Submitted to *Management Science*
manuscript

Near-Optimal Bayesian Ambiguity Sets for Distributionally Robust Optimization

Vishal Gupta
Data Science and Operations, USC Marshall School of Business, Los Angeles, CA, 90089, guptavis@usc.edu

We propose a Bayesian framework for assessing the relative strengths of data-driven ambiguity sets in distributionally robust optimization (DRO) when the underlying distribution is defined by a finite-dimensional parameter. The key idea is to measure the relative size between a candidate ambiguity set and a specific asymptotically optimal set. As the amount of data grows large, this asymptotically optimal set is the smallest convex ambiguity set that satisfies a novel Bayesian robustness guarantee that we introduce. This guarantee is defined with respect to a given class of constraints and is a Bayesian analog of more common frequentist feasibility guarantees from the DRO literature. Using this framework, we prove that many popular existing ambiguity sets are significantly larger than the asymptotically optimal set for constraints that are concave in the ambiguity. By contrast, we construct new ambiguity sets that are tractable, satisfy our Bayesian robustness guarantee and are at most a small, constant factor larger than the asymptotically optimal set; we call these sets *Bayesian near-optimal*. We further prove that asymptotically, solutions to DRO models with our Bayesian near-optimal sets enjoy strong frequentist robustness properties, despite their smaller size. Finally, our framework yields guidelines for practitioners selecting between competing ambiguity set proposals in DRO. Computational evidence in portfolio allocation using real and simulated data confirms that our framework, although motivated by asymptotic analysis in a Bayesian setting, provides practical insight into the performance of various DRO models with finite data under frequentist assumptions.

Key words: robust optimization, data-driven optimization, Bayesian statistics

History: This paper was first submitted in July 2015 and underwent three revisions. It was accepted for publication on 24 May 2018.

1. Introduction

Many applications in decision-making under uncertainty can be modeled as optimization problems where constraints may depend on both the decision variables $\mathbf{x}$ and the distribution $\mathbb{P}^*$ of some

random variables $\tilde{\boldsymbol{\xi}}$. For example, in inventory management problems, constraints on the probability of stock-outs depend on both the ordering policy ($\mathbf{x}$) and the distribution of future demand ($\mathbb{P}^*$). Generically, we can write such constraints as $g(\mathbb{P}^*, \mathbf{x}) \leq 0$ for some function g.

The difficulty is that $\mathbb{P}^*$ is rarely known in practice. At best, we have a dataset $\mathcal{S} = \{\hat{\boldsymbol{\xi}}^1, \ldots, \hat{\boldsymbol{\xi}}^N\}$ drawn from $\mathbb{P}^*$. The distributionally robust optimization (DRO) approach to such problems is to construct an *ambiguity set* $\mathcal{P}(\mathcal{S})$ of potential distributions $\mathbb{P}$ and replace the constraint $g(\mathbb{P}^*, \mathbf{x}) \leq 0$ with the robust constraint

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{S})} g(\mathbb{P}, \mathbf{x}) \leq 0, \tag{1}$$

which depends on $\mathcal{P}(\mathcal{S})$. Despite its seeming complexity, Eq. (1) is computationally tractable for many combinations of g and $\mathcal{P}(\mathcal{S})$ (Ben-Tal et al. (2015), Wiesemann et al. (2014)).

Since the seminal work of Scarf (1958) in inventory control, DRO models with different $\mathcal{P}(\mathcal{S})$ have been proposed for supply-chain design, revenue management, finance, and other applications (see, e.g., Klabjan et al. (2013), Lim and Shanthikumar (2007), Postek et al. (2016)). Empirical evidence confirms that DRO offers benefits over methods that neglect ambiguity in the unknown $\mathbb{P}^*$. This combination of tractability and effectiveness has fueled the increasing popularity of DRO in operations management. However, empirical evidence also suggests that the performance of DRO models crucially depends on the choice of $\mathcal{P}(\mathcal{S})$.

This last observation raises several questions: Is there a "best" possible $\mathcal{P}(\mathcal{S})$? What does "best" mean? If we select an alternative, perhaps simpler ambiguity set for numerical tractability, what is the loss in performance relative to this "best" possible set? Are there simple guidelines for constructing ambiguity sets, selecting between competing proposals and formulating DRO models?

In this work, we propose a novel Bayesian framework for analyzing ambiguity sets in data-driven DRO to answer these questions. Our analysis requires two key assumptions:

Assumption 1. *$\mathbb{P}^*$ is defined by a finite-dimensional parameter, i.e., $\mathbb{P}^* = \mathbb{P}_{\boldsymbol{\theta}^*}$ for some $\boldsymbol{\theta}^* \in \Theta \subseteq \mathbb{R}^d$.*

Assumption 2. *For any fixed $\mathbf{x}$, the function g is closed and concave in $\boldsymbol{\theta}$.*

Let $\mathcal{G}$ denote the set of functions satisfying A2.

A1 is sufficiently general to include a number of special cases of DRO, including when $\mathbb{P}_{\boldsymbol{\theta}^*}$ belongs to a parametric class such as normal distributions; when $\mathbb{P}_{\boldsymbol{\theta}^*}$ is non-parametric but has known, finite, discrete support; or when $\mathbb{P}_{\boldsymbol{\theta}^*}$ is a finite mixture model with known components. Importantly, A1 allows us to rewrite Eq. (1) (by redefining g and $\mathcal{P}(\mathcal{S})$) as

$$g(\boldsymbol{\theta}, \mathbf{x}) \leq 0 \quad \forall \boldsymbol{\theta} \in \mathcal{P}(\mathcal{S}), \qquad \text{with } \mathcal{P}(\mathcal{S}) \subseteq \mathbb{R}^d. \tag{2}$$

A2 is also mild. Practically, many constraints found in DRO applications are concave in $\boldsymbol{\theta}$ (cf. Ex. 1). This observation may not be surprising, as determining the feasibility of a fixed $\mathbf{x}$ in Eq. (2) for non-concave g requires maximizing a non-concave objective and may be numerically challenging.

Under A1 and A2, it is possible to meaningfully define a notion of "best" and quantify the relative strength of different sets. The key idea is to identify the smallest convex ambiguity set that satisfies a novel Bayesian robustness guarantee (see Def. 2). By smallest, we mean that the set is a subset of any other convex set which also satisfies this guarantee. We use this set as a benchmark to assess the relative size of other ambiguity sets.

We define our Bayesian robustness guarantee in Sec. 2. It is defined with respect to a given class of functions g and is a Bayesian analogue of a standard (frequentist) guarantee (Def. 1) used to measure the robustness of sets in the literature (Ben-Tal et al. 2009, Bertsimas et al. 2017a,b).

Our use of size to proxy performance, however, is less standard and motivated by Eq. (2). If one ambiguity set is a subset of another, the smaller set always yields solutions with better objective values. This improvement entails no loss in robustness if both sets satisfy the same robustness guarantee over the same class of functions. In this sense, the smallest set that satisfies this guarantee is "optimal." To emphasize our Bayesian robustness guarantee, we call such a set *Bayesian optimal*.

We prove that although a Bayesian optimal set for $\mathcal{G}$ need not exist for finite N, it always exists under mild assumptions as $N \to \infty$. For many popular ambiguity sets, we can calculate their size relative to this asymptotically Bayesian optimal set explicitly. Intuitively, this relative size provides a good metric for choosing between competing proposals when N is large.

Perhaps surprisingly, we prove that popular proposals for ambiguity sets based upon frequentist confidence regions are relatively large. This includes, for example, the ϕ-divergence sets of Ben-Tal et al. (2013) and the elliptical set of Zhu et al. (2014). Indeed, the relative ratio of such ambiguity set's size to the asymptotically Bayesian optimal set's size scales like $\Omega(\sqrt{d})$. (Recall that $d = \dim(\boldsymbol{\theta})$.) By contrast, we construct novel ambiguity sets that satisfy our Bayesian robustness property over $\mathcal{G}$ and are at most a small, constant factor (independent of d) larger than the asymptotically Bayesian optimal set for $\mathcal{G}$. We call these new sets *Bayesian near-optimal* for $\mathcal{G}$.

This distinction in size has an important practical consequence: When d is moderate to large, replacing the ambiguity set in many popular DRO models with one of our Bayesian near-optimal variants can improve performance while providing a similar robustness guarantee. We say "similar" because, strictly speaking, our robustness guarantee holds under Bayesian assumptions, while traditional ambiguity sets offer a frequentist guarantee.

Although developing a complete theory that reconciles the Bayesian and frequentist perspectives on ambiguity set construction is still open, we provide initial results comparing these viewpoints in Sec. 5. We highlight ways in which traditional (frequentist) ambiguity sets provide additional protection beyond our near-optimal Bayesian variants and also argue that in practical applications, this additional protection may be unnecessary, depending on one's goals. In particular, we prove that as $N \to \infty$, the *solutions* to DRO models with our Bayesian ambiguity sets often satisfy strong frequentist robustness properties, similar to the solutions using frequentist variants, despite our sets' smaller sizes. In Sec. 6, we study this phenomenon numerically and show that even for moderate N, solutions to DRO models using our Bayesian near-optimal sets often exhibit good frequentist behavior. Collectively, these features suggest Bayesian near-optimal sets may be attractive alternatives to traditional ambiguity sets for some applications.

A key idea in our work is that our near-optimal constructions exploit the concave structure of g. By contrast, popular frequentist proposals for ambiguity sets based on confidence regions do not exploit any such structure (Bertsimas et al. 2017b). While some authors have exploited concavity

to prove certain DRO models are tractable (Postek et al. (2016)), to our knowledge, we are the first to exploit concavity in constructing ambiguity sets. This concavity is crucial to our results. We prove that any ambiguity set that satisfies our Bayesian robustness property for even one specific convex, quadratic function of the uncertainty must be approximately as large as frequentist sets for large N. In other words, without concavity, the size advantage of our near-optimal sets disappears.

Our results parallel ideas in traditional robust optimization. In that context, it is well known that one can construct uncertainty sets that satisfy a robustness guarantee for concave g and that these sets are generally much smaller than frequentist confidence regions (see, e.g., Ben-Tal et al. (2009), Chen et al. (2007), Bertsimas et al. (2017a)). There is, however, no notion of an "optimal" set and no theoretical quantification of how much smaller these sets may be.

To the best of our knowledge, this parallel has not been utilized in constructing ambiguity sets for DRO. A possible explanation for this gap is that it is mathematically challenging to apply techniques from traditional robust optimization to the frequentist framework for DRO. One of our contributions is to show that the Bayesian viewpoint overcomes this difficulty.

Finally, we note that there is a stream of literature relating traditional and distributionally robust optimization to regularization in statistics, e.g., Xu and Mannor (2012), Fertis (2009), Xu et al. (2012). By convex duality, there is a bijection between ambiguity sets in DRO and positively homogenous, convex regularizers (see, e.g., Gotoh et al. (2015), Lam (2016)). Thus, insofar as our results concern picking "good" ambiguity sets for DRO, they can also be interpreted as picking "good" regularizers via this bijection.

To summarize our main contributions:

1. We prove that as $N \to \infty$, there exists a smallest-possible, convex ambiguity set that satisfies a Bayesian analogue of a common frequentist robustness property for all $g \in \mathcal{G}$. We term this set *asymptotically Bayesian optimal* for $\mathcal{G}$. Such sets need not exist for finite N.
2. We propose new ambiguity sets that, for finite N, satisfy our Bayesian robustness property for all $g \in \mathcal{G}$ and are tractable. Solutions to DRO problems using our new sets converge to

solutions of the full-information stochastic optimization problem (where $\mathbb{P}_{\boldsymbol{\theta}^*}$ is known) as $N \to \infty$. Most importantly, we prove that our new sets are at most a small, explicit, constant factor larger than the Bayesian asymptotically optimal set as $N \to \infty$. We term such sets *Bayesian near-optimal* for $\mathcal{G}$.

3. By contrast, we prove that any ambiguity set that satisfies the frequentist guarantee for all $g \in \mathcal{G}$ must be much larger; there exist directions in which these sets are at least $\Omega(\sqrt{d})$ times larger than the asymptotically Bayesian optimal set. When $\mathbb{P}_{\boldsymbol{\theta}^*}$ has known, finite, discrete support, we strengthen this result, showing that the class of ϕ-divergence ambiguity sets is at least $\Omega(\sqrt{d})$ times larger than the Bayesian asymptotically optimal set in *every* direction.
4. We prove that under mild assumptions, as $N \to \infty$, solutions to DRO problems with our Bayesian sets are feasible with respect to the true (unknown) constraint with high *frequentist* probability. Thus, for large N, Bayesian near-optimal sets may offer less conservative solutions than frequentist sets while providing solutions with similar frequentist properties.
5. We prove that concavity is essential to the above size distinction. Specifically, if we require that an ambiguity set satisfies our Bayesian robustness guarantee for a particular convex, quadratic function, then the set must be comparably large to existing frequentist proposals.
6. We provide computational evidence in portfolio allocation using real and simulated data, confirming that despite being motivated by Bayesian assumptions and asymptotic analysis, our theoretical results give practical insight into the empirical performance of DRO models in frequentist settings for moderate to large N. In particular, our near-optimal sets can significantly outperform existing proposals in this application. We propose general guidelines for selecting ambiguity sets in Appendix C.

1.1. Notations

Ordinary lowercase letters (e.g., p_i, θ_i) denote scalars, boldfaced lowercase letters (e.g., $\mathbf{p}$, $\boldsymbol{\theta}$) denote vectors, boldfaced capital letters denote matrices (e.g., $\mathbf{A}$), and calligraphic capital letters (e.g., $\mathcal{X}$, $\mathcal{S}$) denote sets. A superscript tilde (e.g., $\tilde{\theta}_i$, $\tilde{\boldsymbol{\theta}}$, $\tilde{\mathcal{S}}$) denotes a random quantity. Let $\mathbf{e}_i$ denote the i-th coordinate vector and $\mathbf{e}$ denote a vector of ones.

For any $\mathcal{P} \subseteq \mathbb{R}^d$ and $\alpha > 0$, let $\mathcal{P} + \mathbf{v} \equiv \{\mathbf{p} + \mathbf{v} : \mathbf{p} \in \mathcal{P}\}$ denote translation and $\alpha\mathcal{P} \equiv \{\alpha\mathbf{p} : \mathbf{p} \in \mathcal{P}\}$ denote dilation. Let $\text{ri}(\mathcal{P}) = \{\boldsymbol{\theta} \in \mathcal{P} : \forall \mathbf{z} \in \mathcal{P}, \exists \lambda > 1 \text{ s.t. } \lambda\boldsymbol{\theta} + (1-\lambda)\mathbf{z} \in \mathcal{P}\}$ denote the relative interior of $\mathcal{P}$ (cf. Bertsekas 1999). Finally, for any positive definite matrix $\mathbf{M}$, define the norm $\|\mathbf{y}\|_{\mathbf{M}} \equiv \sqrt{\mathbf{y}^T\mathbf{M}^{-1}\mathbf{y}}$. When $\mathbf{M}$ is positive semidefinite and a generalized inverse $\mathbf{M}^{-1}$ is clear from context, let $\|\mathbf{y}\|_{\mathbf{M}}$ denote the corresponding semi-norm.

2. Model Setup and Background

We study a single constraint of the form Eq. (2) (uncertain objectives can be studied via an epigraphic formulation). Let $\mathcal{X}(\mathcal{P}) = \{\mathbf{x} : g(\boldsymbol{\theta}, \mathbf{x}) \leq 0, \forall \boldsymbol{\theta} \in \mathcal{P}\}$. Let $\tilde{\boldsymbol{\xi}} \sim \mathbb{P}_{\boldsymbol{\theta}^*}$ be a random variable, where $\mathbb{P}_{\boldsymbol{\theta}^*}$ is defined by a fixed, unknown $\boldsymbol{\theta}^* \in \Theta \subseteq \mathbb{R}^d$.[1] Throughout, we assume Θ is convex. Let $\tilde{\mathcal{S}} = \{\tilde{\boldsymbol{\xi}}^1, \ldots, \tilde{\boldsymbol{\xi}}^N\}$ denote our data, where $\tilde{\mathcal{S}} \sim \mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}^*}$ and $\mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}^*}$ is fully defined by $\boldsymbol{\theta}^*$. For example, when $\tilde{\mathcal{S}}$ is drawn i.i.d. from $\mathbb{P}_{\boldsymbol{\theta}^*}$, $\mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}^*} = \prod_{j=1}^N \mathbb{P}_{\boldsymbol{\theta}^*}$. Since our key results will not require this independence, we prefer the notation $\mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}^*}$, and when $\boldsymbol{\theta}^*$ clear from context, we write $\mathbb{P}_{\mathcal{S}}$.

As mentioned, we adopt a Bayesian viewpoint of DRO, assuming $\boldsymbol{\theta}^*$ is the realization of a random variable $\tilde{\boldsymbol{\theta}} \sim \mathbb{P}_{\tilde{\boldsymbol{\theta}}}$, where $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$ is a prior supported on Θ. For any $\mathcal{S}$, $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}$ denotes the posterior distribution of $\tilde{\boldsymbol{\theta}}$. Most of our results do not depend on the choice of prior. In practice, one might take $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$ to be a suitably uninformative prior, such as the uniform distribution if Θ is compact.

Ambiguity sets in data-driven DRO are functions $\mathcal{P}(\cdot)$ that send $\mathcal{S} \mapsto \mathcal{P}(\mathcal{S}) \subseteq \Theta$. Their "robustness" is typically quantified via a feasibility guarantee. Fix any ϵ, $0 < \epsilon < 0.5$.

Definition 1 (Frequentist Feasibility). The function $\mathcal{P}(\cdot)$ satisfies the frequentist feasibility guarantee at level ϵ for $g \in \mathcal{G}$ if $\mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}^*}\left(g(\boldsymbol{\theta}^*, \mathbf{x}) \leq 0,\ \forall \mathbf{x} \in \mathcal{X}(\mathcal{P}(\tilde{\mathcal{S}}))\right) \geq 1 - \epsilon$ for any $\boldsymbol{\theta}^* \in \Theta$.

Def. 1 is a key motivation for DRO; it asserts that *any* $\mathbf{x}$ that is robust feasible with respect to $\mathcal{P}(\tilde{\mathcal{S}})$ in Eq. (2) is feasible with respect to the unknown $\mathbb{P}_{\boldsymbol{\theta}^*}$ with probability at least $1 - \epsilon$. Ideally, this guarantee will hold for a large class of functions g. For example, Ben-Tal et al. (2013) shows that ϕ-divergence sets satisfy this property for all measurable g whenever $\mathbb{P}_{\boldsymbol{\theta}^*}$ has known, finite, discrete support, while Delage and Ye (2010) shows that a specific ambiguity set based on the first

[1] We briefly discuss extensions of our main results to the case where $\boldsymbol{\theta}^*$ is infinite dimensional in Appendix B.

two moments of a distribution satisfies this property for all measurable g whenever $\mathbb{P}_{\boldsymbol{\theta}^*}$ has bounded support. Similarly, Bertsimas et al. (2017b) presents several ambiguity sets based on hypothesis tests, with each set satisfying this property for different classes of g under various assumptions on $\mathbb{P}_{\boldsymbol{\theta}^*}$, including all measurable functions, all separable functions, and certain polynomial functions.

Next, we introduce a novel Bayesian analogue of Def. 1.

