

$$\max_{j=1,\ldots,N} \left( \frac{j}{N} - \mathbb{P}_{0,i}\left(\tilde{u}_i \leq \hat{u}_i^{(j)}\right)\right) + \max_{j=1,\ldots,N} \left( \mathbb{P}_{0,i}\left(\tilde{u}_i < \hat{u}_i^{(j)}\right) - \frac{j-1}{N}\right) > V_{1-\alpha}.$$

Cramer von-Mises (CvM) Test: The CvM test rejects the null hypothesis at level $\alpha$ if

$$\frac{1}{12N^2} + \frac{1}{N}\sum_{j=1}^{N}\left(\frac{2j-1}{2N} - \mathbb{P}_{0,i}\left(\tilde{u}_i \le \hat{u}_i^{(j)}\right)\right)^2 > (T_{1-\alpha})^2\,.$$

Watson (W) Test: The W test rejects the null hypothesis at level $\alpha$ if

$$\frac{1}{12N^2} + \frac{1}{N}\sum_{j=1}^{N}\left(\frac{2j-1}{2N} - \mathbb{P}_{0,i}\left(\tilde{u}_i \le \hat{u}_i^{(j)}\right)\right)^2 - \left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}_{0,i}\left(\tilde{u}_i \le \hat{u}_i^{(j)}\right) - \frac{1}{2}\right)^2 > (U_{1-\alpha})^2.$$

Anderson-Darling (AD) Test: The AD test rejects the null hypothesis at level $\alpha$ if

$$-1 - \sum_{j=1}^{N}\frac{2j-1}{N^2}\left(\log\left(\mathbb{P}_{0,i}(\tilde{u}_i \le \hat{u}_i^{(j)})\right) + \log\left(1 - \mathbb{P}_{0,i}\left(\tilde{u}_i \le \hat{u}_i^{(N+1-j)}\right)\right)\right) > (A_{1-\alpha})^2$$

Tables of the thresholds above are readily available (e.g., [47, and references therein]).

As described in [15], the confidence regions of these tests can be expressed in the form

$$\mathcal{P}_i^{EDF} = \left\{\mathbb{P}_i \in \theta\left[\hat{u}_i^{(0)}, \hat{u}_i^{(N+1)}\right] : \exists \zeta \in \mathbb{R}^N,\ \mathbb{P}_i\left(\tilde{u}_i \le \hat{u}_i^{(j)}\right) = \zeta_i,\ \mathbf{A}_S\boldsymbol{\zeta} - \mathbf{b}_S \in \mathcal{K}_S\right\},$$

where the the matrix $\mathbf{A}_S$, vector $\mathbf{b}_S$ and cone $\mathcal{K}_S$ depend on the choice of test. Namely,

$$\begin{aligned}
\mathcal{K}_K &= \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2N} : \min_i x_i + \min_i y_i \ge 0\},\\
\mathbf{b}_K &= \begin{pmatrix} \frac{1}{N} - V_{1-\alpha}/2 \\ \vdots \\ \frac{N}{N} - V_{1-\alpha}/2 \\ -\frac{0}{N} - V_{1-\alpha}/2 \\ \vdots \\ -\frac{N-1}{N} - V_{1-\alpha}/2 \end{pmatrix}, \quad \mathbf{A}_K = \begin{pmatrix} [\mathbf{I}_N] \\ [-\mathbf{I}_N] \end{pmatrix},\\
\mathcal{K}_{CvM} &= \{\mathbf{x} \in \mathbb{R}^N, t \in \mathbb{R}_+ : \|\mathbf{x}\| \le t\},\\
\mathbf{b}_{CvM} &= \begin{pmatrix} \sqrt{N(T_{1-\alpha}^2)^2 - \frac{1}{2N}} \\ \frac{1}{2N} \\ \frac{3}{2N} \\ \vdots \\ \frac{2N-1}{2N} \end{pmatrix}, \quad \mathbf{A}_{CvM} = \begin{pmatrix} 0\cdots 0 \\ [\mathbf{I}_N] \end{pmatrix},
\end{aligned} \tag{69}$$