Definition 2 (Posterior Feasibility). The set $\mathcal{P}(\mathcal{S})$ satisfies the posterior feasibility guarantee at level ϵ for g if $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(g(\tilde{\boldsymbol{\theta}}, \mathbf{x}) \leq 0\right) \geq 1-\epsilon$ for all $\mathbf{x} \in \mathcal{X}(\mathcal{P}(\mathcal{S}))$. The function $\mathcal{P}(\cdot)$ satisfies the posterior guarantee if $\mathcal{P}(\mathcal{S})$ satisfies the posterior guarantee for all $\mathcal{S}$.

The posterior feasibility guarantee also asserts that *any* $\mathbf{x}$ that is robust feasible with respect to $\mathcal{P}(\mathcal{S})$ will be feasible with respect to the unknown $\mathbb{P}_{\boldsymbol{\theta}^*}$ with probability at least $1-\epsilon$. The difference from Def. 1 is the meaning of this probability. The frequentist probability in Def. 1 fixes the ground-truth $\boldsymbol{\theta}^*$ and considers the probability over repeated (random) draws of potential datasets $\tilde{\mathcal{S}}$, i.e., $\mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}^*}$. Thus, the frequentist framework is sometimes described as "repeated sampling." By contrast, the posterior probability in Def. 2 fixes the realized $\mathcal{S}$ and considers the probability over the residual uncertainty in the unknown realization of $\tilde{\boldsymbol{\theta}}$, i.e., $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}$.

The relative merits of frequentist vs. Bayesian modeling have been fiercely debated in the statistics literature (see Efron and Hastie (2016, Chapt. 2, 3) for a modern viewpoint and references). From a DRO perspective, an example may help to clarify some modeling consequences: Consider an inventory manager stocking many similar products based upon historical demand data. Demand for product k follows $\mathbb{P}_{\tilde{\boldsymbol{\theta}}^k}$, where $\tilde{\boldsymbol{\theta}}^k$ is unknown. The manager's a priori knowledge about typical demand profiles, e.g., that demand for a typical product is between 10 and 20 units per month, is accurately encoded by the prior $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$, i.e., she assumes $\tilde{\boldsymbol{\theta}}^k$ are realizations of independent draws from $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$. For each k, the manager has historical data $\mathcal{S}^k$, which she models as a realization of $\tilde{\mathcal{S}}^k \sim \mathbb{P}_{\mathcal{S}|\tilde{\boldsymbol{\theta}}^k}$. Finally, suppose Eq. (2) is a constraint controlling the probability of a stockout. Consequently, for each k, she uses data $\mathcal{S}^k$ to form $\mathcal{P}(\mathcal{S}^k)$, solves Eq. (2), and stocks accordingly.

Fix a ground truth parameter $\boldsymbol{\theta}^*$ and only consider products k with $\tilde{\boldsymbol{\theta}}^k = \boldsymbol{\theta}^*$. For what proportion of these products do we expect her to stock-out? This setup approximately mirrors the frequentist

framework; the ground truth $\boldsymbol{\theta}^*$ is fixed, and we observe many data draws $\mathcal{S}^k$, each from $\mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}^*}$ (repeated sampling). If $\mathcal{P}(\cdot)$ satisfies the frequentist guarantee and there are many products with $\tilde{\boldsymbol{\theta}}^k = \boldsymbol{\theta}^*$, then we expect a stockout on no more than $\epsilon\%$ of these products.

Now, instead, fix some potential data realization $\mathcal{S}$ and only consider products k with $\mathcal{S}^k = \mathcal{S}$. For what proportion of *these* products do we expect her to stock-out? This setup approximately mirrors the Bayesian framework; the data $\mathcal{S}$ are fixed, and we consider residual uncertainty in each $\tilde{\boldsymbol{\theta}}^k$ as a realization of $\tilde{\boldsymbol{\theta}}$. If $\mathcal{P}(\cdot)$ satisfies the posterior feasibility guarantee and there are many products with $\mathcal{S}^k = \mathcal{S}$, then we expect a stockout on no more than $\epsilon\%$ of these products.

Arguably, both guarantees have limited relevance since in real world scenarios, few products may satisfy $\tilde{\boldsymbol{\theta}}^k = \boldsymbol{\theta}^*$ or $\mathcal{S}^k = \mathcal{S}$. However, the guarantees are useful insofar as $\boldsymbol{\theta}^*$ and $\mathcal{S}$ describe a product's specific context. Depending on the application, either guarantee may be of interest. Sec. 5 provides a more formal comparison of these two guarantees in context of DRO.

2.1. Tractability of Robust Constraints

The tractability of Eq. (2) under A2 is well studied. Ben-Tal et al. (2015) prove that for non-empty, convex, compact $\mathcal{P}(\mathcal{S})$ satisfying a mild regularity condition[2], Eq. (2) is equivalent to

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

Here, g_* denotes the partial concave conjugate of g, and $\delta^*(\mathbf{v}|\ \mathcal{P})$ denotes the support function of $\mathcal{P}$. These are respectively defined as

$$g_*(\mathbf{v}, \mathbf{x}) \equiv \inf_{\boldsymbol{\theta}} \left\{ \boldsymbol{\theta}^T \mathbf{v} - g(\boldsymbol{\theta}, \mathbf{x}) \right\}, \quad \delta^*(\mathbf{v}|\ \mathcal{P}) \equiv \sup_{\boldsymbol{\theta} \in \mathcal{P}} \mathbf{v}^T \boldsymbol{\theta}.$$

For many g, including bi-affine and conic quadratic representable functions, $g_*(\mathbf{v}, \mathbf{x})$ admits a simple, computationally tractable description. (We refer readers to Ben-Tal et al. (2015), Bertsimas et al. (2017a), Postek et al. (2016) for details and examples.) Consequently, under A2, to prove that Eq. (2) is computationally tractable for any such g, it suffices to show that we can solve the optimization defining $\delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}))$ tractably. This optimization only involves linear functions of $\boldsymbol{\theta}$. In what follows, we will say that $\mathcal{P}(\mathcal{S})$ is tractable whenever evaluating $\delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}))$ is tractable.

[2] An example of a sufficient regularity condition is that $\text{ri}(\mathcal{P}(\mathcal{S})) \cap \text{ri}(\text{dom}(g(\cdot, \mathbf{x}))) \neq \emptyset$ for all $\mathbf{x}$.

2.2. Examples

We recast some examples from the DRO literature in our framework by specifying $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$ and confirming concavity of $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta}, \cdot)$. In what follows, we utilize our framework to assess the strength of various ambiguity sets for these examples.

Example 1 (Finite, discrete support). Suppose $\tilde{\boldsymbol{\xi}}$ has known, finite, discrete support, i.e., $\tilde{\boldsymbol{\xi}} \in \{\mathbf{a}^1, \ldots, \mathbf{a}^d\}$, but that $\mathbb{P}_{\boldsymbol{\theta}^*}(\tilde{\boldsymbol{\xi}} = \mathbf{a}^i)$ is uncertain for $i = 1, \ldots, d$. Suppose also that $\tilde{\mathcal{S}}$ is drawn i.i.d. from $\mathbb{P}_{\boldsymbol{\theta}^*}$. Ben-Tal et al. (2013), Klabjan et al. (2013), Postek et al. (2016), Bertsimas et al. (2017b) study DRO problems involving these unknown probabilities with applications in portfolio allocation and inventory management and propose various ambiguity sets $\mathcal{P}(\mathcal{S})$.

We cast this setting in our framework by letting $\theta_j^* \equiv \mathbb{P}_{\boldsymbol{\theta}^*}(\tilde{\boldsymbol{\xi}} = \mathbf{a}^j)$ for $j = 1, \ldots, d$, and $\Theta \equiv \Delta_d = \{\boldsymbol{\theta} \in \mathbb{R}_+^d : \mathbf{e}^T\boldsymbol{\theta} = 1\}$. We adopt a Dirichlet prior for $\tilde{\boldsymbol{\theta}}$. Recall that $\tilde{\boldsymbol{\theta}}$ follows a Dirichlet distribution with parameter $\boldsymbol{\tau}'$ if it admits the probability density $f_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}) = B(\boldsymbol{\tau}')^{-1} \prod_{i=1}^d \theta_i^{\tau_i' - 1}$, where $\tau_i' > 0$ for all i and $B(\boldsymbol{\tau}')$ is a normalizing constant. The Dirichlet distribution is a conjugate prior in this setting, meaning $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\tilde{\mathcal{S}}}$ is also Dirichlet with updated parameter $\boldsymbol{\tau}$, $\tau_i = \tau_i' + \sum_{j=1}^N \mathbb{I}(\hat{\boldsymbol{\xi}}^j = \mathbf{a}^i)$. When $\boldsymbol{\tau}' = \mathbf{e}$, the Dirichlet distribution is a uniform distribution, a common uninformative prior.

As observed in Postek et al. (2016), most common constraints involve $g \in \mathcal{G}$:

Expectation and Chance Constraints: For any function $v(\tilde{\boldsymbol{\xi}}, \mathbf{x})$, the constraint $\mathbb{E}^{\mathbb{P}_{\boldsymbol{\theta}}}[v(\tilde{\boldsymbol{\xi}}, \mathbf{x})] \leq 0$ is equivalent to $\sum_{j=1}^d \theta_j v(\mathbf{a}^j, \mathbf{x}) \leq 0$, which is linear, and therefore concave, in $\boldsymbol{\theta}$. Chance constraints are a special case of expectations since $\mathbb{P}_{\boldsymbol{\theta}}(v(\tilde{\boldsymbol{\xi}}, \mathbf{x}) \leq 0) = \mathbb{E}^{\mathbb{P}_{\boldsymbol{\theta}}}[\mathbb{I}(v(\tilde{\boldsymbol{\xi}}, \mathbf{x}) \leq 0)]$.

Conditional Value at Risk and Spectral Risk Measures: For any function $v(\tilde{\boldsymbol{\xi}}, \mathbf{x})$, the conditional value at risk of $v(\tilde{\boldsymbol{\xi}}, \mathbf{x})$ at level λ is defined by $\mathrm{CVaR}_\lambda^{\mathbb{P}_{\boldsymbol{\theta}}}(v(\tilde{\boldsymbol{\xi}}, \mathbf{x})) \equiv \min_\beta \left\{\beta + \frac{1}{\lambda}\mathbb{E}^{\mathbb{P}_{\boldsymbol{\theta}}}[v(\tilde{\boldsymbol{\xi}}, \mathbf{x}) - \beta]^+\right\}$. Conditional value at risk is a popular risk measure in financial applications. Since expectations are linear in $\boldsymbol{\theta}$, $\mathrm{CVaR}_\lambda^{\mathbb{P}_{\boldsymbol{\theta}}}$ is the minimum of a set of linear functions and, hence, concave in $\boldsymbol{\theta}$. Spectral risk measures are generalizations of $\mathrm{CVaR}_\lambda^{\mathbb{P}_{\boldsymbol{\theta}}}$. Under suitable regularity conditions, a spectral risk measure $\rho(v(\tilde{\boldsymbol{\xi}}, \mathbf{x}))$ can be rewritten as $\int_0^1 \mathrm{CVaR}_\lambda^{\mathbb{P}_{\boldsymbol{\theta}}}(v(\tilde{\boldsymbol{\xi}}, \mathbf{x}))\nu(d\lambda)$ for some measure ν (Noyan and Rudolf 2014). As a positive combination of concave functions of $\boldsymbol{\theta}$, spectral risk measures are also concave.

Mean Absolute Deviation: Certain statistical measures are also concave in $\boldsymbol{\theta}$. For example, the mean absolute deviation from the median, $\mathbb{E}^{\mathbb{P}_\theta}[|v(\tilde{\boldsymbol{\xi}},\mathbf{x}) - \text{Median}(v(\tilde{\boldsymbol{\xi}},\mathbf{x}))|]$, can be rewritten as $\min_\beta \mathbb{E}^{\mathbb{P}_\theta}[|v(\tilde{\boldsymbol{\xi}},\mathbf{x}) - \beta|]$, which is the minimum of linear functions in $\boldsymbol{\theta}$ and, hence, concave.

There are examples of natural constraints that are not concave in $\boldsymbol{\theta}$. For example, bounds on coefficient of variation are generally non-concave, although they can sometimes be reformulated to be concave (see Postek et al. (2016)).

Example 2 (Finite mixtures of known distributions). Suppose instead that $\tilde{\boldsymbol{\xi}}$ follows a mixture distribution, i.e., $\tilde{\boldsymbol{\xi}} \sim \sum_{i=1}^d \theta_i^* F_i$, where each F_i is a known distribution function but $\boldsymbol{\theta}^* \in \Delta_d$ is unknown. Zhu and Fukushima (2009), Zhu et al. (2014) propose ambiguity sets for $\boldsymbol{\theta}^*$ and formulate DRO problems for particular financial applications. In their applications, each F_i represents the distribution of asset returns under a possible future market scenario i.

This example generalizes Ex. 1 and similarly maps to our framework by taking $\Theta = \Delta_d$. We again propose a Dirichlet prior for $\tilde{\boldsymbol{\theta}}$. In this setting, the posterior distribution is not Dirichlet and must be determined numerically, e.g., using MCMC methods (Gelman et al. 2014, Chapt. 11-12). Both open-source and commercial implementations of these methods are widely available. All examples of concave constraints from Ex. 1 remain concave in this setting.

Exs. 1 and 2 utilize very flexible classes of distributions and are general purpose. Appendix A details several other, more specialized examples leveraging parametric distributions, including Gaussian and time-series models, assortment optimization under the multinomial logit model, and pricing under generalized linear models.

3. Constructing Bayesian Ambiguity Sets

We first use A2 to characterize the Bayesian feasibility guarantee geometrically. This theorem was proven in a different context in Bertsimas et al. (2017a).

Theorem 1. *Fix any $\mathcal{S}$ and suppose $\mathcal{P}(\mathcal{S})$ is non-empty, closed, and convex. Then, $\mathcal{P}(\mathcal{S})$ satisfies the posterior feasibility guarantee for all $g \in \mathcal{G}$ at level ϵ if, and only if,*

$$\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\big(\mathbf{v}^T\tilde{\boldsymbol{\theta}} \le \delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}))\big) \ge 1-\epsilon \quad \forall \mathbf{v} \in \mathbb{R}^d. \tag{4}$$

Leveraging Thm. 1, we adapt the approach of Bertsimas et al. (2017a) for constructing uncertainty sets in traditional robust optimization to construct novel Bayesian ambiguity sets in DRO.

Define $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) \equiv \inf\{t : \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}^T\tilde{\boldsymbol{\theta}} \leq t) \geq 1-\epsilon\}$ to be the posterior value at risk of $\mathbf{v}^T\tilde{\boldsymbol{\theta}}$. From Eq. (4), $\mathcal{P}(\mathcal{S})$ satisfies the posterior feasibility guarantee for all $g \in \mathcal{G}$ at level ϵ if, and only if,

$$\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) \leq \delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S})) \quad \forall \mathbf{v} \in \mathbb{R}^d. \tag{5}$$

Thus, to construct an ambiguity set that satisfies the posterior feasibility guarantee, it suffices to 1) compute a closed, convex, positively homogenous upper bound $\phi(\mathbf{v})$ to $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ and 2) identify the ambiguity set for which $\phi(\mathbf{v})$ is the support function.[3]

Recall from Nedic et al. (2003) that, for any two sets,

$$\mathcal{P}_1 \subseteq \mathcal{P}_2 \iff \delta^*(\mathbf{v}|\ \mathcal{P}_1) \leq \delta^*(\mathbf{v}|\ \mathcal{P}_2) \quad \forall \mathbf{v} \in \mathbb{R}^d. \tag{6}$$

Thus, tighter upper bounds in Eq. (5) yield smaller ambiguity sets. A "tightest" upper bound would yield an "optimal" set.

Definition 3. We say that a $\mathcal{P}(\cdot)$ that satisfies the posterior feasibility guarantee at level ϵ for all $g \in \mathcal{G}$ is *Bayesian optimal* for $\mathcal{G}$ if, for any $\mathcal{S}$, $\mathcal{P}(\mathcal{S})$ is a subset of any other ambiguity set that satisfies the posterior feasibility guarantee at level ϵ for that $\mathcal{S}$ and all $g \in \mathcal{G}$.

Theorem 2. *A Bayesian optimal ambiguity set for $\mathcal{G}$ at level ϵ exists if, and only if, $VaR^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ is convex for all $\mathcal{S}$. When it exists, this set is unique and satisfies Eq.* (5) *with equality.*

Although it is possible to describe sufficient conditions on $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ for convexity of $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}$, in practice, these conditions are too restrictive to be useful.[4] Typically, $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ is non-convex.

[3] The existence of such a set is guaranteed by the bijection between closed, positively homogenous convex functions and closed, convex sets in convex analysis. See Nedic et al. (2003).

[4] For example, one can show that an optimal set exists if $\mathbb{P}_{\boldsymbol{\theta}}$ belongs to an exponential family and is log-concave and symmetric in $\boldsymbol{\theta}$.

3.1. A General Construction

Fortunately, there is a rich literature on upper-bounding $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ when it is non-convex. As observed in Bertsimas et al. (2017a), any of these bounds can be used to construct an ambiguity set that satisfies the posterior guarantee. We illustrate this idea using a bound proven in El Ghaoui et al. (2003). In our Bayesian context, given $\mathcal{S}$, the bound is

$$\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) \leq \mathbf{v}^T\boldsymbol{\mu}_N + \sqrt{\frac{1-\epsilon}{\epsilon}}\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}_N\mathbf{v}} \quad \forall \mathbf{v} \in \mathbb{R}^d, \tag{7}$$

where $\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N$ are the posterior mean and covariance of $\tilde{\boldsymbol{\theta}}$ given $\mathcal{S}$. When $\boldsymbol{\Sigma}_N$ is invertible, we can define for any $\Gamma > 0$,

$$\mathcal{P}^*(\mathcal{S}, \Gamma) \equiv \left\{ \boldsymbol{\theta} \in \Theta : \frac{1}{\sqrt{N}} \|\boldsymbol{\theta} - \boldsymbol{\mu}_N\|_{\boldsymbol{\Sigma}_N} \leq \Gamma \right\}. \tag{8}$$

We will see shortly that for Exs. 1 and 2, $\boldsymbol{\Sigma}_N$ is not invertible. Indeed, non-invertibility will occur whenever the affine dimension Θ is less than d. To remedy this, suppose Θ belongs to an r-dimensional affine subspace. By possibly permuting the indices, we assume without loss of generality that $\theta_1, \ldots, \theta_r$ span this space, i.e., there exists $\boldsymbol{\beta} \in \mathbb{R}^{d-r}$, $\mathbf{A} \in \mathbb{R}^{(d-r)\times r}$ such that

$$\boldsymbol{\theta}_{r+1,d} = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\theta}_{1,r}, \quad \forall \boldsymbol{\theta} \in \Theta, \tag{9}$$

where $\boldsymbol{\theta}_{1,r}$ are the first r components of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{r+1,d}$ are the remaining components. Define

$$\boldsymbol{\Sigma}_N^{-1} \equiv \begin{pmatrix} \boldsymbol{\Sigma}_{1,r}^{-1} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{0} \end{pmatrix}, \tag{10}$$

where $\boldsymbol{\Sigma}_{1,r}$ is the restriction of $\boldsymbol{\Sigma}_N$ to its first r rows and columns. By construction, $\boldsymbol{\Sigma}_N^{-1}$ inverts $\boldsymbol{\Sigma}_N$ on the space spanned by the first r components. When $\boldsymbol{\Sigma}_N$ is not invertible, we interpret Eq. (8) via this generalized inverse. Then, using Eq. (7) in the previous schema yields the following:

THEOREM 3. $\mathcal{P}^*\left(\cdot, \sqrt{\frac{1-\epsilon}{\epsilon N}}\right)$ *satisfies the posterior feasibility guarantee at level* ϵ *for all* $g \in \mathcal{G}$.

REMARK 1. $\mathcal{P}^*(\mathcal{S}, \Gamma)$ is tractable for any $\Gamma > 0$ whenever we can separate over Θ tractably (El Ghaoui et al. 2003). For example, when Θ is a polyhedron or SOCP representable, $\delta^*(\mathbf{v}|\ \mathcal{P}^*(\mathcal{S}, \Gamma))$ is also SOCP representable. When $\Theta = \mathbb{R}^d$, $\delta^*(\mathbf{v}|\ \mathcal{P}^*(\mathcal{S}, \Gamma))$ equals the righthand side of Eq. (7).

REMARK 2. Our definition of $\mathbf{\Sigma}_N^{-1}$ uses the basis $\mathbf{e}_1, \ldots, \mathbf{e}_r$. Other bases yield equivalent representations $\mathcal{P}^*(\mathcal{S}, \Gamma)$. Our choice simplifies exposition. Note that $\|\mathbf{y}\|_{\mathbf{\Sigma}_N}$ and $\|\mathbf{y}\|_{\mathbf{\Sigma}_N^{-1}}$ define semi-norms on $\mathbb{R}^d$, but define true norms on the linear subspace spanned by $\Theta - \boldsymbol{\theta}^*$.

Eq. (7) is only one of many possible upper bounds for $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ that can be used to create an ambiguity set. A computational benefit of $\mathcal{P}^*(\mathcal{S}, \Gamma)$ is that it only depends on the posterior mean and covariance, which are easily calculated by MCMC, rather than the full posterior distribution.

3.2. Ambiguity Sets for Distributions with Finite, Discrete Support

Recall Ex. 1, in which $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}$ is Dirichlet with parameter $\boldsymbol{\tau}$, and define $\tau_0 \equiv \sum_{i=1}^d \tau_i$.

THEOREM 4. *Suppose $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}$ is a Dirichlet distribution with parameter $\boldsymbol{\tau} > \mathbf{0}$. Then,*

1. *For $d = 2$, $VaR^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ is convex for all $\mathcal{S}$. The Bayesian optimal ambiguity set for $\mathcal{G}$ is*

$$\left\{ \lambda \begin{pmatrix} \beta_{1-\epsilon}(\tau_1, \tau_2) \\ 1 - \beta_{1-\epsilon}(\tau_1, \tau_2) \end{pmatrix} + (1-\lambda) \begin{pmatrix} 1 - \beta_{1-\epsilon}(\tau_1, \tau_2) \\ \beta_{1-\epsilon}(\tau_1, \tau_2) \end{pmatrix} \;:\; 0 \le \lambda \le 1 \right\},$$

   where $\beta_{1-\epsilon}(\tau_1, \tau_2)$ is the $1-\epsilon$-quantile of a Beta distribution with parameters τ_1, τ_2.

2. *For $d \ge 3$, there exist $\mathcal{S}$ such that $VaR^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ is non-convex. Consequently, there does not exist an optimal ambiguity set for $\mathcal{G}$.*

Since $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}}(\mathbf{v})$ may be non-convex, we seek convex upper bounds. Note that

$$\mu_{N,i} = \frac{\tau_i}{\tau_0}, \quad \mathbf{\Sigma}_N = \frac{1}{\tau_0 + 1} \left(\mathrm{diag}(\boldsymbol{\mu}_N) - \boldsymbol{\mu}_N \boldsymbol{\mu}_N^T \right), \tag{11}$$

are the posterior mean and covariance and that $\mathbf{\Sigma}_N$ is singular since $\mathbf{\Sigma}_N \mathbf{e} = \mathbf{0}$, corresponding to the fact that $\mathbf{e}^T \tilde{\boldsymbol{\theta}} = 1$ almost surely. Applying Eq. (10) yields

$$\mathbf{\Sigma}_N^{-1} \equiv (1 + \tau_0) \begin{pmatrix} \mathrm{diag}(\boldsymbol{\mu}_{N,-})^{-1} + \mu_{N,d}^{-1} \mathbf{e}\mathbf{e}^T & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}, \tag{12}$$

where $\boldsymbol{\mu}_{N,-}$ is the restriction of $\boldsymbol{\mu}_N$ to its first $d-1$ components. Define

$$\mathcal{P}^{\chi^2}(\mathcal{S}, \Gamma) \equiv \left\{ \boldsymbol{\theta} \in \Delta_d : \sum_{i=1}^d \frac{(\theta_i - \mu_{N,i})^2}{\mu_{N,i}} \le \Gamma^2 \right\}. \tag{13}$$

Substituting Eqs. (11) and (12) into Thm. 3 proves the following:

COROLLARY 1. *Under Ex. 1,* $\mathcal{P}^{\chi^2}\left(\cdot,\ \sqrt{\frac{1-\epsilon}{\epsilon(\tau_0+1)}}\right)$ *satisfies the posterior feasibility guarantee at level* ϵ *for all* $g \in \mathcal{G}$.

The proposed set resembles the χ^2-ambiguity set from Klabjan et al. (2013), Ben-Tal et al. (2013), Bertsimas et al. (2017a), etc. An important difference is the radius of the set: $\sqrt{\frac{1/\epsilon-1}{\tau_0+1}}$. In each of the previous works, the proposed radius is $\sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{N}}$. Here, $\chi^2_{d-1,1-\epsilon}$ is the $1-\epsilon$ quantile of a chi-square random variable with $d-1$ degrees of freedom. We expect $\tau_0 = O(N)$. Thus, our ambiguity set can be much smaller than this existing proposal, especially for large d, and still satisfy a posterior feasibility guarantee (see also Fig. 1). This is a first example of the general phenomenon that we discuss in detail in Secs. 4.3 and 5.

Corollary 1 utilizes the general-purpose Thm. 3. Define

$$\mathcal{P}^{KL}(\mathcal{S},\Gamma) \equiv \left\{ \boldsymbol{\theta} \in \Delta_d : \sum_{i=1}^{d} \mu_{N,i} \log\left(\frac{\mu_{N,i}}{\theta_i}\right) \leq \Gamma^2 \right\}. \tag{14}$$

By exploiting specific properties of the Dirichlet distribution, we have the following:

THEOREM 5. *Under Ex. 1,* $\mathcal{P}^{KL}\left(\cdot,\ \sqrt{\frac{\log(\frac{1}{\epsilon})}{\tau_0}}\right)$ *satisfies the posterior guarantee at level* ϵ *for all* $g \in \mathcal{G}$.

This set resembles the relative entropy set in Ben-Tal et al. (2013) and Bertsimas et al. (2017b) but enjoys a smaller radius: $\sqrt{\frac{\log(1/\epsilon)}{\tau_0}}$ (cf. Fig. 1). In previous works, the proposed radius is $\sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{2N}}$. (Again, $\tau_0 = O(N)$.) This is a second example of the aforementioned phenomenon.

REMARK 3. Ben-Tal et al. (2013) establish the tractability of $\mathcal{P}^{KL}(\mathcal{S},\Gamma)$ using an exponential cone optimization problem, which is polynomial-time solvable but numerically challenging. Bertsimas et al. (2017a) observe that for N sufficiently large, $\mathcal{P}^{KL}(\tilde{\mathcal{S}},\Gamma) \subseteq \mathcal{P}^{\chi^2}\left(\tilde{\mathcal{S}}, \Gamma + O(\sqrt{d}N^{-3/2})\right)$, $\mathbb{P}_{\mathcal{S}}$-a.s. This motivates heuristically replacing $\mathcal{P}^{KL}\left(\cdot, \sqrt{\log(1/\epsilon)/\tau_0}\right)$ with $\mathcal{P}^{\chi^2}\left(\cdot, \sqrt{\log(1/\epsilon)/\tau_0}\right)$ in applications, since the latter can be treated as a simpler second order cone optimization problem.

4. Asymptotics and Relative Size

Although optimal sets need not exist for finite N, asymptotically, an essentially optimal set does exist. Recall the classical Bernstein-von Mises Theorem (Chen 1985, Van der Vaart 2000).[5]

[5] As stated, the theorem slightly differs from, but is equivalent to, Thm. 10.1 of Van der Vaart (2000). See pg. 144 of that work for proof of the equivalence.

THEOREM 6 **(Bernstein-von Mises Theorem)**. *Suppose $\boldsymbol{\theta}^* \in ri(\Theta)$ and let the prior be absolutely continuous in a neighborhood of $\boldsymbol{\theta}^*$ with continuous, positive density at $\boldsymbol{\theta}^*$. Then, under mild regularity conditions,*

$$\sup_{\mathcal{A}} \left| \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\tilde{\mathcal{S}}}(\sqrt{N}(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\mu}}_N) \in \mathcal{A}) - \mathbb{P}(\tilde{\boldsymbol{\zeta}} \in \mathcal{A}) \right| \to_{\mathbb{P}_{\mathcal{S}}} 0, \qquad \text{as } N \to \infty,$$

where the supremum is taken over all measurable subsets $\mathcal{A}$ of Θ, $\tilde{\boldsymbol{\zeta}} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$ and $\mathcal{I}(\boldsymbol{\theta}^*)$ denotes the Fisher information matrix of $\mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}^*}$.

Intuitively, the Bernstein-von Mises Theorem describes the convergence of the random probability distribution $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\tilde{\mathcal{S}}}$ to the deterministic probability distribution, i.e., $\mathcal{N}(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$. This convergence is with respect to the data-generating distribution $\mathbb{P}_{\mathcal{S}}$. In particular, recall that $\tilde{\mathcal{S}}$ is random, drawn according to $\mathbb{P}_{\mathcal{S}}$. Since $\tilde{\mathcal{S}}$ is random, the posterior distribution $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\tilde{\mathcal{S}}}$ (which depends on $\tilde{\mathcal{S}}$) is also random. Although analyzing performance with respect to this random measure may be challenging, the theorem asserts that $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\tilde{\mathcal{S}}}$ converges to a known normal distribution, enabling simple asymptotic approximations.

The target normal distribution depends on the matrix $\mathcal{I}(\boldsymbol{\theta}^*)$. Explicit formulas for $\mathcal{I}(\boldsymbol{\theta})$ exist in terms of $\mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}}$. We will not need these formulas and, hence, omit them. We note, however, that if the affine dimension of Θ is less than d, $\mathcal{I}(\boldsymbol{\theta}^*)$ is singular, but the theorem is still valid for $\tilde{\boldsymbol{\theta}}_{1,r}$ (cf. Eq. (9)). Let $\mathcal{I}(\boldsymbol{\theta}^*)_{1,r} \in \mathbb{R}^{r \times r}$ denote the Fisher information matrix of $\boldsymbol{\theta}^*_{1,r}$ and define

$$\mathcal{I}(\boldsymbol{\theta}^*)^{-1} \equiv \begin{pmatrix} \mathcal{I}(\boldsymbol{\theta}^*)_{1,r} & \mathcal{I}(\boldsymbol{\theta}^*)_{1,r}\mathbf{A}^T \\ \mathbf{A}\mathcal{I}(\boldsymbol{\theta}^*)_{1,r} & \mathbf{A}\mathcal{I}(\boldsymbol{\theta}^*)_{1,r}\mathbf{A}^T \end{pmatrix}.$$

Then, the theorem is also valid as stated with $\tilde{\boldsymbol{\zeta}}$ having a degenerate normal distribution.

Thm. 6 is sometimes called the "Bayesian Central Limit Theorem." Like the traditional Central Limit Theorem, the requisite regularity conditions are very mild but are somewhat technical to state formally.[6] Under similar mild conditions, the posterior mean and covariance are consistent,

[6] In our setting, one set of sufficient conditions is that the map $\theta \mapsto \mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}}$ be differentiable in quadratic mean at $\boldsymbol{\theta}^*$ and that $\boldsymbol{\mu}_N$ be an asymptotically efficient estimator. For proof, see Thm. 10.1 of Van der Vaart (2000) and the discussion just preceding the theorem combined with Lemma 10.6 of the same work. See also Chen (1985), Van der Vaart (2000) for other sufficient conditions.

i.e., $\tilde{\boldsymbol{\mu}}_N \to_{\mathbb{P}_\mathcal{S}} \boldsymbol{\theta}^*$, $N\tilde{\boldsymbol{\Sigma}}_N \to_{\mathbb{P}_\mathcal{S}} \mathcal{I}(\boldsymbol{\theta}^*)^{-1}$ (see, e.g., Diaconis and Freedman (1986) for an even stronger result). These regularity conditions do not require i.i.d. data. For example, Thm. 6 applies to the auto-regressive, time-series model of Ex. EC.1 (Chatfield 2013). Thus, some authors, such as Gelman et al. (2014), advocate that unless the model is one of a few well-known pathological cases, it is reasonable to simply assume Thm. 6 and posterior mean consistency hold in practice rather explicitly validating the regularity conditions. To avoid unnecessary technicalities, we do the same:

ASSUMPTION 3. *The conclusion of Thm. 6 holds, and* $(\tilde{\boldsymbol{\mu}}_N, N\tilde{\boldsymbol{\Sigma}}_N) \to_{\mathbb{P}_\mathcal{S}} (\boldsymbol{\theta}^*, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$.

For each of Exs. 1 to EC.2, sufficient conditions for A3 to hold can be found in Geyer and Meeden (2013), McLachlan and Peel (2004), Chatfield (2013), Gelman et al. (2014), respectively. Finally, although A3 describes convergence in probability for the total variation distance and posterior mean, for many models both convergences, actually hold almost surely. See Geyer and Meeden (2013) for a proof in the case of Ex. 1.

4.1. Asymptotically Optimal and Near-Optimal Bayesian Ambiguity Sets

We next use A3 to characterize the asymptotics of ambiguity sets. Since Thm. 6 does not require i.i.d. data, our asymptotic results do not require independence, and since the limiting distribution in Thm. 6 does not depend on the specific choice of prior, our asymptotic results also do not depend on the specific choice of prior. Recall $\mathcal{P}^*(\cdot)$ as defined in Eq. (8).

THEOREM 7. *Assuming A3 and* $\boldsymbol{\theta}^* \in ri(\Theta)$, *as* $N \to \infty$,

$$\sup_{\mathbf{v}\in\mathbb{R}^d:\|\mathbf{v}\|=1} \sqrt{N} \left| VaR^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\tilde{\mathcal{S}}}(\mathbf{v}) - \mathbf{v}^T\tilde{\boldsymbol{\mu}}_N - z_{1-\epsilon}\|\mathbf{v}\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}} \right| \to_{\mathbb{P}_\mathcal{S}} 0, \tag{15}$$

where $z_{1-\epsilon}$ *is the* $1-\epsilon$ *quantile of a standard normal distribution. Consequently, for any* $0<\kappa<1$, *the* $\mathbb{P}_\mathcal{S}$*-probability of both of the following events tends to* 1 *as* $N \to \infty$*:*

1. $\mathcal{P}^*(\tilde{\mathcal{S}}, (1+\kappa)z_{1-\epsilon}/\sqrt{N})$ *satisfies the posterior feasibility guarantee at level* ϵ *for* $\mathcal{G}$.
2. $\mathcal{P}^*(\tilde{\mathcal{S}}, (1-\kappa)z_{1-\epsilon}/\sqrt{N})$ *is a subset of any other convex ambiguity set* $\mathcal{P}(\tilde{\mathcal{S}})$ *that satisfies the posterior feasibility guarantee at level* ϵ *for* $\mathcal{G}$.

In words, Thm. 7 asserts that as $N \to \infty$, $\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N})$ is essentially a Bayesian optimal set for $\mathcal{G}$. Any other set that satisfies the posterior feasibility guarantee for all $g \in \mathcal{G}$ eventually contains a small contraction of $\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N})$. Any small inflation of $\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N})$ eventually satisfies the posterior feasibility guarantee for all $g \in \mathcal{G}$. Observe that the theorem makes no claim for finite N; $\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N})$ generally will *not* satisfy the posterior feasibility guarantee for finite N. Nonetheless, we can use $\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N})$ as a benchmark to measure the relative size of other ambiguity sets that *do* satisfy the posterior guarantee for finite N.

Definition 4. We say $\mathcal{P}(\cdot)$ is *α-near-optimal* for $\mathcal{G}$ if there exists a non-random α (not depending on N or d, but perhaps depending on ϵ) such that as $N \to \infty$,

$$\mathbb{P}_{\mathcal{S}}\left(\mathcal{P}(\tilde{\mathcal{S}}) - \boldsymbol{\mu}_N \subseteq \alpha\Big(\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}) - \boldsymbol{\mu}_N\Big)\right) \to 1.$$

If no such α exists, we say $\mathcal{P}(\cdot)$ is not near-optimal for $\mathcal{G}$.

From Thm. 7, with $\mathbb{P}_{\mathcal{S}}$-probability tending to 1, an α-near-optimal set is asymptotically no more than α times larger than any other set that satisfies a posterior guarantee, justifying our terminology "near-optimal." We require that α not depend on d because in many applications, d can be large relative to N, causing sets with d dependence to also be large. Perhaps surprisingly, our general purpose ambiguity set is near-optimal.

Theorem 8. *The set* $\mathcal{P}^*\left(\cdot, \sqrt{\frac{1-\epsilon}{\epsilon N}}\right)$ *is* $\frac{\sqrt{1/\epsilon-1}}{z_{1-\epsilon}}$*-near-optimal for* $\mathcal{G}$*.*

Fig. 1 shows the constant $\frac{\sqrt{1/\epsilon-1}}{z_{1-\epsilon}}$ for some typical values of ϵ.

4.2. Near-Optimal Sets for Distributions with Finite, Discrete Support

Under Ex. 1, Thm. 8 proves that $\mathcal{P}^{\chi^2}\left(\cdot, \sqrt{\frac{1-\epsilon}{\epsilon(\tau_0+1)}}\right)$ is $\sqrt{\frac{1-\epsilon}{\epsilon z_{1-\epsilon}}}$-near-optimal for $\mathcal{G}$. We prove an analogous result for $\mathcal{P}^{KL}(\cdot)$.

Theorem 9. *Under setup of Ex. 1 with* $\boldsymbol{\theta}^* \in ri(\Theta)$*,* $\mathcal{P}^{KL}\left(\cdot, \sqrt{\log(1/\epsilon)/(\tau_0+2)}\right)$ *is* $\frac{\sqrt{2\log(1/\epsilon)}}{z_{1-\epsilon}}$*-near-optimal for* $\mathcal{G}$*.*

The radius of $\mathcal{P}^{KL}$ in Thm. 9 differs from that in Thm. 5 by the asymptotically negligible scaling $\sqrt{\frac{\tau_0+2}{\tau_0}} = 1 + O(N^{-1/2})$. Thus, we consider the sets to be comparable.

For comparison to Thm. 8, we include the constant of Thm. 9 in Fig. 1. Neither constant dominates the other: for $\epsilon < .219$, $\frac{\sqrt{2\log(1/\epsilon)}}{z_{1-\epsilon}} < \frac{\sqrt{1/\epsilon-1}}{z_{1-\epsilon}}$, and the reverse holds for larger ϵ.