$$\mathcal{K}_W = \{\mathbf{x} \in \mathbb{R}^{N+1}, t \in \mathbb{R}_+ : \|\mathbf{x}\| \le t\}, \quad \mathbf{b}_W = \begin{pmatrix} -\frac{1}{2} + \left(\frac{N}{24} - \frac{N}{2}(U_{1-\alpha})^2\right) \\ -\frac{1}{2} - \left(\frac{N}{24} - \frac{N}{2}(U_{1-\alpha})^2\right) \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$\mathbf{A}_{U_N} = \begin{pmatrix} \frac{1-N}{2N} & \frac{3-N}{2N} \cdots \frac{N-1}{2N} \\ \frac{N-1}{2N} & \frac{N-3}{2N} \cdots \frac{1-N}{2N} \\ \left[\mathbf{I}_N - \frac{1}{N}\mathbf{E}_N\right] & \end{pmatrix}, \tag{70}$$

$$\mathcal{K}_{AD} = \left\{(z, \mathbf{x}, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}_+^{2N} : |z| \le \prod_{i=1}^{N} (x_i y_i)^{\frac{2i-1}{2N^2}}\right\}, \quad \mathbf{b}_{AD} = \begin{pmatrix} e^{-(A_{1-\alpha})^2 - 1} \\ 0 \\ \vdots \\ 0 \\ -1 \\ \vdots \\ -1 \end{pmatrix},$$

$$\mathbf{A}_{AD} = \begin{pmatrix} 0 \cdots 0 \\ [I_N] \\ [-\tilde{I}_N] \end{pmatrix}, \tag{71}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix, $\tilde{\mathbf{I}}_N$ is the skew identity matrix ($[\tilde{\mathbf{I}}_N]_{ij} = \mathbb{I}[i = N - j]$), and $\mathbf{E}_N$ is the $N \times N$ matrix of all ones.

Let $\mathcal{K}^*$ denote the dual cone to $\mathcal{K}$. By specializing Theorem 10 of Bertsimas et al. [15], we obtain the following theorem, paralleling Theorem 15.

**Theorem 17** *Suppose $g(u)$ is monotonic and right-continuous, and let $\mathcal{P}^S$ denote the confidence region of any of the above EDF tests.*

$$\begin{aligned} \sup_{\mathbb{P}_i \in \mathcal{P}_i^{EDF}} \mathbb{E}^{\mathbb{P}_i}[g(\tilde{u}_i)] = \min_{\mathbf{r},\mathbf{c}} \quad & \mathbf{b}_S^T \mathbf{r} + c_{N+1} \\ s.t. \quad & -\mathbf{r} \in \mathcal{K}_S^*, \ \ \mathbf{c} \in \mathbb{R}^{N+1}, \\ & \left(\mathbf{A}_S^T \mathbf{r}\right)_j = c_j - c_{j+1} \quad \forall j = 1, \ldots, N, \\ & c_j \ge g\left(\hat{u}_i^{(j-1)}\right), \ \ c_j \ge g\left(\hat{u}_i^{(j)}\right), \quad j = 1, \ldots, N+1. \end{aligned} \tag{72}$$

$$\begin{aligned} = \max_{\mathbf{z},\mathbf{q}^L,\mathbf{q}^R,\mathbf{p}} \quad & \sum_{j=0}^{N+1} p_j g\left(\hat{u}_i^{(j)}\right) \\ s.t. \quad & \mathbf{A}_S \mathbf{z} - \mathbf{b}_S \in \mathcal{K}_S, \ \ \mathbf{q}^L, \mathbf{q}^R, \mathbf{p} \in \mathbb{R}_+^{N+1} \\ & q_j^L + q_j^R = z_j - z_{j-1}, \ \ j = 1, \ldots, N, \end{aligned}$$

$$
\begin{aligned}
& q_{N+1}^L + q_{N+1}^R = 1 - z_N \\
& p_0 = q_1^L, \quad p_{N+1} = q_{N+1}^R, \quad p_j = q_{j+1}^L + q_j^R, \quad j = 1, \ldots, N,
\end{aligned} \tag{73}
$$

*where* $\mathbf{A}_S, \mathbf{b}_S, \mathcal{K}_S$ *are the appropriate matrix, vector and cone to the test. Moreover, when* $g(u)$ *is non-decreasing (resp. non-increasing), there exists an optimal solution where* $\mathbf{q}^L = \mathbf{0}$ *(resp.* $\mathbf{q}^R = \mathbf{0}$*) in* (73).