	χ^2	KL	ϕ-Div : $\frac{\sqrt{\chi^2_{d,1-\epsilon}}}{z_{1-\epsilon}}$			
ϵ	$\frac{\sqrt{1/\epsilon-1}}{z_{1-\epsilon}}$	$\frac{\sqrt{2\log(1/\epsilon)}}{z_{1-\epsilon}}$	d=3	d=5	d=10	d=20
0.3	2.91	2.96	3.65	4.70	6.55	9.10
0.2	2.38	2.13	2.56	3.21	4.36	5.95
0.1	2.34	1.67	1.95	2.37	3.12	4.16
0.05	2.65	1.49	1.70	2.02	2.60	3.41
0.01	4.28	1.30	1.45	1.67	2.07	2.63
0.001	10.23	1.20	1.31	1.47	1.76	2.18

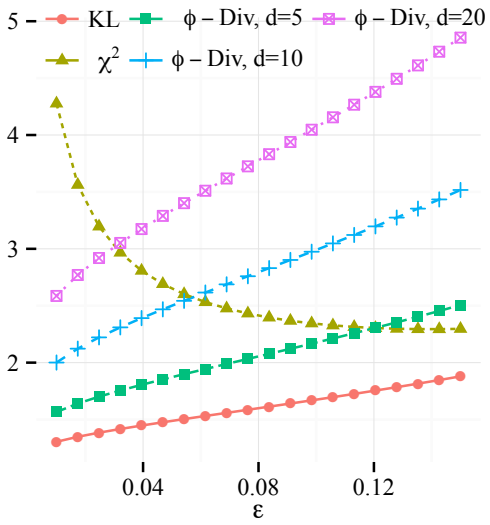

Figure 1 **The table shows the size of various ambiguity sets relative to the asymptotically Bayesian optimal set for $\mathcal{G}$. The graph plots these relative sizes for varying ϵ. Throughout, $\mathcal{P}^{KL}\left(\mathcal{S}, \sqrt{\log(1/\epsilon)/\tau_0}\right)$ is denoted "KL", $\mathcal{P}^{\chi^2}\left(\mathcal{S}, \sqrt{\frac{1-\epsilon}{\epsilon(\tau_0+1)}}\right)$ is denoted "χ^2" and $\mathcal{P}^{\phi}\left(\mathcal{S}, \sqrt{\frac{\phi''(1)\chi^2_{d-1,1-\epsilon}}{2N}}\right)$ is denoted "ϕ-Div" for $d=5,10,20$.**

4.3. Sub-optimality of Credible Regions for $\mathcal{G}$

A common approach to constructing ambiguity sets is to choose $\mathcal{P}(\mathcal{S})$ so that it contains the true, unknown distribution with high probability. In our Bayesian context, such sets are called *credible regions*, i.e., they satisfy $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\tilde{\boldsymbol{\theta}} \in \mathcal{P}(\mathcal{S})) \geq 1-\epsilon$ for any $\mathcal{S}$. (The frequentist analogue of a credible region is a confidence region and will be discussed in Sec. 5.) Credible regions satisfy the posterior feasibility guarantee for all measurable g, not just $g \in \mathcal{G}$, since $\mathbf{x} \in \mathcal{X}(\mathcal{P}(\mathcal{S}))$ and $\tilde{\boldsymbol{\theta}} \in \mathcal{P}(\mathcal{S})$ imply that $g(\tilde{\boldsymbol{\theta}}, \mathbf{x}) \leq 0$ for any g.

Despite the popularity of this approach, credible regions cannot be near-optimal for $\mathcal{G}$.

THEOREM 10. *Suppose $\mathcal{P}(\mathcal{S})$ is a credible region for all $\mathcal{S}$, $\boldsymbol{\theta}^* \in ri(\Theta)$. Let r be the affine dimension of Θ and fix $\alpha < \frac{\sqrt{\chi^2_{r,1-\epsilon}}}{z_{1-\epsilon}}$. Then, under A3, with $\mathbb{P}_{\mathcal{S}}$-probability tending to 1,*

$$\mathcal{P}(\tilde{S}) - \tilde{\boldsymbol{\mu}}_N \not\subseteq \alpha\left(\mathcal{P}^*(\tilde{S}, z_{1-\epsilon}/\sqrt{N}) - \tilde{\boldsymbol{\mu}}_N\right). \tag{16}$$

In particular, if r scales with d, $\mathcal{P}(\cdot)$ is not near-optimal for $\mathcal{G}$.

In Exs. 1 and 2, $r = d-1$, so credible regions cannot be near-optimal for $\mathcal{G}$. Practically, even for relatively small r, the constant in Thm. 10 can be large (c.f. Fig. 1), suggesting near-optimal variants may offer better performance. We confirm this intuition numerically in Sec. 6.

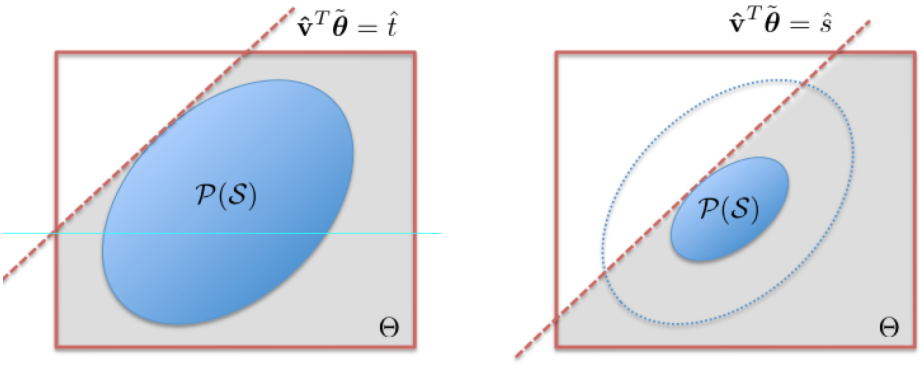

Figure 2 The key intuition behind Thm. 10.

Fig. 2 illustrates the key intuition behind Thm. 10 when $g(\boldsymbol{\theta}, (\mathbf{v}, t)) \equiv \mathbf{v}^T\boldsymbol{\theta} - t$ is a linear function. The left panel shows a credible region $\mathcal{P}(\mathcal{S})$ and a robust feasible pair $(\hat{\mathbf{v}}, \hat{t})$, i.e., $\sup_{\boldsymbol{\theta}\in\mathcal{P}(\mathcal{S})} \hat{\mathbf{v}}^T\boldsymbol{\theta} \leq \hat{t}$. The shaded region represents the sub-level set $\{\boldsymbol{\theta} \in \Theta : g(\boldsymbol{\theta}, (\hat{\mathbf{v}}, \hat{t})) \leq 0\}$, which contains $\mathcal{P}(\mathcal{S})$ and some additional volume. Consequently, $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(g(\tilde{\boldsymbol{\theta}}, (\hat{\mathbf{v}}, \hat{t})) \leq 0) > \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\tilde{\boldsymbol{\theta}} \in \mathcal{P}(\mathcal{S})) = 1 - \epsilon$, and this inequality can be very loose depending on how much mass lies in the shaded region outside $\mathcal{P}(\mathcal{S})$.

By contrast, the right panel shows a near-optimal set $\mathcal{P}(\mathcal{S})$, with robust feasible pair $(\hat{\mathbf{v}}, \hat{s})$. By construction (cf. Thm. 1), $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(g(\tilde{\boldsymbol{\theta}}, (\hat{\mathbf{v}}, \hat{s})) \leq 0) \geq 1 - \epsilon$. This probability consists of both $\mathcal{P}(\mathcal{S})$ *and* the shaded area outside $\mathcal{P}(\mathcal{S})$. Thus, $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\tilde{\boldsymbol{\theta}} \in \mathcal{P}(\mathcal{S})) < 1 - \epsilon$; we have a set that satisfies the posterior feasibility guarantee but is much smaller than a credible region. Thm. 10 asserts that this case is in fact typical and that, asymptotically, the ratio of sizes is $\Omega(\sqrt{r})$ in at least one direction.

We next specialize and strengthen Thm. 10 for some of our previous examples.

4.4. The Size of ϕ-Divergence Ambiguity Sets

A popular class of ambiguity sets for Ex. 1 is based upon ϕ-divergences (see Ben-Tal et al. (2013)). Given a function $\phi(t)$ such that $\phi(t)$ is convex for $t \geq 0$ and $\phi(1) = 0$, the ϕ-divergence between two vectors $\mathbf{p}, \mathbf{q}$ is $\sum_{i=1}^{d} q_i \phi\left(\frac{p_i}{q_i}\right)$. Thus, ϕ-divergences resemble distance metrics. Given a ϕ-divergence, consider the ambiguity set $\mathcal{P}^{\phi}(\mathcal{S}, \Gamma) \equiv \left\{\boldsymbol{\theta} \in \Delta_d : \sum_{i=1}^{d} \mu_{N,i} \phi\left(\frac{\theta_i}{\mu_{N,i}}\right) \leq \Gamma^2\right\}$. This set generalizes many other popular ambiguity sets. For example, with $\phi(t) = (t-1)^2$, $\mathcal{P}^{\phi}(\mathcal{S}, \Gamma) \equiv \mathcal{P}^{\chi^2}(\mathcal{S}, \Gamma)$ and when $\phi(t) = t \log t - t + 1$, $\mathcal{P}^{\phi}(\mathcal{S}, \Gamma) \equiv \mathcal{P}^{KL}(\mathcal{S}, \Gamma)$.

Ben-Tal et al. (2013) observes that if $\phi''(t)$ exists in a neighborhood of 1, $\mathbb{P}_{\mathcal{S}}\left(\boldsymbol{\theta}^* \in \mathcal{P}^{\phi}\left(\tilde{\mathcal{S}}, \sqrt{\frac{\phi''(1)\chi^2_{d-1,1-\epsilon}}{2N}}\right)\right) \to 1-\epsilon$.[7] Ben-Tal et al. (2013) also proves that ϕ-divergence sets are tractable. These two features have made ϕ-divergence sets with this radius very popular.

One can show that ϕ-divergence sets with this radius are asymptotically Bayesian credible regions. By Thm. 10, there exist directions in which the set is $\Omega(\sqrt{r})$ larger than a Bayesian optimal set. We prove a stronger result; ϕ-divergence sets are $\Omega(\sqrt{r})$ larger than a Bayesian optimal set in *all* directions simultaneously:

THEOREM 11. *Under the setup of Ex. 1, suppose $\phi''(t)$ exists in a neighborhood of 1. Fix any $\alpha < \frac{\sqrt{\chi^2_{d-1,1-\epsilon}}}{z_{1-\epsilon}}$. Then, for N sufficiently large,*

$$\alpha\big(\mathcal{P}^*(\mathcal{S}, z_{1-\epsilon}/\sqrt{N}) - \boldsymbol{\mu}_N\big) \subseteq \mathcal{P}^{\phi}\left(\mathcal{S}, \sqrt{\frac{\phi''(1)\chi^2_{d-1,1-\epsilon}}{2N}}\right) - \boldsymbol{\mu}_N, \quad \forall \mathcal{S}.$$

In other words, $\mathcal{P}^{\phi}\left(\cdot, \sqrt{\frac{\phi''(1)\chi^2_{d-1,1-\epsilon}}{2N}}\right)$ is not Bayesian near-optimal for $\mathcal{G}$.

4.5. Sub-Optimality of the Ambiguity Sets of Zhu and Fukushima (2009), Zhu et al. (2014).

Zhu and Fukushima (2009) and Zhu et al. (2014) both propose ambiguity sets for Ex. 2 in slightly different applications. Zhu and Fukushima (2009) considers a non-data-driven setting and proposes using $\mathcal{P} = \Delta_d$ to bound worst-case conditional value at risk for portfolio optimization problems. In the absence of any data or probabilistic assumptions, this set is the only ambiguity set that offers a feasibility guarantee. When data is available, it is very large relative to the Bayesian optimal set:

THEOREM 12. *Under the setup of Ex. 2, for any $\mathcal{S}$,*

$$\left(\frac{\sqrt{\tau_0+1}}{z_{1-\epsilon}} \min_i \sqrt{\mu_{N,i}}\right) \left(\mathcal{P}^*(\mathcal{S}, z_{1-\epsilon}/\sqrt{N}) - \boldsymbol{\mu}_N\right) \subseteq (\Delta_d - \boldsymbol{\mu}_N)$$

Moreover, under A3, if $\boldsymbol{\theta}^* \in ri(\Theta)$, Δ_d is not near-optimal for $\mathcal{G}$.

By contrast, Zhu et al. (2014) proposes the set $\mathcal{P}^{\chi^2}(\mathcal{S}, \sqrt{\chi^2_{d,1-\epsilon}/N})$ and argues that it is asymptotically a credible region under some regularity conditions on the mixture components. A sub-optimality bound for $\mathcal{P}^{\chi^2}(\mathcal{S}, \sqrt{\chi^2_{d,1-\epsilon}/N})$ follows directly from Thm. 11. Indeed, the proof of Thm. 11 does not utilize the support of $\tilde{\boldsymbol{\xi}}$. Consequently, it also readily applies to Ex. 2.

[7] More precisely, the authors observe this when $\boldsymbol{\tau}' = \mathbf{0}$, but the asymptotics are the same for other choices of $\boldsymbol{\tau}'$.

4.6. Consistency of Optimal Solutions

Thus far, we have focused on the geometry of ambiguity sets. We next investigate the asymptotic properties of solutions to DRO problems. Our results are closely related to those in Bertsimas et al. (2017b) but neither imply nor are implied by those results. Indeed, we treat multiple uncertain constraints that are concave in a finite dimensional $\boldsymbol{\theta}^*$, while Bertsimas et al. (2017b) focuses on an uncertain linear objective in a potentially infinite dimensional parameter.

To be concrete, consider the following (full-information) optimization problem:

$$\begin{aligned} \mathbf{P}: \quad z^* = \min_{\mathbf{x}\in\mathcal{C}} \quad & g_0(\boldsymbol{\theta}^*, \mathbf{x}) \\ \text{s.t.} \quad & g_l(\boldsymbol{\theta}^*, \mathbf{x}) \leq 0, \ l = 1, \ldots, L, \end{aligned} \tag{17}$$

where $g_l \in \mathcal{G}$ for $l = 0, \ldots, L$, and $\mathcal{C}$ is compact, not depending on $\boldsymbol{\theta}^*$. Its robust counterpart is

$$\begin{aligned} \mathbf{P_N}: \quad \tilde{z}_N = \min_{\mathbf{x}\in\mathcal{C}} \quad & \max_{\boldsymbol{\theta}\in\mathcal{P}(\tilde{\mathcal{S}})} g_0(\boldsymbol{\theta}, \mathbf{x}) \\ \text{s.t.} \quad & g_l(\boldsymbol{\theta}, \mathbf{x}) \leq 0, \ \forall \boldsymbol{\theta} \in \mathcal{P}(\tilde{\mathcal{S}}), \ \ l = 1, \ldots, L, \end{aligned} \tag{18}$$

where $\mathcal{P}(\tilde{\mathcal{S}})$ is a non-empty, convex, compact ambiguity set. Let $\mathcal{O}^*$, $\tilde{\mathcal{O}}_N$ denote the set of optimal solutions to each problem. We write $\tilde{\mathcal{O}}_N$ instead of $\mathcal{O}_N$ to emphasize the randomness of $\tilde{\mathcal{S}}$.

Following Shapiro and Ruszczyński (2003), define the deviation between two sets by $d(\mathcal{A}, \mathcal{B}) \equiv \sup_{\mathbf{x}\in\mathcal{A}} \inf_{\mathbf{y}\in\mathcal{B}} \|\mathbf{x} - \mathbf{y}\|$. Recall that g is equicontinuous in $\boldsymbol{\theta}$ over $\mathbf{x} \in \mathcal{C}$ if, for any $\boldsymbol{\theta}_0 \in \Theta$ and $\epsilon > 0$, there exists a δ such that for any $\boldsymbol{\theta} \in \Theta$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta$, $\sup_{\mathbf{x}\in\mathcal{C}} |g(\boldsymbol{\theta}, \mathbf{x}) - g(\boldsymbol{\theta}_0, \mathbf{x})| \leq \epsilon$. Finally, let $cl(\mathcal{A})$ denote the closure of $\mathcal{A}$. The next theorem proves that $\tilde{\mathcal{O}}_N$ "converges" to a subset of $\mathcal{O}^*$.

THEOREM 13. *Suppose*

i) g_l *is equicontinuous in* $\boldsymbol{\theta}$ *over* $\mathbf{x} \in \mathcal{C}$ *and continuous in* $\mathbf{x}$ *for every* $\boldsymbol{\theta} \in \Theta$.

ii) There exists $\alpha_N = o(N^{1/2})$ *such that* $\mathcal{P}(\tilde{\mathcal{S}}) - \tilde{\boldsymbol{\mu}}_N \subseteq \alpha_N \left(\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}) - \tilde{\boldsymbol{\mu}}_N \right)$, $\mathbb{P}_{\mathcal{S}}$*-a.s.*

iii) $\{\mathbf{x} \in \mathcal{C} : g_l(\boldsymbol{\theta}^*, \mathbf{x}) \leq 0, \ l = 1, \ldots, L\} = cl(\{\mathbf{x} \in \mathcal{C} : g_l(\boldsymbol{\theta}^*, \mathbf{x}) < 0, \ l = 1, \ldots, L\})$.

iv) $(\tilde{\boldsymbol{\mu}}_N, N\tilde{\boldsymbol{\Sigma}}_N) \to (\boldsymbol{\theta}^*, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$, $\mathbb{P}_{\mathcal{S}}$*-a.s.*

Then, under A3, $\tilde{z}_N \to z^*$ *and* $d(\tilde{\mathcal{O}}_N, \mathcal{O}^*) \to 0$, $\mathbb{P}_{\mathcal{S}}$*-a.s.*

Condition ii) is strictly weaker than requiring $\mathcal{P}(\tilde{\mathcal{S}})$ be α-near-optimal for $\mathcal{G}$. In particular, near-optimal sets satisfy Condition ii) with $\alpha_N = O(1)$. Moreover, Condition iii) is very mild and is satisfied, e.g., if $g_l(\theta^*, \mathbf{x})$ is convex in $\mathbf{x}$ for $l = 1, \ldots, L$, and there exists a Slater point. Finally, as mentioned, most models of interest, such as Exs. 1 and EC.1, satisfy Condition iv).

An important consequence of Thm. 13 is that if Eq. (17) admits a unique optimal solution, any sequence of optimal solutions to Eq. (18) converges almost surely to this solution. We leverage this property to relate Bayesian and frequentist ambiguity sets in the next section.

5. Comparing Bayesian and Frequentist Ambiguity Sets

We next contrast sets that satisfy Def. 1 and Def. 2.

5.1. Frequentist Feasibility and Confidence Regions

Bertsimas et al. (2017b) observe that many frequentist proposals for $\mathcal{P}(\cdot)$ are confidence regions, i.e., they satisfy

$$\mathbb{P}_{\mathcal{S}}(\boldsymbol{\theta}^* \in \mathcal{P}(\tilde{\mathcal{S}})) \geq 1 - \epsilon. \tag{19}$$

If $\mathcal{P}(\cdot)$ is a confidence region, then it satisfies the frequentist guarantee at level ϵ for all measurable g since for any $\mathcal{S}$, $\mathbf{x} \in \mathcal{X}(\mathcal{P}(\mathcal{S}))$ and $\boldsymbol{\theta}^* \in \mathcal{P}(\mathcal{S})$ implies $g(\boldsymbol{\theta}^*, \mathbf{x}) \leq 0$. We improve this result.

THEOREM 14. *Suppose $\mathcal{P}(\mathcal{S})$ is closed and convex for any $\mathcal{S}$. Then, $\mathcal{P}(\cdot)$ is a confidence region if, and only if, it satisfies the frequentist guarantee at level ϵ for the function $g(\boldsymbol{\theta}, (\mathbf{v}, t)) = \mathbf{v}^T\boldsymbol{\theta} - t$.*

Since the given function is in $\mathcal{G}$, any ambiguity set that satisfies the frequentist guarantee for all $g \in \mathcal{G}$ is a confidence region and automatically satisfies the frequentist guarantee for the larger class of all measurable g. Thus, loosely speaking, sets that satisfy the frequentist guarantee for all $g \in \mathcal{G}$ offer "more protection" than those that satisfy a posterior guarantee for all $g \in \mathcal{G}$.

This protection comes at a cost. Confidence regions are comparably large to Bayesian credible regions and are typically not near-optimal.

We prove this claim using the maximum likelihood estimator: $\tilde{\boldsymbol{\theta}}^{MLE} \in \arg\max_{\boldsymbol{\theta}\in\Theta} \log \mathbb{P}_{\mathcal{S}}(\tilde{\mathcal{S}})$. Under mild regularity conditions,[8]

$$\sqrt{N}(\tilde{\boldsymbol{\theta}}^{MLE} - \boldsymbol{\theta}^*) \to_d \mathcal{N}\left(0, \mathcal{I}(\boldsymbol{\theta}^*)^{-1}\right), \tag{20}$$

where $\to_d$ denotes convergence in distribution. Note the similarity between Eq. (20) and Thm. 6.

THEOREM 15. *Suppose that both Eq.* (20) *and A3 hold,* $\mathcal{P}(\cdot)$ *is a confidence region, and* $\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} \to_{\mathbb{P}_{\mathcal{S}}} 0$. *Let* r *be the affine dimension of* Θ, *and fix any* $0 < \alpha < \sqrt{\chi^2_{r,1-\epsilon}}/z_{1-\epsilon}$. *Then,*

$$\mathbb{P}_{\mathcal{S}}\left(\mathcal{P}(\tilde{\mathcal{S}}) - \tilde{\boldsymbol{\mu}}_N \not\subseteq \alpha(\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}) - \tilde{\boldsymbol{\mu}}_N)\right) > 0.$$

for all N *sufficiently large. In particular, if* r *scales with* d, $\mathcal{P}(\cdot)$ *is not near-optimal for* $\mathcal{G}$.

Many models satisfy the assumption on $\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N}$ in the theorem. For example, in Ex. 1, a direct computation yields $\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} = O_{\mathbb{P}_{\mathcal{S}}}(N^{-1/2})$. Hence, in this setting, confidence regions cannot be near-optimal.

Thm. 15 is analogous to Thm. 10; both establish that sets which protect against all measurable functions are not near-optimal for $\mathcal{G}$ and are comparably sized. Said another way, when considering all measurable functions, existing frequentist ambiguity sets are essentially the smallest possible (in our Bayesian framework). Thus, Thms. 10 and 15 also partially explain the size advantage of Bayesian near-optimal sets over confidence regions; unlike confidence regions, near-optimal sets only protect against $g \in \mathcal{G}$ and, thus, can be smaller.

5.2. Critical Role of Concavity

The class of functions $\mathcal{G}$ is such that there exist Bayesian ambiguity sets that satisfy the posterior feasibility guarantee for all $g \in \mathcal{G}$ that are smaller than credible regions. We next prove that any class of functions with this property cannot contain a specific convex function, defined below. This theorem highlights the critical role of concavity to constructing Bayesian ambiguity sets smaller than credible regions.

[8] A set of possible sufficient conditions is that $\boldsymbol{\theta} \to \mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}}$ is differentiable in quadratic mean at $\boldsymbol{\theta}^*$, $\boldsymbol{\theta}^* \in ri(\Theta)$, $\tilde{\boldsymbol{\theta}}^{MLE}(\tilde{\mathcal{S}}) \to_{\mathbb{P}_{\mathcal{S}}} \boldsymbol{\theta}^*$ and that there exists a measurable function $\dot{\ell}$ with $\mathbb{E}_{\theta^*}[\dot{\ell}^2] < \infty$ such that for every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ in a neighborhood of $\boldsymbol{\theta}^*$, $|\log d\mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}_1}(\xi) - \log d\mathbb{P}_{S|\boldsymbol{\theta}_2}(\xi)| \le \dot{\ell}(\xi)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$ (Van der Vaart 2000, Chapt 5.5).

THEOREM 16. *Suppose $\mathcal{P}(\cdot)$ satisfies a posterior feasibility guarantee for the function $g(\boldsymbol{\theta}, \mathbf{x}) = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathcal{I}(\boldsymbol{\theta}^*)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - x^2$, that A3 holds, and that $\sqrt{N}\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\theta}^*\| \to_{\mathbb{P}_\mathcal{S}} 0$. Let r be the affine dimension of Θ, and fix any $\alpha < \frac{\sqrt{\chi^2_{r,1-\epsilon}}}{z_{1-\epsilon}}$. Then, with $\mathbb{P}_\mathcal{S}$-probability tending to 1, $\mathcal{P}(\tilde{\mathcal{S}}) - \tilde{\boldsymbol{\mu}}_N \not\subseteq \alpha\left(\mathcal{P}^*\left(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}\right) - \tilde{\boldsymbol{\mu}}_N\right)$. In other words, $\mathcal{P}(\cdot)$ is not near-optimal.*

Thm. 16 is analogous to Thm. 14. Both theorems identify a particular test function such that any ambiguity set that protects against that test function is necessarily large.

5.3. Robustness to Solution Method

Frequentist ambiguity sets also enjoy an added degree of robustness over Bayesian near-optimal sets with respect to the algorithm used to compute an optimal solution.[9] Specifically, consider fixing an algorithm for solving $\mathbf{P_N}$ (cf. Eq. (18)). Given $\tilde{\mathcal{S}}$, this algorithm returns an optimal solution $\tilde{\mathbf{x}}$ of $\mathbf{P_N}$, so that $\tilde{\mathbf{x}} \in \mathcal{X}(\mathcal{P}(\tilde{\mathcal{S}})$, $\mathbb{P}_\mathcal{S}$-a.s. The solution $\tilde{\mathbf{x}}$ depends on $\tilde{\mathcal{S}}$, but may also depend on other sources of randomness, e.g., if the algorithm is randomized or leverages additional data beyond $\tilde{\mathcal{S}}$.

Given our algorithm, a natural frequentist guarantee we might seek for the l^{th} constraint is

$$\mathbb{P}\left(g_l(\boldsymbol{\theta}^*, \tilde{\mathbf{x}}) \leq 0\right) \geq 1 - \epsilon, \quad \forall \boldsymbol{\theta}^* \in \Theta, \tag{21}$$

which guarantees feasibility of $\tilde{\mathbf{x}}$ with high probability across multiple draws of the data and any additional randomness.[10] Notice that if $\mathcal{P}(\cdot)$ satisfies the frequentist feasibility guarantee in Def. 1, then Eq. (21) holds for *any* $\tilde{\mathbf{x}}$ since $\tilde{\mathbf{x}} \in \mathcal{X}(\mathcal{P}(\tilde{\mathcal{S}}))$ almost surely.

Now consider the Bayesian perspective. The Bayesian analogue of Eq. (21) is

$$\mathbb{P}\left(g_l(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{x}}) \leq 0\right) \geq 1 - \epsilon, \tag{22}$$

which guarantees feasibility of $\tilde{\mathbf{x}}$ with high probability across multiple draws of the data, $\tilde{\boldsymbol{\theta}}$, and any other randomness in the algorithm. Unlike the frequentist case, however, $\mathcal{P}(\cdot)$ satisfying the posterior feasibility guarantee in Def. 2 does not imply that Eq. (22) is satisfied for *any* $\tilde{\mathbf{x}}$. Indeed,

[9] We thank the anonymous Associate Editor for his or her insightful comments and questions, which inspired the results in this subsection.

[10] Since the probability is with respect to all sources of randomness, we drop the subscript on $\mathbb{P}$.

if $\tilde{\mathbf{x}}$ depends on information about $\tilde{\boldsymbol{\theta}}$ not present in $\tilde{\mathcal{S}}$, Eq. (22) may not hold. The key requirement is that $\tilde{\mathbf{x}}$ be conditionally independent of $\tilde{\boldsymbol{\theta}}$ given $\tilde{\mathcal{S}}$.

THEOREM 17. *Suppose $\mathcal{P}(\cdot)$ satisfies a posterior feasibility guarantee for all $g \in \mathcal{G}$, and let $\tilde{\mathbf{x}}$ be a solution $\mathbf{P_N}$ (cf. Eq. (18)) given data $\tilde{\mathcal{S}}$. Then, if $\tilde{\mathbf{x}} \perp\!\!\!\perp \tilde{\boldsymbol{\theta}} \mid \tilde{\mathcal{S}}$, then Eq. (22) holds. Furthermore, there exist instances of $\mathbf{P_N}$ with $\mathcal{P}(\cdot)$ that are near-optimal for $\mathcal{G}$ and solutions $\tilde{\mathbf{x}}$ such that $\tilde{\mathbf{x}} \not\perp\!\!\!\perp \tilde{\boldsymbol{\theta}} \mid \tilde{\mathcal{S}}$ such that Eq. (22) does not hold.*

Conditional independence of $\tilde{\mathbf{x}}$ and $\tilde{\boldsymbol{\theta}}$ often holds in practice. For example, if $\tilde{\mathcal{S}}$ is the sole source of randomness in $\tilde{\mathbf{x}}$, then $\tilde{\mathbf{x}}$ is conditionally constant and, hence, conditionally independent of $\tilde{\boldsymbol{\theta}}$. Similarly, if solutions to $\mathbf{P_N}$ are unique $\mathbb{P}_{\mathcal{S}}$-a.s., $\tilde{\mathbf{x}}$ is constant given $\tilde{\mathcal{S}}$, and, hence, conditionally independent of $\tilde{\boldsymbol{\theta}}$. A case where conditional independence will not hold may be when $\mathbf{P_N}$ admits multiple optima and an independent hold-out data set $\mathcal{S}'$ whose distribution depends on the realization of $\tilde{\boldsymbol{\theta}}$ is used to choose between optima.

In any case, Thm. 17 also partially explains the size difference between confidence regions and our Bayesian near-optimal sets; unlike Bayesian near-optimal sets, frequentist confidence regions are completely robust to the choice of solution method, and, hence, necessarily larger.

5.4. Frequentist Properties of Solutions with Bayesian Ambiguity Sets

In summary, the frequentist feasibility guarantee is arguably a stronger property than the posterior feasibility guarantee; it guarantees feasibility for all measurable functions g for any solution method and, hence, requires larger sets. If, however, one is only interested in the (frequentist) guarantee Eq. (21), for a specific instance of Problem $\mathbf{P}$ (c.f. Eq. (17)), we argue that this additional strength may be unnecessary for large N. Indeed, near-optimal Bayesian ambiguity sets can sometimes asymptotically achieve Eq. (21), the desired *frequentist* outcome, despite their smaller size. We study this claim empirically in Sec. 6 and prove it formally for a special case:

THEOREM 18. *Suppose that*

i) $\mathcal{P}(\cdot)$ satisfies a posterior feasibility guarantee for $\mathcal{G}$.

ii) A3 and Eq. (20) hold.

iii) $\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} \to_{\mathbb{P}_{\mathcal{S}}} 0$.

iv) The assumptions of Thm. 13 hold.

v) Problem $\mathbf{P}$ *has a unique optimal solution.*

vi) Each constraint $g_l(\boldsymbol{\theta}, \mathbf{x}) \leq 0$, $l = 0, \ldots, L$, *in* $\mathbf{P}$ *is linear in* $\boldsymbol{\theta}$.

Let $\tilde{\mathbf{x}}_N$ *be a robust optimal solution to* $\tilde{\mathbf{P}}_{\mathbf{N}}$, *where we have suppressed the dependence of* $\tilde{\mathbf{x}}_N$ *on* $\tilde{\mathcal{S}}$. *Then,* $\limsup \mathbb{P}_{\mathcal{S}}(g_l(\boldsymbol{\theta}^*, \tilde{\mathbf{x}}_N) \geq 0) \leq \epsilon$, *for* $l = 1, \ldots, L$.

Comparing to Eq. (21), the theorem asserts that DRO problems with Bayesian ambiguity sets that satisfy a posterior feasibility guarantee yield solutions that asymptotically achieve Eq. (21). The theorem assumes the full-information problem $\mathbf{P}$ has a unique solution; this assumption ensures that asymptotically $\tilde{\mathbf{x}}_N$ is conditionally independent of $\tilde{\boldsymbol{\theta}}$ given $\tilde{\mathcal{S}}$.

The remaining regularity conditions in the theorem are not as restrictive as they perhaps seem. Previously, we argued that Conditions ii), iii), and iv) are satisfied by many statistical models. Condition vi) is seemingly most stringent. However, in applications such as Exs. 1 and 2, all expectation and chance constraints are linear and, hence, satisfy the assumption. Outside these particular settings, for any $g \in \mathcal{G}$,

$$g(\boldsymbol{\theta}^*, \mathbf{x}) \leq 0 \iff \inf_{\mathbf{v}} \mathbf{v}^T \boldsymbol{\theta}^* - g_*(\mathbf{v}, \mathbf{x}) \leq 0 \iff \exists (\mathbf{v}, t) \text{ s.t. } \mathbf{v}^T \boldsymbol{\theta}^* \leq t,\ t \geq g_*(\mathbf{v}, \mathbf{x}).$$

Thus, given an instance of $\mathbf{P}$, by i) rewriting the objective epigraphically, ii) introducing auxiliary variables $(\mathbf{v}_l, t_l)$ for $l = 0, \ldots, L$, iii) replacing each constraint with $\mathbf{v}_l^T \boldsymbol{\theta}^* \leq t_l$ and iv) augmenting $\mathcal{C}$ with the new constraint $t_l \leq g_{l*}(\mathbf{v}_l, \mathbf{x})$, we *almost* obtain an instance of the requisite form. We write "almost" because we must verify that we can restrict the new auxiliary variables $(\mathbf{v}_l, t_l)$ to a compact set. This restriction can often be argued via ad hoc bounds on the subgradients of g_l.

Overall, we believe Thm. 18 to be a compelling argument to consider Bayesian ambiguity sets as alternatives to frequentist sets for $g \in \mathcal{G}$, especially for moderate to large N. We propose some general guidelines for practitioners choosing among ambiguity sets in Appendix C.

6. Computational Experiments

We now present numerical experiments based on synthetic and real data.[11] We are interested in the following questions: Do DRO solutions using Bayesian near-optimal sets exhibit good *frequentist* properties for finite N? Does our theoretical analysis of size yield useful insight into the performance of DRO solutions? How sensitive are our results to misspecification of the Bayesian model?

We focus on portfolio allocation. Portfolio allocation has been widely studied in the data-driven DRO literature (see, e.g., Delage and Ye (2010), Postek et al. (2016), Wozabal (2014), Bertsimas et al. (2017b), Rujeerapaiboon et al. (2016)) because it is well known that methods that neglect ambiguity in $\boldsymbol{\theta}^*$ can perform poorly (see, e.g., Michaud (1989), DeMiguel et al. (2009a,b), Lim et al. (2011), El Karoui et al. (2011)). Specifically, we consider the nominal optimization problem

$$\max_{\mathbf{x}\in\mathbb{R}^n_+:\mathbf{e}^T\mathbf{x}\le 1} \quad \left\{\mathbb{E}^{\mathbb{P}_{\boldsymbol{\theta}^*}}[\mathbf{x}^T\tilde{\boldsymbol{\xi}}] \;:\; \mathrm{CVaR}_\epsilon^{\mathbb{P}_{\boldsymbol{\theta}^*}}(\mathbf{x}^T\tilde{\boldsymbol{\xi}}) \le \Gamma\right\} \tag{23}$$

and its robust counterpart

$$\begin{aligned}\max_{\mathbf{x}\in\mathbb{R}^n_+:\mathbf{e}^T\mathbf{x}\le 1} \min_{\boldsymbol{\theta}\in\mathcal{P}(\mathcal{S})} \quad & \mathbb{E}^{\mathbb{P}_{\boldsymbol{\theta}}}[\mathbf{x}^T\tilde{\boldsymbol{\xi}}] \\ \text{s.t.} \quad & \mathrm{CVaR}_\epsilon^{\mathbb{P}_{\boldsymbol{\theta}}}(\mathbf{x}^T\tilde{\boldsymbol{\xi}}) \le \Gamma, \quad \forall \boldsymbol{\theta}\in\mathcal{P}(\mathcal{S})\end{aligned} \tag{24}$$

for various ambiguity sets. We fix $\epsilon = 10\%$ throughout.

To facilitate comparisons to existing methods, we adopt the setup of Ex. 1. In particular, we consider $\mathcal{P}^{KL}(\mathcal{S}, \sqrt{\log(1/\epsilon)/\tau_0})$ (denoted "KL"), $\mathcal{P}^{KL}(\mathcal{S}, \sqrt{\chi^2_{d,1-\epsilon}/(2N)})$ (denoted "KL$_C$"), $\mathcal{P}^{\chi^2}(\mathcal{S}, \sqrt{\frac{1-\epsilon}{\epsilon(\tau_0+1)}})$ (denoted "χ^2"), and $\mathcal{P}^{\chi^2}(\mathcal{S}, \sqrt{\chi^2_{d,1-\epsilon}/N})$ (denoted "χ^2_C"). In each case, the subscript C indicates the confidence-region variant of the set, instead of the Bayesian near-optimal one. Unless otherwise specified, we adopt the uninformative prior $\tau' = \mathbf{e}$. For comparison, we also consider three non-DRO approaches to portfolio allocation:

- The sample average approximation (SAA) of Eq. (23), which replaces $\mathbb{P}_{\boldsymbol{\theta}^*}$ with the empirical distribution of the data: $\hat{\mathbb{P}}(\mathcal{A}) = \frac{1}{N}\sum_{j=1}^N \mathbb{I}(\hat{\boldsymbol{\xi}}^j \in \mathcal{A})$ for all measurable sets $\mathcal{A}$.

[11] Julia code for running each of the following experiments is available at `https://github.com/vgupta1/AmbiguitySets`.

Table 1 Summary statistics for individual industry portfolios.