*Proof* Apply Theorem 10 of Bertsimas et al. [15] and observe that since $g(u)$ is monotonic and right continuous,

$$
c_j \geq \sup_{u \in \left(\hat{u}_i^{(j-1)},\, \hat{u}_i^{(j)}\right]} g(u) \iff c_j \geq g\left(\hat{u}_i^{(j-1)}\right), \; c_j \geq g\left(\hat{u}_i^{(j)}\right).
$$

Take the dual of this (finite) conic optimization problem to obtain the given maximization formulation.

To prove the last statement, suppose first that $g(u)$ is non-decreasing and fix some $j$. If $g(\hat{u}_i^{(j)}) > g(\hat{u}_i^{(j-1)})$, then by complementary slackness, $\mathbf{q}^L = 0$. If $g(\hat{u}_i^{(j)}) = g(\hat{u}_i^{(j-1)})$, then given any feasible $(q_j^L, q_j^R)$, the pair $(0, q_j^L + q_j^R)$ is also feasible with the same objective value. Thus, without loss of generality, $\mathbf{q}^L = 0$. The case where $g(u)$ is non-increasing is similar. □

*Remark 22* At optimality of (73), $\mathbf{p}$ can be considered a probability distribution, supported on the points $\hat{u}_i^{(j)}$ $j = 0, \ldots, N+1$. This distribution is analogous to $\mathbf{q}^L(\Gamma), \mathbf{q}^R(\Gamma)$ for the KS test.

In the special case of the $K$ test, we can solve (73) explicitly to find this worst-case distribution.

**Corollary 1** *When* $\mathcal{P}_i^{EDF}$ *refers specifically to the K test in Theorem 17 and if g is monotonic, we have*

$$
\sup_{\mathbb{P}_i \in \mathcal{P}_i^{EDF}} \mathbb{E}^{\mathbb{P}_i}\left[g(\tilde{u}_i)\right] = \max\left(\sum_{j=0}^{N+1} q_j^L\left(\Gamma^K\right) g\left(\hat{u}_i^{(j)}\right), \sum_{j=0}^{N+1} q_j^R\left(\Gamma^K\right) g\left(\hat{u}_i^{(j)}\right)\right). \tag{74}
$$

*Proof* One can check that in the case of the $K$ test, the maximization formulation given is equivalent to (48) with $\Gamma^{KS}$ replaced by $\Gamma^K$. Following the proof of Theorem 15 yields the result. □

*Remark 23* One an prove that $\Gamma^K \geq \Gamma^{KS}$ for all $N, \alpha$. Consequently, $\mathcal{P}_i^{KS} \subseteq \mathcal{P}_i^K$. For practical purposes, one should thus prefer the KS test to the K test, as it will yield smaller sets.

We can now generalize Theorem 6. For each of K, CvM, W and AD tests, define the (finite dimensional) set

$$\mathcal{P}_i^{EDF} = \{\mathbf{p} \in \mathbb{R}_+^{N+2} : \exists \mathbf{q}^L, \mathbf{q}^R \in \mathbb{R}_+^{N+2}, \mathbf{z} \in \mathbb{R}^N \text{ s.t. } \mathbf{p}, \mathbf{q}^L, \mathbf{q}^R, \mathbf{z} \text{ are feasible in (73)}\}, \tag{75}$$

using the appropriate $\mathbf{A}_S, \mathbf{b}_S, \mathcal{K}_S$.

**Theorem 18** *Suppose $\mathbb{P}^*$ is known to have independent components, with supp$(\mathbb{P}^*) \subseteq [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$.*

*i) With probability at least $1-\alpha$ over the sample, the family $\{\mathcal{U}_\epsilon^I : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee, where*

$$\mathcal{U}_\epsilon^I = \Bigg\{ \mathbf{u} \in \mathbb{R}^d : \exists \mathbf{p}^i \in \mathcal{P}_i^{EDF},\ \mathbf{q}^i \in \Delta_{N+2},\ i = 1 \ldots, d, \\ \sum_{j=0}^{N+1} \hat{u}_i^{(j)} q_j^i = u_i\ i = 1, \ldots, d,\ \sum_{i=1}^{d} D\left(\mathbf{q}^i, \mathbf{p}^i\right) \leq \log(1/\epsilon) \Bigg\}. \tag{76}$$

*ii) In the special case of the K test, the above formulation simplifies to (21) with $\Gamma^{KS}$ replaced by $\Gamma^K$.*

The proof of the first part is entirely analogous to Theorem 6, but uses Theorem 17 to evaluate the worst-case expectations. The proof of the second part follows by applying Corollary 1. We omit the details.