	Dec. 2008 - Dec. 2014			Mar. 1998 - Dec 2014		
	Mean	Std	CVaR	Mean	Std	CVaR
Business Equipment	1.74	4.74	6.78	0.72	7.71	13.55
Chemicals	1.38	4.41	6.65	0.77	4.26	7.54
Consumer Durables	2.29	8.80	11.21	0.62	7.82	12.65
Energy	0.78	5.62	10.10	0.98	5.93	9.50
Healthcare	1.68	3.80	4.96	0.72	4.00	6.92
Manufacturing	1.73	6.07	9.51	0.98	5.82	10.01
Finance	1.38	6.33	11.17	0.56	5.78	10.37
Consumer Non-Durables	1.48	3.41	4.83	0.77	3.54	6.26
Other	1.53	5.48	8.83	0.54	5.21	9.60
Wholesale/Retail	1.70	3.99	5.47	0.77	4.52	7.75
Telecom	1.70	4.22	6.17	0.44	5.59	10.05
Utilities	1.14	3.57	6.06	0.81	4.36	7.58

- The "naive" diversification portfolio (Naive), which invests $1/d$ in each asset. DeMiguel et al. (2009b), Wozabal (2014) shows that this portfolio performs surprisingly well and enjoys strong robustness properties. Since this portfolio is typically infeasible in Eq. (23) for reasonable values of Γ, we implement $\mathbf{x}^{naive} = \mathbf{e}\min(1/d, \Gamma/\mathrm{CVaR}_{\epsilon}^{\hat{\mathbb{P}}}(\mathbf{e}^T\boldsymbol{\xi}))$ in what follows, i.e., we scale down the Naive portfolio if necessary to make it feasible for the empirical distribution.
- The minimum-variance portfolio (MinVar), which solves $\min_{\mathbf{x}\in\mathbb{R}^n_+:\mathbf{e}^T\mathbf{x}=1} \mathrm{Var}^{\hat{\mathbb{P}}}(\mathbf{x}^T\tilde{\boldsymbol{\xi}})$, where $\mathrm{Var}^{\hat{\mathbb{P}}}(\cdot)$ denotes the empirical variance (see DeMiguel et al. (2009b)). Since this portfolio is also typically infeasible in Eq. (23), we scale it similarly to the Naive portfolio.

Our data are based upon the historical returns of 12 industry portfolios available from French (2015). These 12 portfolios can be seen as proxies for index funds, and we will refer to them loosely as indices. Table 1 provides some summary statistics for each index over the two time periods most relevant for our analysis. We remark that the covariance matrix for these 12 indices is approximately low-rank; the first eigenvalue accounts for 63% of the total eigenspectrum. The first three eigenvalues account for approximately 80%. These features are typical of financial data.

Before presenting the details of our experiments, we summarize our main findings:

- Portfolios constructed from our Bayesian near-optimal sets are feasible with frequentist probability approximately $1-\epsilon$ in this application. The approximation error shrinks rapidly as N grows large and is negligibly small for moderate N.

- As predicted, sets with smaller asymptotic size tend to yield better optimization solutions. In particular, our Bayesian near-optimal sets significantly outperform their confidence-region variants in this application for both synthetic and real data, with similar robustness properties.
- These features do not strongly depend on the choice of prior. Specifically, for small N, priors that assign a small probability mass to the true distribution but a large probability mass to an incorrect distribution may yield sets with poor frequentist performance. However, in this particular application, we find that the strength in such a prior belief must be very large before the loss in performance makes traditional variants a preferable choice. Moreover, the loss in performance is attenuated in N. For large N, most priors yield sets with good performance.

6.1. Dependence on N

We first study the dependence on N under frequentist assumptions with synthetic data. Specifically, we take the true distribution to be uniformly distributed on the 72 points described by the monthly returns of our indices from Dec. 2008 to Dec. 2014. Then, for varying N, we simulate N data points from this distribution, use these data to construct one of our ambiguity sets, and solve Eq. (24) with $\Gamma = 3\%$. We repeat this procedure 1000 times, each time using the true distribution to compute the true expected return and CVaR of a portfolio. Notice that this repeated sampling setup accords precisely with the frequentist viewpoint; $\boldsymbol{\theta}^*$ is *fixed*, but the data $\tilde{\mathcal{S}}$ change between simulations.

Fig. 3 displays the expected returns and CVaRs over repeated random draws of the data along with error bars at the 90% and 10% quantiles. We draw attention to several features:

- The SAA, Naive, and MinVar portfolios frequently incur more than allocated 3% risk. Indeed, even for very large N, these portfolios exceed the threshold approximately 50% of the time. For smaller N, the returns are also highly unstable, i.e., the error bars are very large. These are well-documented drawbacks of SAA (see, e.g., Bertsimas et al. (2017b)) and are inherited by the Naive and MinVar portfolios because of the need to respect the constraint in Eq. (23).
- By contrast, the data-driven DRO models with the confidence-region-based ambiguity sets safely maintain a risk below 3% but are very conservative. The very large error bars for N

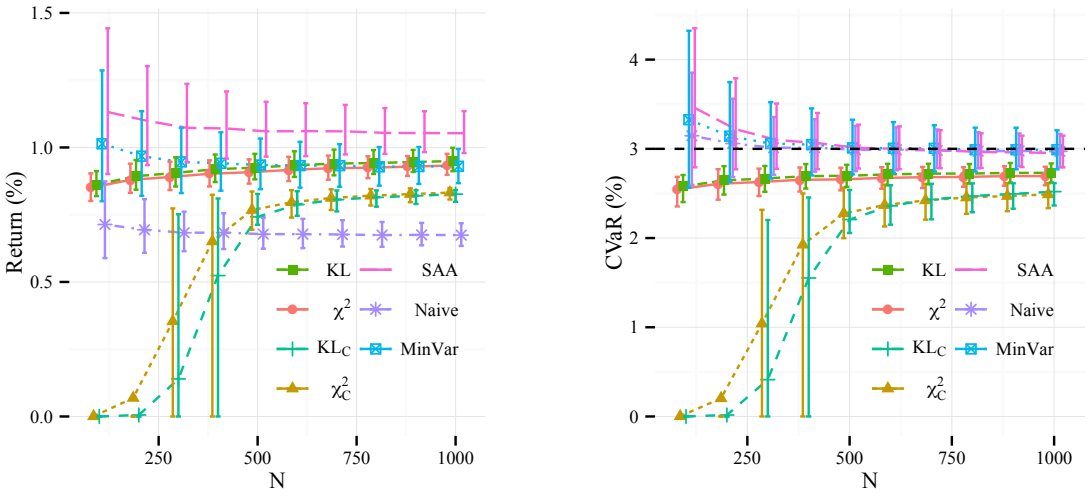

Figure 3 **Return and risk for increasing data N from Sec. 6.1.**

near 250 occur because, for some data realizations, the only portfolio that the model can safely guarantee will be feasible is $\mathbf{x} = \mathbf{0}$, i.e., to not invest. Table EC.1 in Appendix E.2 details the percentage of runs that return the zero portfolio and the standard deviation of $\|\mathbf{x}\|$ for each method. One can confirm that both correlate strongly with the error bars in Fig. 3.

- Finally, our Bayesian near-optimal sets perform much more strongly. They maintain a risk below 3% at least $1-\epsilon = 90\%$ of the time (under *frequentist* sampling), but the top error bars are fairly close to the budget, i.e., they are not overly conservative. As a consequence, their expected return is also much higher and reasonably close to the SAA and MinVar returns. Unlike SAA and MinVar, however, the returns are very stable as seen by the small error bars.

These findings suggest our Bayesian sets have good frequentist performance for moderate N.

6.2. Non-Uniform $\boldsymbol{\theta}^*$

A possible criticism of the previous simulations is that $\boldsymbol{\theta}^* = \frac{1}{d}\mathbf{e}$ is uniform, which may be unrealistic in some applications. Thus, we repeat the above experiments for a non-uniform $\boldsymbol{\theta}^*$ formed by clustering the historical data to form "typical" market scenarios. The results of these experiments agree qualitatively with the above. See Appendix E.1 for details.

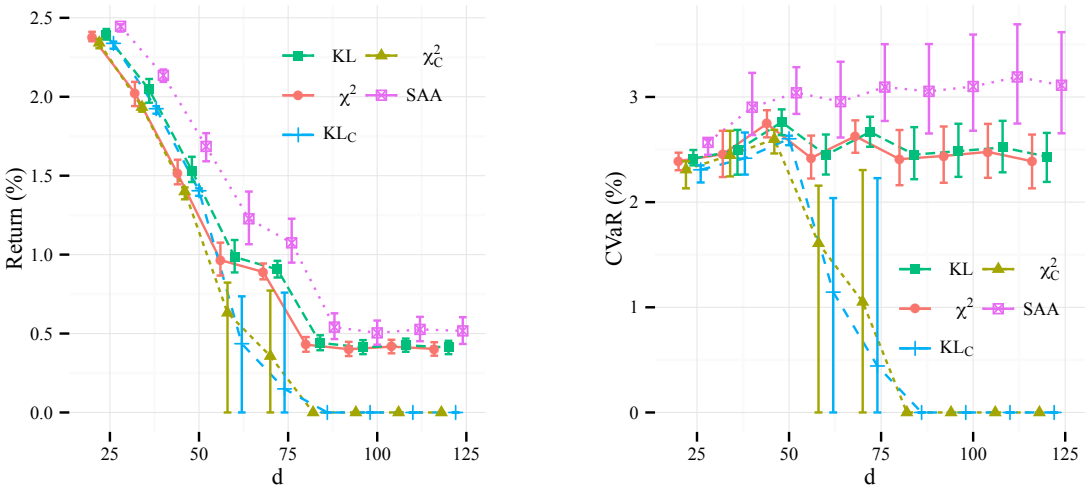

Figure 4 **Return and risk for increasing support size d from Sec. 6.3.**

6.3. Dependence on d

An important implication of our theoretical results is that DRO models using credible or confidence regions may not perform well if d is large. We next study the dependence on d for synthetic data. Specifically, we take the true distribution to be supported on the most recent d monthly returns of our 12 indices and repeatedly sample $N = 300$ data points from this distribution. We form portfolios from this sample and repeat the entire procedure 1000 times, each time recording each portfolio's performance with respect to the true underlying distribution. Fig. 4 presents summary statistics for various d. The Naive and MinVar portfolios perform similarly to the SAA portfolio and are omitted for clarity. Fig. EC.4 in Appendix E.2 shows all portfolios.

As expected, as d increases for a fixed N, all methods perform worse; there is relatively less data to learn a more complicated distribution. More interestingly, the performance degrades more quickly for some methods. Namely, as d increases, the DRO models with confidence-region-based uncertainty sets quickly degrade. For d near 100, they converge to investing in the most-conservative $\mathbf{x} = \mathbf{0}$ portfolio. Similarly, although the SAA portfolio maintains a reasonably good return, as d grows, it violates the risk bound more frequently. By contrast, our Bayesian near-optimal ambiguity sets maintain a return fairly close to the SAA return and safely maintain a risk below 3%. These

observations are fairly robust to the choice of N. See Fig. EC.5 in Appendix E.2 for a similar experiment with $N = 700$ data points.

These experiments confirm the theoretically predicted behavior and are a strong argument for preferring Bayesian near-optimal sets over frequentist confidence regions when d is large.

6.4. Prior Specification

We next study the sensitivity to the choice of prior. Thm. 6 ensures that as $N \to \infty$, this choice becomes irrelevant, but it is less clear what the effect is for finite N. Intuitively, an "ideal" prior would place most of its mass in a neighborhood of the true, unknown $\boldsymbol{\theta}^*$, causing the posterior distribution to concentrate at the true value. For example, in the model of Sec. 6.1, the "ideal" prior is $\boldsymbol{\tau}' = \hat{\tau}_0 \mathbf{e}$ with $\hat{\tau}_0 \to \infty$. The uninformative prior $\boldsymbol{\tau}' = \mathbf{e}$ is less ideal since it puts equal weight on the true $\boldsymbol{\theta}^*$ and other incorrect models. Even less ideal priors might weigh incorrect models more heavily than the true model, e.g., the prior $(\hat{\tau}_0, 1, \ldots, 1)$ with τ_0' very large.

As a first test, we fix $N = 300$ and take the true distribution to be as in Sec. 6.1. We consider various priors of this last form $(\hat{\tau}_0, 1, \ldots, 1)$ as $\hat{\tau}_0$ increases. We stress that for large $\hat{\tau}_0$, this is a highly informative prior that places relatively small weight on the true distribution and most of its weight on incorrect distributions. We choose to scale the first component instead of some other component since in the limit, as $\hat{\tau}_0 \to \infty$, robust portfolios built with our sets and this prior converge to the zero portfolio, which has the worst possible (true) return of any portfolio under the true distribution. Thus, this is a very poor choice of prior.

Fig. 5 shows the return and CVaR with respect to the true distribution over 1000 draws of the data for the portfolio built with $\mathcal{P}^{KL}(\mathcal{S}, \sqrt{\log(1/\epsilon)/\tau_0})$. Clearly, as $\hat{\tau}_0$ increases, the prior weight on the true $\boldsymbol{\theta}^*$ decreases, and the performance suffers. It is not until $\hat{\tau}_0 = 175$, however, that the portfolio incurs more than 3% risk for more than an ϵ fraction of sample paths. In other words, for $\hat{\tau}_0 \leq 175$, despite the choice of prior, the portfolio is still feasible with high *frequentist* probability. We can interpret the value $\hat{\tau}_0 = 175$ via pseudocounts (Gelman et al. 2014); it would take $175 - 1 = 174$ data points for a Bayesian starting from an uninformative, uniform prior to update her belief

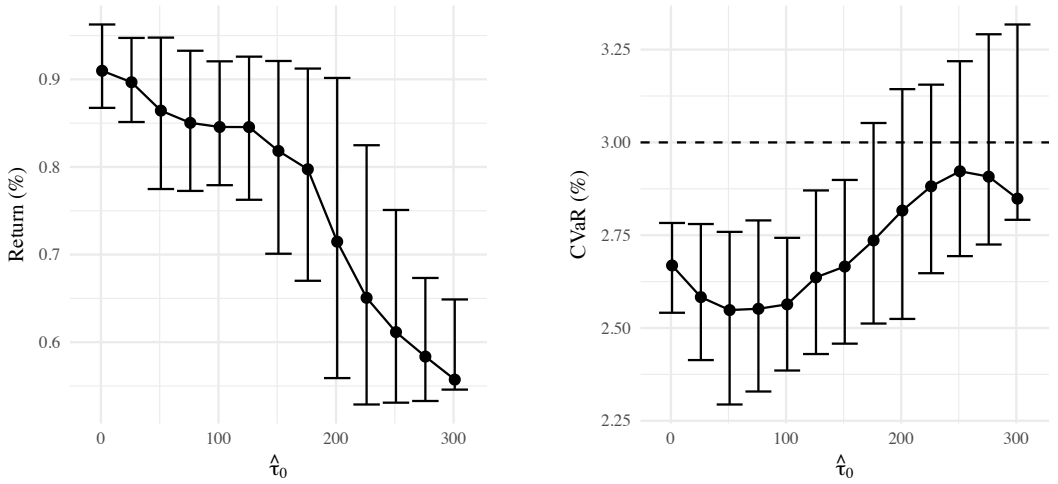

Figure 5 **Return and risk of the "KL" portfolio with an increasingly strong, misinformed prior from Sec. 6.4.**

to the given prior. Compared to $N = 300$, this value suggests we have $174/300 = 58\%$ as much confidence in our prior as we do in our data. Similarly, at $\hat{\tau}_0 = 300$, the average return falls below the average return of the corresponding confidence-region-based set (cf. Fig. 3). A value of $\hat{\tau}_0 = 300$ suggests having $299/300 \approx 100\%$ as much confidence in our prior distribution as we do in our data. Thus, both metrics require very strong prior beliefs on a very poor choice of prior before performance degrades significantly.

These effects are attenuated as N increases. In Fig. 6, we consider the performance of $\mathcal{P}^{KL}(\mathcal{S}, \sqrt{\log(1/\epsilon)/\tau_0})$ with the above informative prior with $\hat{\tau}_0 = 175$ as we increase N. As predicted by Thms. 13 and 18, the performance improves steadily despite the poor choice of prior.

Finally, as a more global assessment of prior sensitivity, we consider the performance of our portfolios for randomly generated priors of various strengths. Specifically, we take the true distribution to be as in Sec. 6.1 with $N = 300$. We consider sequentially $\tau_0' \in [102, 147, 222, 297, 372, 447, 522]$ (corresponding to priors with strengths $[10\%, 25\%, \ldots, 150\%]$ of our data). For each value of τ_0', we generate 100 random priors such that $(\tau_1'/\tau_0', \ldots, \tau_d'/\tau_0')$ are uniformly distributed on the simplex. For each prior, we then compute the performance of our portfolios using sets built from that prior over 1000 repeated samples of data. We consider the average return and CVaR of the portfolio (computed with respect to the true distribution) over these 1000 repeated samples as a good

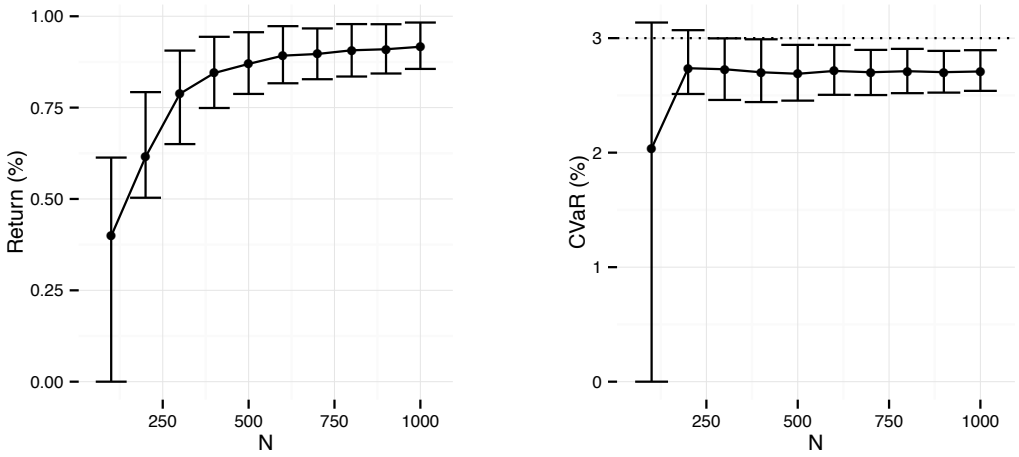

Figure 6 **Return and risk of the "KL" portfolio using prior** $(175, 1, \ldots, 1)$ **as** N **increases from Sec. 6.4.**

proxy for the performance of the method with respect to that prior.[12] Table EC.2 in Appendix E.2 presents summary statistics for this performance across the various priors. As can be seen, although our portfolios can perform poorly for a pathologically bad prior, as in Fig. 5, for most priors, the (frequentist) performance is still relatively good; our portfolios outperform DRO portfolios built with traditional confidence regions even if the strength of the prior belief is large.

6.5. Historical Case Study

Finally, we consider back-testing portfolios built from our sets using real data from Mar. 1998 to Dec. 2014. For each month, we construct portfolios assuming (potentially incorrectly) that the true distribution is supported on the most recent 36 monthly returns and that these returns represent i.i.d. draws from this distribution. We form portfolios using each of our sets and record the realized return from these portfolios using the upcoming month's return. We set $\Gamma = 6\%$, since lower values cause the confidence-region sets to uniformly invest in the portfolio $\mathbf{x} = \mathbf{0}$.

In reality, the true distribution is unlikely to have our assumed support, and the data are unlikely to be independent. Thus, this experiment is a strong test of our methods under model misspecification. Table 2 shows summary statistics for each method over the entire data window.

[12] With 1000 samples, the standard error on the average return and average CVaR is smaller than the last significant digit shown in Table EC.2.

Table 2 **Realized performance from Mar. 1998 to Dec. 2014 from Sec. 6.5. Target CVaR is** 6%.

Method	Avg. Return	Std. Dev	VaR	CVaR	Turnover
χ^2	0.40	2.98	2.40	4.93	0.23
KL	0.41	3.39	3.19	5.95	0.26
χ^2_C	(0.02)	0.76	-	0.09	0.06
KL_C	(0.04)	0.53	-	0.05	0.03
SAA	0.52	4.79	4.97	8.72	0.34
Naive	0.59	4.28	4.06	7.94	0.03
MinVar	0.77	5.65	5.82	10.85	0.19

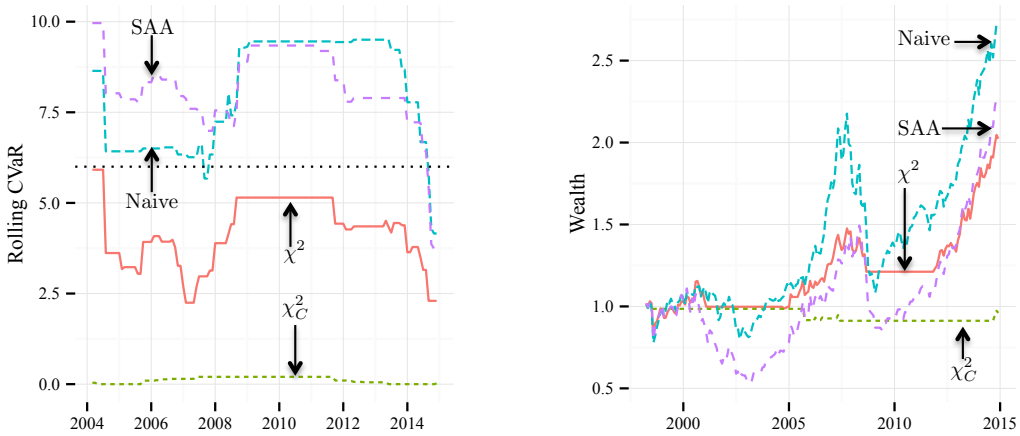

Figure 7 **The left panel shows rolling estimate CVaR using a trailing window of 72 months. The right panel shows the cumulative wealth of each portfolio from Mar. 1998 through Dec. 2014.**

As with our synthetic data, although SAA, MinVar and Naive yield high returns, they exceed the budget for CVaR significantly. By contrast, the confidence-region-based sets maintain a very low CVaR but are so conservative that they frequently do not invest at all, yielding an overall negative return. Our Bayesian near-optimal sets strike a reasonable compromise; they are below the risk threshold and earn a positive return. These observations accord with our synthetic data.

To better understand the portfolio's risk profiles, we plot in the left panel of Fig. 7 an estimate of the CVaR of each portfolio using a trailing window of 72 months. For clarity, we only plot a subset of portfolios; Fig. EC.7 in Sec. E.2 displays all portfolios. The CVaR of the SAA, MinVaR, and Naive portfolios incur much more than the target 6% risk, and the confidence-region variants occur far less.

In the right panel of Fig. 7, we plot the cumulative wealth of an investor who follows each strategy from Mar. 1998 through Dec. 2014 (see also Fig. EC.6 in Sec. E.2). In highly volatile

markets, our near-optimal sets recognize the potential risk and choose not to invest (see, e.g., the large flat regions between 2002 and 2005 or around 2010). Consequently, in down-markets, such as between 2002 and 2005, they outperform SAA. In very strong up-markets, though, they are not as aggressive and, hence, underperform relative to SAA (see the peaks in the graph).

Finally, although our optimization problem Eq. (24) does not explicitly control for multi-period transaction costs, we note that the average monthly turnover for the Bayesian near-optimal sets is much smaller than that for the SAA portfolio in Table 2. Practically, this corresponds to lower transaction costs.

7. Conclusion

In this paper, we introduced a novel Bayesian framework to study the relative strengths of ambiguity sets in data-driven robust optimization. We propose a new class of ambiguity sets that enjoy the usual tractability and asymptotic convergence guarantees but are $\Omega(\sqrt{d})$ smaller than existing proposals while satisfying a Bayesian analogue of a typical robustness property. These results have important implications for using DRO models in practice. Namely, replacing traditional ambiguity sets with our new Bayesian variants may yield significantly better performance with little loss in robustness, especially when d and N are moderate to large.

Acknowledgments

We would like to thank the three anonymous reviewers, associate editor and area editor for their comments and suggestions on previous drafts of this manuscript.

References

Ben-Akiva, M.E., S.R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*, vol. 9. MIT Press.

Ben-Tal, A., D. Den Hertog, A. De Waegenaere, B. Melenberg, G. Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* **59**(2) 341–357.

Ben-Tal, A., D. Den Hertog, J.P. Vial. 2015. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming* **149**(1-2) 265–299.

Ben-Tal, A., L. El Ghaoui, A. Nemirovski. 2009. *Robust Optimization*. Princeton University Press.

Bertsekas, D.P. 1999. *Nonlinear Programming*. Athena Scientific, Belmont.

Bertsimas, D., V. Gupta, N. Kallus. 2017a. Data-driven robust optimization. *Mathematical Programming* doi:10.1007/s10107-017-1125-8.

Bertsimas, D., V. Gupta, N. Kallus. 2017b. Robust sample average approximation. *Mathematical Programming* doi:10.1007/s10107-017-1174-z.

Bickel, P.J., B.J.K Kleijn. 2012. The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics* **40**(1) 206–237.

Castillo, I., R. Nickl. 2013. Nonparametric Bernstein–von Mises theorems in gaussian white noise. *The Annals of Statistics* **41**(4) 1999–2028.

Castillo, I., R. Nickl. 2014. On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics* **42**(5) 1941–1969.

Chatfield, C. 2013. *The Analysis of Time Series: An Introduction*. CRC press.

Chen, C.F. 1985. On asymptotic normality of limiting density functions with Bayesian implications. *Journal of the Royal Statistical Society. Series B (Methodological)* 540–546.

Chen, X., M. Sim, P. Sun. 2007. A robust optimization perspective on stochastic programming. *Operations Research* **55**(6) 1058–1071.

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

DeMiguel, V., L. Garlappi, F.J. Nogales, R. Uppal. 2009a. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science* **55**(5) 798–812.

DeMiguel, V., L. Garlappi, R. Uppal. 2009b. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies* **22**(5) 1915–1953.

Diaconis, P., D. Freedman. 1986. On the consistency of Bayes estimates. *The Annals of Statistics* 1–26.

Efron, B., T. Hastie. 2016. *Computer Age Statistical Inference*, vol. 5. Cambridge University Press.

El Ghaoui, L., M. Oks, F. Oustry. 2003. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research* **51**(4) 543–556.

El Karoui, N., A.E.B. Lim, G.Y. Ban. 2011. Performance-based regularization in mean-cvar portfolio optimization. *arXiv preprint arXiv:1111.2091* .

Fertis, A.G. 2009. A robust optimization approach to statistical estimation problems. Ph.D. thesis, Massachusetts Institute of Technology.

Freedman, D. 1999. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics* 1119–1140.

French, K. 2015. Downloadable data library: 12 industry portfolios. URL `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. Online; accessed 1-June-2015.

Friedman, J., T. Hastie, R. Tibshirani. 2001. *The Elements of Statistical Learning*, vol. 1. Springer Series in Statistics Springer, Berlin.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Rubin. 2014. *Bayesian Data Analysis*, vol. 2. Taylor & Francis.

Geyer, C., G. Meeden. 2013. Asymptotics for constrained Dirichlet distributions. *Bayesian Analysis* **8**(1) 89–110.

Ghosal, S., J.K. Ghosh, A.W. Van Der Vaart. 2000. Convergence rates of posterior distributions. *Annals of Statistics* 500–531.

Gotoh, J.Y., M.J. Kim, A. Lim. 2015. Robust empirical optimization is almost the same as mean-variance optimization. *Available at SSRN 2827400* .

Graves, S.C. 1999. A single-item inventory model for a nonstationary demand process. *Manufacturing & Service Operations Management* **1**(1) 50–61. doi:10.1287/msom.1.1.50.

Kim, Y., J. Lee. 2004. A Bernstein-von Mises theorem in the nonparametric right-censoring model. *Annals of Statistics* 1492–1512.

Klabjan, D., D. Simchi-Levi, M. Song. 2013. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management* **22**(3) 691–710.

Lam, H. 2016. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research* **41**(4) 1248–1275.

Lee, H.L., K.C. So, C.S. Tang. 2000. The value of information sharing in a two-level supply chain. *Management Science* **46**(5) 626–643.

Lim, A.E.B., J.G. Shanthikumar. 2007. Relative entropy, exponential utility, and robust dynamic pricing. *Operations Research* **55**(2) 198–214.

Lim, A.E.B., J.G. Shanthikumar, G.Y. Ban. 2011. Conditional value-at-risk in portfolio optimization: Coherent but fragile. *Operations Research Letters* **39**(3) 163–171.

Lo, A.Y. 1983. Weak convergence for Dirichlet processes. *Sankhyā: The Indian Journal of Statistics, Series A* 105–111.

Luong, H.T. 2007. Measure of bullwhip effect in supply chains with autoregressive demand process. *European Journal of Operational Research* **180**(3) 1086 – 1097. doi:http://dx.doi.org/10.1016/j.ejor.2006.02.050.

McLachlan, G., D. Peel. 2004. *Finite Mixture Models*. John Wiley & Sons.

Michaud, R.O. 1989. The Markowitz optimization enigma: Is "optimized" optimal? *Financial Analysts Journal* **45**(1) 31–42.

Nedic, A., D.P. Bertsekas, A.E. Ozdaglar. 2003. *Convex Analysis and Optimization*. Athena Scientific.

Nemirovski, A., A. Shapiro. 2006. Convex approximations of chance constrained programs. *SIAM Journal on Optimization* **17**(4) 969–996.

Noyan, N., G. Rudolf. 2014. Kusuoka representations of coherent risk measures in general probability spaces. *Annals of Operations Research* 1–15.

Phillips, R.L. 2005. *Pricing and Revenue Optimization*. Stanford University Press.

Postek, K., D. den Hertog, B. Melenberg. 2016. Computationally tractable counterparts of distributionally robust constraints on risk measures. *SIAM Review* **58**(4) 603–650.

Rivoirard, V., J. Rousseau. 2012. Bernstein-von Mises theorem for linear functionals of the density. *The Annals of Statistics* **40**(3) 1489–1523.

Rujeerapaiboon, N., D. Kuhn, W. Wiesemann. 2016. Robust growth-optimal portfolios. *Management Science* **62**(7) 2090–2109.

Rusmevichientong, P., H. Topaloglu. 2012. Robust assortment optimization in revenue management under the multinomial logit choice model. *Operations Research* **60**(4) 865–882.

Scarf, H. 1958. A min-max solution of an inventory problem. K.J. Arrow, S. Karlin, H. Scarf, eds., *Studies in the Mathematical Theory of Inventory and Production*. Sanford University Press, Stanford, 201–209.

Shapiro, A., D. Dentcheva, A.P. Ruszczyński. 2014. *Lectures on Stochastic Programming: Modeling and Theory*, vol. 16. SIAM.

Shapiro, A., A.P Ruszczyński. 2003. *Stochastic Programming*. Elsevier.

Talluri, K., G. Van Ryzin. 2004. Revenue management under a general discrete choice model of consumer behavior. *Management Science* **50**(1) 15–33.

Van der Vaart, A.W. 2000. *Asymptotic Statistics*, vol. 3. Cambridge University Press.

Wiesemann, W., D. Kuhn, M. Sim. 2014. Distributionally robust convex optimization. *Operations Research* **62**(6) 1358–1376.

Wozabal, D. 2014. Robustifying convex risk measures for linear portfolios: A nonparametric approach. *Operations Research* **62**(6) 1302–1315.

Xu, H., C. Caramanis, S. Mannor. 2012. A distributional interpretation of robust optimization. *Mathematics of Operations Research* **37**(1) 95–110.

Xu, H., S. Mannor. 2012. Robustness and generalization. *Machine Learning* **86**(3) 391–423.

Zhu, S., M. Fan, D. Li. 2014. Portfolio management with robustness in both prediction and decision: A mixture model based learning approach. *Journal of Economic Dynamics and Control* **48** 1–25.

Zhu, S., M. Fukushima. 2009. Worst-case conditional value-at-risk with application to robust portfolio management. *Operations Research* **57**(5) 1155–1168.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Online Appendix: Near-Optimal Bayesian Ambiguity Sets for Distributionally Robust Optimization

Appendix A: Additional Parametric Examples of Our Framework

In this appendix, we describe additional examples from the DRO literature that can be recast in our framework by using parametric distributions.

EXAMPLE EC.1 (GAUSSIAN AND TIME-SERIES MODELS). Suppose $\tilde{\boldsymbol{\xi}} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ is normally distributed, but $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$ are unknown and estimated from data. Linear chance constraints can be cast in our framework by noting that $\mathbb{P}_{\boldsymbol{\theta}^*}(\mathbf{a}(\mathbf{x})^T\tilde{\boldsymbol{\xi}} > \mathbf{b}(\mathbf{x})) \leq \kappa \iff \mathbf{a}(\mathbf{x})^T\boldsymbol{\mu}^* + z_{1-\kappa}\sqrt{\mathbf{a}(\mathbf{x})^T\boldsymbol{\Sigma}^*\mathbf{a}(\mathbf{x})} - \mathbf{b}(\mathbf{x}) \leq 0$, which is concave in $\boldsymbol{\theta}^* = (\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ for any functions $\mathbf{a}(\mathbf{x}), \mathbf{b}(\mathbf{x})$. Here, $\Theta = \mathbb{R}^d \times \mathbb{S}_+^d$, where $\mathbb{S}_+^d$ is the cone of positive semidefinite matrices. A typical prior for $\tilde{\boldsymbol{\theta}}$ is the normal-inverse-Wishart prior since if $\tilde{\mathcal{S}}$ is drawn i.i.d., the posterior is also a normal-inverse-Wishart distribution (see Gelman et al. (2014)).

Similarly, expected quadratic constraints can be cast in our framework by noting that $\mathbb{E}_{\boldsymbol{\theta}^*}[\mathbf{a}(\mathbf{x})^T\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}^T\mathbf{a}(\mathbf{x}) + \mathbf{b}(\mathbf{x})^T\tilde{\boldsymbol{\xi}}] \leq \kappa \iff \mathbf{a}(\mathbf{x})^T(\boldsymbol{\Sigma}^* + \boldsymbol{\mu}^*\boldsymbol{\mu}^{*T})\mathbf{a}(\mathbf{x}) + \mathbf{b}(\mathbf{x})^T\boldsymbol{\mu}^* - \kappa \leq 0$, which is concave in the transformed parameter $\boldsymbol{\theta}^* = (\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* + \boldsymbol{\mu}^*\boldsymbol{\mu}^{*T})$ for any $\mathbf{a}(\mathbf{x}), \mathbf{b}(\mathbf{x})$. Again, we take $\Theta = \mathbb{R}^d \times \mathbb{S}_+^d$. Many choices of prior are possible, including specifying that $\tilde{\boldsymbol{\theta}}' = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ follows a normal-inverse-Wishart prior and computing the posterior of $\tilde{\boldsymbol{\theta}}$ by MCMC.

Although multivariate normality may seem contrived, it frequently occurs in time series and forecasting applications. For example, consider an autoregressive process, $\tilde{\xi}_t = \sum_{i=1}^p \theta_i^*\tilde{\xi}_{t-i} + \eta_t$, for $t = 1, \ldots,$ and where η_t represents a random shock at time t. Various authors have used autoregressive processes to model demand and sales (Lee et al. (2000), Luong (2007) and references therein), typically assuming η_t are i.i.d., mean-zero, Gaussian random variables with variance σ^{*2}. Under these assumptions, $\tilde{\xi}_t|\tilde{\xi}_1, \ldots, \tilde{\xi}_{t-1} \sim \mathcal{N}(\sum_{i=1}^p \theta_i^*\tilde{\xi}_{t-i}, \sigma^{*2})$. Similar results hold for other time-series models, such as the ARIMA process of Graves (1999),; vector autoregressive processes; and some linear-state models that occur frequently in econometrics, control, and finance.

We stress that in the autoregressive case and most forecasting applications, the data $\mathcal{S}$ are not drawn i.i.d. from $\mathbb{P}_{\boldsymbol{\theta}^*}$; future realizations depend on the past. Nonetheless, this case can be analyzed within our framework (see Sec. 4.1).

Example EC.2 (Multinomial Logit Model). In the assortment optimization problem, a retailer can offer any subset of d products to consumers. Each consumer purchases at most one item from the assortment. If she purchases item i, the retailer earns revenue r_i, $i = 1, \ldots, d$; otherwise, the retailer earns nothing (the "no-purchase" option). The retailer may strategically choose to not offer certain low-revenue products to induce consumers to purchase higher revenue substitutes. The assortment optimization problem seeks a revenue-maximizing assortment for a given model of consumer choice behavior.

A common behavioral choice model in the operations management literature is the multinomial logit model (Talluri and Van Ryzin 2004), which posits that a consumer assigns utility $\log(\theta_i^*) + \tilde{\eta}_i$ to item i and utility $\tilde{\eta}_0$ to the no-purchase option. The $\tilde{\eta}_i$ are independent Gumbel random variables with mean 0 and scale 1. After assigning utilities, the consumer chooses the item (or no-purchase option) corresponding to the highest utility. The parameters $\boldsymbol{\theta}^* \in \mathbb{R}^d_{++}$ represent preference weights. Under this model the expected revenue for offering assortment $\mathcal{A} \subseteq \{1, \ldots, d\}$ is $\frac{\sum_{i \in \mathcal{A}} r_i \theta_i^*}{1 + \sum_{j \in \mathcal{A}} \theta_j^*}$ (Ben-Akiva and Lerman (1985)).

Rusmevichientong and Topaloglu (2012) proposes and studies the corresponding DRO formulation

$$\max_{\mathcal{A} \subseteq \{1, \ldots d\}} \min_{\boldsymbol{\theta} \in \mathcal{P}} \frac{\sum_{i \in \mathcal{A}} r_i \theta_i}{1 + \sum_{j \in \mathcal{A}} \theta_j}, \tag{EC.1}$$

where $\mathcal{P}$ is one of several non data-driven ambiguity sets. We can cast this problem in our framework by letting $\Theta = \mathbb{R}^d_{++}$ and adopting any prior on $\tilde{\boldsymbol{\theta}}$ (the posterior will be computed by MCMC). We need only show Eq. (EC.1) can be written in the form of Eq. (2). This is readily accomplished by first writing the problem epigraphically as

$$\max_{t, \mathcal{A} \subseteq \{1, \ldots d\}} \quad \left\{ t \ : \ \frac{\sum_{i \in \mathcal{A}} r_i \theta_i}{1 + \sum_{j \in \mathcal{A}} \theta_j} \geq t, \quad \forall \boldsymbol{\theta} \in \mathcal{P} \right\},$$

and then rewriting the constraint as $\sum_{j \in \mathcal{A}} (r_j - t) \theta_j \geq t$, which is concave in $\boldsymbol{\theta}$. As an aside, it is not immediately obvious that Eq. (EC.1) is tractable since it involves optimizing over the 2^d subsets $\mathcal{A}$, but Rusmevichientong and Topaloglu (2012) proves this problem can be solved efficiently by restricting attention to so-called revenue-ordered assortments.

Example EC.3 (Pricing under Generalized Linear Models). Parametric demand modeling is commonplace in pricing and revenue management applications (see (Phillips 2005, Chapt. 3)). Many typical demand functions are special cases of generalized linear models (GLMs), possibly after a transformation of variables. GLMs are popular since they can be fit efficiently using maximum likelihood estimation or

Bayesian methods and often give rise to tractable optimization models. (See Gelman et al. (2014) for a Bayesian overview of GLM theory.)