*Remark 24* In contrast to our definition of $\mathcal{U}_\epsilon^I$ using the KS test, we know of no simple algorithm for evaluating $\delta^*(\mathbf{v}|\mathcal{U}_\epsilon^I)$ when using the CvM, W, or AD tests. (For the K test, the same algorithm applies but with $\Gamma^K$ replacing $\Gamma^{KS}$.) Although it still polynomial time to optimize over constraints $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^I) \leq t$ for these tests using interior-point solvers for conic optimization, it is more challenging numerically.

## References


1. Acerbi, C., Tasche, D.: On the coherence of expected shortfall. J. Bank. Financ. **26**(7), 1487–1503 (2002)
2. Arlot, S., Celisse, A., et al.: A survey of cross-validation procedures for model selection. Stat. Surv. **4**, 40–79 (2010)
3. Bandi, C., Bertsimas, D.: Tractable stochastic analysis in high dimensions via robust optimization. Math. Program. **134**(1), 23–70 (2012)
4. Bandi, C., Bertsimas, D., Youssef, N.: Robust queueing theory. Oper. Res. **63**(3), 676–700 (2012)
5. Ben-Tal, A., Den Hertog, D., Vial, J.P.: Deriving robust counterparts of nonlinear uncertain inequalities. Math. Program. **149**, 1–35 (2012)
6. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust Optimization. Princeton University Press, Princeton (2009)
7. Ben-Tal, A., Golany, B., Nemirovski, A., Vial, J.: Retailer-supplier flexible commitments contracts: a robust optimization approach. Manuf. Serv. Oper. Manag. **7**(3), 248–271 (2005)

8. Ben-Tal, A., Hazan, E., Koren, T., Mannor, S.: Oracle-based robust optimization via online learning. Oper. Res. **63**(3), 628–638 (2015)
9. Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. Manag. Sci. **59**(2), 341–357 (2013)
10. Ben-Tal, A., Nemirovski, A.: Robust solutions of linear programming problems contaminated with uncertain data. Math. Program. **88**(3), 411–424 (2000)
11. Bertsekas, D., Nedi, A., Ozdaglar, A.: Convex Analysis and Optimization. Athena Scientific, Belmont (2003)
12. Bertsimas, D., Brown, D.: Constructing uncertainty sets for robust linear optimization. Oper. Res. **57**(6), 1483–1495 (2009)
13. Bertsimas, D., Dunning, I., Lubin, M.: Reformulations versus cutting planes for robust optimization (2014). http://www.optimization-online.org/DB_HTML/2014/04/4336.html
14. Bertsimas, D., Gamarnik, D., Rikun, A.: Performance analysis of queueing networks via robust optimization. Oper. Res. **59**(2), 455–466 (2011)
15. Bertsimas, D., Gupta, V., Kallus, N.: Robust sample average approximation (2013). arxiv:1408.4445
16. Bertsimas, D., Sim, M.: The price of robustness. Oper. Res. **52**(1), 35–53 (2004)
17. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
18. Calafiore, G., El Ghaoui, L.: On distributionally robust chance-constrained linear programs. J. Optim. Theory Appl. **130**(1), 1–22 (2006)
19. Calafiore, G., Monastero, B.: Data-driven asset allocation with guaranteed short-fall probability. In: American Control Conference (ACC), 2012, pp. 3687–3692. IEEE (2012)
20. Campi, M., Car, A.: Random convex programs with l_1-regularization: sparsity and generalization. SIAM J. Control Optim. **51**(5), 3532–3557 (2013)
21. Campi, M., Garatti, S.: The exact feasibility of randomized solutions of uncertain convex programs. SIAM J. Optim. **19**(3), 1211–1230 (2008)
22. Chen, W., Sim, M., Sun, J., Teo, C.: From CVaR to uncertainty set: implications in joint chance-constrained optimization. Oper. Res. **58**(2), 470–485 (2010)
23. Chen, X., Sim, M., Sun, P.: A robust optimization perspective on stochastic programming. Oper. Res. **55**(6), 1058–1071 (2007)
24. David, H., Nagaraja, H.: Order Statistics. Wiley Online Library, New York (1970)
25. Delage, E., Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. Oper. Res. **58**(3), 596–612 (2010)
26. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap, vol. 57. CRC Press, Boca Raton (1993)
27. Embrechts, P., Höing, A., Juri, A.: Using copulae to bound the value-at-risk for functions of dependent risks. Financ. Stoch. **7**(2), 145–167 (2003)
28. Esfahani, P.M., Kuhn, D.: Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. Preprint (2015). arXiv:1505.05116
29. Goldfarb, D., Iyengar, G.: Robust portfolio selection problems. Math. Oper. Res. **28**(1), 1–38 (2003)
30. Grötschel, M., Lovász, L., Schrijver, A.: The ellipsoid method and its consequences in combinatorial optimization. Combinatorica **1**(2), 169–197 (1981)
31. Hastie, T., Friedman, J., Tibshirani, R.: The Elements of Statistical Learning, vol. 2. Springer, Berlin (2009)
32. Jager, L., Wellner, J.A.: Goodness-of-fit tests via phi-divergences. Ann. Stat. **35**(5), 2018–2053 (2007)
33. Kingman, J.: Some inequalities for the queue GI/G/1. Biometrika **49**(3/4), 315–324 (1962)
34. Klabjan, D., Simchi-Levi, D., Song, M.: Robust stochastic lot-sizing by means of histograms. Prod. Oper. Manag. **22**(3), 691–710 (2013)
35. Lehmann, E., Romano, J.: Testing Statistical Hypotheses. Texts in Statistics. Springer, Berlin (2010)
36. Lindley, D.: The theory of queues with a single server. In: Mathematical Proceedings of the Cambridge Philosophical Society, vol. 48, pp. 277–289. Cambridge University Press, Cambridge (1952)
37. Lobo, M., Vandenberghe, L., Boyd, S., Lebret, H.: Applications of second-order cone programming. Linear Algebra Appl. **284**(1), 193–228 (1998)
38. Mutapcic, A., Boyd, S.: Cutting-set methods for robust convex optimization with pessimizing oracles. Optim. Methods Softw. **24**(3), 381–406 (2009)
39. Natarajan, K., Dessislava, P., Sim, M.: Incorporating asymmetric distributional information in robust value-at-risk optimization. Manag. Sci. **54**(3), 573–585 (2008)
40. Nemirovski, A.: Lectures on modern convex optimization. In: Society for Industrial and Applied Mathematics (SIAM). Citeseer (2001)

41. Nemirovski, A., Shapiro, A.: Convex approximations of chance constrained programs. SIAM J. Optim. **17**(4), 969–996 (2006)
42. Rice, J.: Mathematical Statistics and Data Analysis. Duxbury press, Pacific Grove (2007)
43. Rockafellar, R., Uryasev, S.: Optimization of conditional value-at-risk. J. Risk **2**, 21–42 (2000)
44. Rusmevichientong, P., Topaloglu, H.: Robust assortment optimization in revenue management under the multinomial logit choice model. Oper. Res. **60**(4), 865–882 (2012)
45. Shapiro, A.: On duality theory of conic linear problems. In: Goberna, M.Á., López, M.A. (eds.) Semi-infinite Programming, pp. 135–165. Springer, Berlin (2001)
46. Shawe-Taylor, J., Cristianini, N.: Estimating the moments of a random vector with applications (2003). http://eprints.soton.ac.uk/260372/1/EstimatingTheMomentsOfARandomVectorWithApplications.pdf
47. Stephens, M.: EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. **69**(347), 730–737 (1974)
48. Thas, O.: Comparing Distributions. Springer, Berlin (2010)
49. Wang, Z., Glynn, P.W., Ye, Y.: Likelihood robust optimization for data-driven newsvendor problems. Tech. rep., Working paper (2009)
50. Wiesemann, W., Kuhn, D., Sim, M.: Distributionally robust convex optimization. Oper. Res. **62**(6), 1358–1376 (2014)