Recall that demand $\tilde{D}$ follows a GLM if the distribution of $\tilde{D}$ belongs to a pre-specified family parametrized by its mean and there exists a set of explanatory variables $\mathbf{f}$ and a link function $h(\cdot)$ such that $h(\mathbb{E}[\tilde{D}]) = \boldsymbol{\beta}^T\mathbf{f}$ for some weights $\boldsymbol{\beta}$. As an example, if we take $\mathbf{f} = (p, 1)$, $h(\cdot) = \log(\cdot)$ and assume $\tilde{D}$ follows a Poisson distribution, we arrive at the log-linear demand model: $\mathbb{E}[\tilde{D}] = \exp(\beta_1 p + \beta_0)$. For the same link and distribution, if we take $\mathbf{f} = (\log p, 1)$, we obtain the log-log or constant elasticity demand model $\mathbb{E}[\tilde{D}] = e^{\beta_0} p^{\beta_1}$. Alternatively, if we take $\mathbf{f} = (p, 1)$, $h(\cdot) = logit(\cdot)$ and assume $\tilde{D}$ follows a Bernoulli distribution, i.e., customers either buy the product or do not, then we obtain the logit-price demand model $\mathbb{E}[\tilde{D}] = 1/(1 + \exp(-\beta_1 p - \beta_0))$. Other examples can be constructed similarly.

We next show how certain DRO problems with GLMs can be analyzed in our framework. To be concrete, consider the revenue optimization problem

$$\max_{\mathbf{x}\in\mathcal{C}} \quad \sum_{k=1}^{n} x_k h_k^{-1}(\boldsymbol{\beta}^{k\,T}\mathbf{x} + \beta_0^k), \tag{EC.2}$$

where x_k is the price of the k-th product, and the demand $\tilde{D}_k$ for product k is assumed to follow a GLM with link h_k and explanatory variables $(\mathbf{x}, 1)$. This model allows for the possibility of both complements and substitutes. The set $\mathcal{C}$ encodes business constraints on possible prices and any transformations of the decision variables necessary to create the explanatory variables. Since the parameters $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^k, \beta_0^k)_{k=1,\ldots,n}$ are estimated from data, we model them via DRO. Rewrite Eq. (EC.2) as

$$\max_{\mathbf{x}\in\mathcal{C}, t_1,\ldots,t_K} \sum_{k=1}^{n} t_k \quad \text{s.t.} \quad h_k(t_k/x_k) \le \boldsymbol{\beta}^{k\,T}\mathbf{x} + \beta_0^k, \quad k = 1, \ldots, n.$$

The robust counterpart is

$$\max_{\mathbf{x}\in\mathcal{C}, t_1,\ldots,t_K} \sum_{k=1}^{n} t_k \quad \text{s.t.} \quad h_k(t_k/x_k) \le \boldsymbol{\beta}^T\mathbf{x} + \beta_0 \quad \forall(\boldsymbol{\beta}, \beta_0) \in \mathcal{P}_k, \quad k = 1 \ldots, n, \tag{EC.3}$$

where the k-th constraint is of the form Eq. (2) with $\boldsymbol{\theta} = (\boldsymbol{\beta}, \beta_0)$.

Appendix B: Extension to Infinite Dimensional Parameter $\boldsymbol{\theta}$

In some applications, it may be more natural to work with an infinite dimensional parameter $\boldsymbol{\theta}$ directly. Such a parameter might be used, e.g., to represent the density of a continuous random variable. In particular, infinite dimensional parameters occur frequently in Bayesian nonparametric methods, and, intuitively, we

might conjecture that since many of our results are dimension independent, they may also hold in these settings. Making this intuition rigorous involves some technical difficulties, although many key elements of our framework pass through untouched.

Specifically, Thms. 1 and 2 hold in the infinite dimensional setting essentially unchanged; the key elements of the proof are concavity and the separating hyperplane theorem, both of which still apply. Thus, we can still use the described approach to construct new Bayesian ambiguity sets.

Unfortunately, however, posterior consistency of the mean and the Bernstein-von Mises Theorem generally do not hold in infinite dimensions, at least not in the form described in Thm. 6. Indeed, even for seemingly well-behaved problems, Bayesian estimates may either not converge or else converge to an incorrect parameter as the amount of data grows large (see Freedman (1999), Ghosal et al. (2000), and references therein). As discussed in Castillo and Nickl (2014), even identifying sufficient conditions under which posterior consistency and the Bernstein-von Mises Theorem hold is non-trivial. For example, it is not immediately clear what should replace the normal distribution in Thm. 6; a Gaussian distribution over infinite dimensional Θ will depend subtly on the topology of Θ. Castillo and Nickl (2014, 2013), Kim and Lee (2004), Rivoirard and Rousseau (2012), Lo (1983), Bickel and Kleijn (2012) each prove a Bernstein-von-Mises-type result under different technical assumptions on the prior with different modes of convergence to slightly different Gaussian objects. In other words, simple, general-purpose, sufficient conditions for a Bernstein-von Mises phenomenon in infinite dimension arguably have yet to be found, and research is ongoing. Thus, while it seems plausible that an analogue of Thm. 7 exists, analyzing this limiting Gaussian object to establish that the posterior Value-at-Risk converges *uniformly* remains technically challenging. This difficulty is the root of the challenge in extending the subsequent asymptotic analysis to the infinite dimensional case.

Similar comments apply to analyzing moment-based ambiguity sets. Bayesian methods require complete specification of the likelihood $\mathbb{P}_{\mathcal{S}|\boldsymbol{\theta}}$, but without additional assumptions, specifying a few moments, such as the mean and covariance, will not uniquely define this distribution. In our opinion, the most natural approach in this case is to model the unknown distribution of ξ semi-parametrically, i.e., let $\boldsymbol{\theta}$ be composed of a finite dimensional component (corresponding to the unknown moments) and an infinite dimensional nuisance component, which completes the specification (see Van der Vaart (2000, Chapt. 25) for an overview of semi-parametric methods). Within this model, Bayesian inference is possible, but similar technical issues around specifying an appropriate prior and asymptotic analysis arise.

Appendix C: Guidelines for Practitioners

Our results have a number of practical implications for using DRO models.

First, in applications where an approximate feasibility guarantee is sufficient, tuning the radius of the uncertainty set via cross-validation to ensure that the set approximately satisfies the posterior feasibility guarantee should perform very well. Our results suggest that a good radius exists of size $O(N^{-1/2})$, independently of d.

Second, in applications that require a provable (Bayesian) feasibility guarantee at level ϵ, constants like those in Thms. 8 to 12 can provide guidelines for choosing between competing ambiguity sets when N is large. For example, in the finite, discrete model of Ex. 1, when computational resources permit using $\mathcal{P}^{KL}\left(\mathcal{S}, \frac{\sqrt{\log(1/\epsilon)}}{\tau_0}\right)$, Fig. 1 suggests that this set should be preferred to $\mathcal{P}^{\phi}\left(\mathcal{S}, \sqrt{\frac{\phi''(1)\chi^2_{d-1,1-\epsilon}}{2N}}\right)$ for any ϕ-divergence, especially when d is large. If, however, computational resources are too limited to use this set, one might prefer a simpler ϕ-divergence set if ϵ and d are small, and $\mathcal{P}^{\chi^2}\left(\mathcal{S}, \sqrt{\frac{1-\epsilon}{\epsilon(\tau_0+1)}}\right)$, otherwise. The exact notion of "small," here, depends on how much sub-optimality the application can permit. What is important is that these constants allow us to partially quantify this sub-optimality and provide a clear rule for choosing a set. Developing similar constants for other models involves straightforward computations with Gaussian distributions via Thm. 7. Moreover, developing new ambiguity sets for custom applications is possible by directly applying the approach of Sec. 3.

Admittedly, for small N, these constants are less informative. The left panel of Fig. EC.1 shows the ratios $\delta^*(\mathbf{v}|\ \mathcal{P})/\mathrm{VaR}^{1-\epsilon}_{\hat{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ for randomly chosen directions $\mathbf{v}$ on a single sample path varying N and sets

$$\mathcal{P}^{\chi^2}\left(\mathcal{S}, \sqrt{\frac{1-\epsilon}{\epsilon(\tau_0+1)}}\right),\ \ \mathcal{P}^{KL}\left(\mathcal{S}, \sqrt{\log(1/\epsilon)/\tau_0}\right),\ \ \mathcal{P}^{\chi^2}\left(\mathcal{S}, \sqrt{\chi^2_{d-1,1-\epsilon}/N}\right),\ \ \mathcal{P}^{KL}\left(\mathcal{S}, \sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{2N}}\right).$$

These last two sets are confidence regions as $N \to \infty$. We have set $\boldsymbol{\theta}^* = \frac{1}{15}\mathbf{e}$, $\epsilon = .1$, and $\boldsymbol{\tau}' = \mathbf{0}$. We draw attention to the following features:

- The finite ratio can be above or below the asymptotic ratio.
- For small N, the ordering between sets may change. For example, although $\mathcal{P}^{KL}\left(\mathcal{S}, \sqrt{\log(1/\epsilon)/\tau_0}\right)$ is asymptotically smaller than $\mathcal{P}^{\chi^2}\left(\mathcal{S}, \sqrt{\chi^2_{d-1,1-\epsilon}/N}\right)$, for $N < 20$, it is larger in certain directions.
- The empirical rate of convergence differs by set and by its size. For example, $\mathcal{P}^{\chi^2}\left(\mathcal{S}, \sqrt{\chi^2_{d-1,1-\epsilon}/N}\right)$, converges almost immediately; $\mathcal{P}^{KL}\left(\mathcal{S}, \sqrt{\log(1/\epsilon)/\tau_0}\right)$ converges more slowly, and $\mathcal{P}^{KL}\left(\mathcal{S}, \sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{2N}}\right)$ converges even more slowly.

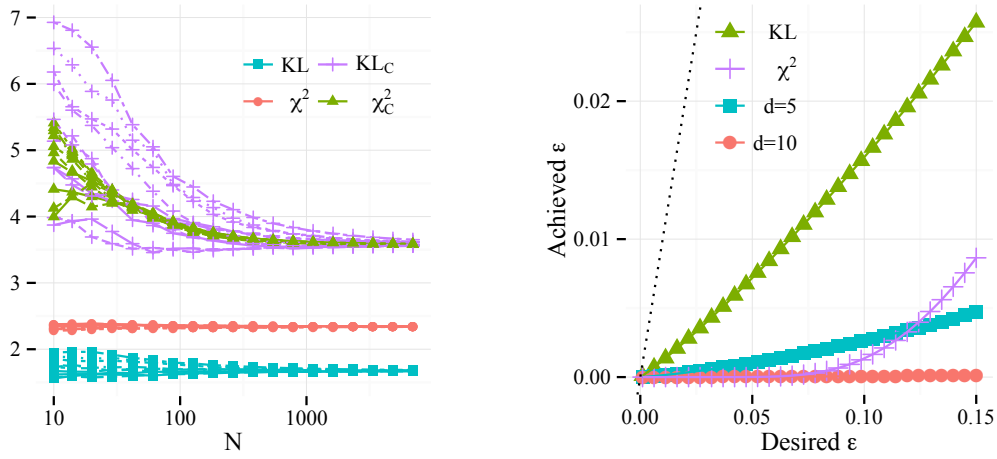

Figure EC.1 **The lefthand panel shows the ratio $\delta^*(\mathbf{v}|\ \mathcal{P})/\mathbf{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ for $\mathcal{P}$ equal to $\mathcal{P}^{KL}\left(\mathcal{S}, \sqrt{\log(1/\epsilon)/\tau_0}\right)$ (denoted "KL"), $\mathcal{P}^{\chi^2}\left(\mathcal{S}, \sqrt{\frac{1-\epsilon}{\epsilon(\tau_0+1)}}\right)$ (denoted"χ^2"), $\mathcal{P}^{KL}\left(\mathcal{S}, \sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{2N}}\right)$ (denoted"KL_C"), or $\mathcal{P}^{\chi^2}\left(\mathcal{S}, \sqrt{\chi^2_{d-1,1-\epsilon}/N}\right)$(denoted "$\chi^2_C$"). We take $\boldsymbol{\theta}^* = \frac{1}{15}\mathbf{e}$, $d = 15$, and $\epsilon = .1$. The righthand panel compares the desired robustness level ϵ with the asymptotically achieved robustness level ϵ' for $\mathcal{P}^{KL}\left(\mathcal{S}, \sqrt{\log(1/\epsilon)/\tau_0}\right)$, $\mathcal{P}^{\chi^2}\left(\mathcal{S}, \sqrt{\frac{1-\epsilon}{\epsilon(\tau_0+1)}}\right)$, and several ϕ-divergence sets for $d = 5, 7, 10$. Closer to the dotted line $\epsilon = \epsilon'$ is "better."**

Moreover, our results help to quantify the level of conservatism in *solutions* to DRO problems as $N \to \infty$ in a frequentist setting. To be precise, although an ambiguity set might be constructed to ensure that a solution is infeasible with (frequentist) probability at most ϵ (the *desired* level), in reality, we expect the solution will be infeasible with probability of at most $\epsilon' < \epsilon$ (the *achieved level*) because of various conservative bounds in the set's construction. By Thm. 18, however, under mild assumptions, $\mathcal{P}^*(\cdot,\ z_{1-\epsilon}/\sqrt{N})$ satisfies the frequentist guarantee at level ϵ asymptotically. Thus, for a candidate ambiguity set, we can compute the smallest ϵ' such that the asymptotically Bayesian optimal set at level ϵ' is still a subset of the candidate set as $N \to \infty$. The gap between ϵ (desired) and ϵ' (achieved) indicates the over-conservatism in the solution.

We illustrate this idea in the case of our sets for finite, discrete support. Consider $\mathcal{P}^{\chi^2}(\mathcal{S}, \sqrt{\frac{1-\epsilon}{\epsilon(\tau_0+1)}})$. A comparison of the radii shows ϵ' should be $1 - \Phi(\sqrt{1/\epsilon - 1})$. Similar computations can be made for the ϕ-divergence sets and our near-optimal variants. We plot these asymptotically achieved robustness levels in the righthand panel of Fig. EC.1. The differences between the desired robustness and actual robustness can be striking, especially for large d. This plot illustrates that $\mathcal{P}^{\chi^2}(\mathcal{S}, \sqrt{\frac{1-\epsilon}{\epsilon(\tau_0+1)}})$ is unsuitable for small ϵ.

Finally, Thm. 16 also has modeling implications. For a given application, there are often multiple possible formulations of the underlying optimization problem, all equally valid approximations of the real, physical

system. Our results suggest favoring formulations in which $g(\boldsymbol{\theta}, \mathbf{x})$ is concave in $\boldsymbol{\theta}$, enabling us to use our new, smaller ambiguity sets instead of larger confidence or credible regions. If such a reformulation is impossible, we should at least favor formulations in which d is small, using feature reduction techniques if necessary to pre-process the data.

Appendix D: Proofs

Proof Thm. 1. Our proof is nearly identical to Bertsimas et al. (2017a) with notation altered for the Bayesian setting. We include it for completeness.

First, suppose $\mathcal{P}(\mathcal{S})$ satisfies the posterior feasibility guarantee for $\mathcal{G}$. For any $\mathbf{v} \in \mathbb{R}^d$, consider $g(\boldsymbol{\theta}, x) = \mathbf{v}^T\boldsymbol{\theta} - x$. Then, $\delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S})) \in \mathcal{X}(\mathcal{P}(\mathcal{S}))$, whereby from the feasibility guarantee, $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}^T\tilde{\boldsymbol{\theta}} \leq \delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}))) \geq 1 - \epsilon$.

For the converse, consider any $g \in \mathcal{G}$ and some $\mathbf{x} \in \mathcal{X}(\mathcal{P}(\mathcal{S}))$. For any $t > 0$, the set $\{\boldsymbol{\theta} \in \Theta : g(\boldsymbol{\theta}, \mathbf{x}) \geq t\}$ is disjoint from $\mathcal{P}(\mathcal{S})$ because $\mathbf{x}$ is robust feasible. Since $g \in \mathcal{G}$, the first set is convex. Meanwhile, the second is convex by assumption. Thus, there exists a strict separating hyperplane $\mathbf{v}^T\boldsymbol{\theta} = v_0$ such that $\boldsymbol{\theta} \in \mathcal{P}(\mathcal{S}) \implies \mathbf{v}^T\boldsymbol{\theta} < v_0$ and $\boldsymbol{\theta} \in \{\boldsymbol{\theta} \in \Theta : g(\boldsymbol{\theta}, \mathbf{x}) \geq t\} \implies \mathbf{v}^T\boldsymbol{\theta} > v_0$. Thus, $v_0 \geq \sup_{\boldsymbol{\theta} \in \mathcal{P}(\mathcal{S})} \mathbf{v}^T\boldsymbol{\theta} = \delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}))$, and

$$\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(g(\tilde{\boldsymbol{\theta}}, \mathbf{x}) \geq t\right) \leq \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}^T\tilde{\boldsymbol{\theta}} > v_0) \leq \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\mathbf{v}^T\tilde{\boldsymbol{\theta}} > \delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}))\right) \leq \epsilon,$$

where the last inequality follows by assumption. Taking the limit as $t \to 0$ yields the result. □

Proof of Thm. 2 ("If" direction) By continuity of probability, $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ is a closed function. It is positively homogenous by construction and convex by assumption. Consequently, there exists a unique, closed, and convex $\mathcal{P}^*$ such that $\delta^*(\mathbf{v}|\ \mathcal{P}^*) = \mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ (Nedic et al. 2003). Notice that $\mathcal{P}^*$ satisfies Eq. (5) with equality. By Thm. 1, $\mathcal{P}^*$ satisfies the posterior feasibility guarantee, and, by Eq. (6), $\mathcal{P}^*$ is a subset of any other convex set that also satisfies this guarantee.

("Only if" direction) Since $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ is non-convex, there exists $\mathbf{v}_1, \mathbf{v}_2$ and $0 < \lambda < 1$ such that

$$\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2) > \lambda\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}_1) + (1-\lambda)\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}_2). \tag{EC.4}$$

Since $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ is positively homogenous, it cannot be that $\mathbf{v}_1 = \alpha\mathbf{v}_2$ for some $\alpha \geq 0$. For $k = 1, 2$, define $\mathcal{P}_k = \{\boldsymbol{\theta} \in \mathbb{R}^d : \mathbf{v}_k^T\boldsymbol{\theta} \leq \mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}_k)\}$. A direct computation yields

$$\delta^*(\mathbf{v}|\ \mathcal{P}_k) = \begin{cases} \alpha\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}_k) & \text{if } \mathbf{v} = \alpha\mathbf{v}_k \text{ for some } \alpha \geq 0 \\ \infty & \text{otherwise.} \end{cases} \tag{EC.5}$$

Notice that $\delta^*(\mathbf{v}|\ \mathcal{P}_k)$ upper bounds $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ so that by (5), both $\mathcal{P}_1, \mathcal{P}_2$ satisfy the posterior feasibility guarantee.

Now, by contradiction, suppose that there exists a Bayesian optimal ambiguity set $\mathcal{P}^*$. From above, $\mathcal{P}^* \subseteq \mathcal{P}_1 \cap \mathcal{P}_2$. Let $\overline{\mathbf{v}} \equiv \lambda \mathbf{v}_1 + (1-\lambda)\mathbf{v}_2$. Then, since $\mathcal{P}^*$ satisfies a posterior guarantee,

$$\begin{aligned}
\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\overline{\mathbf{v}}) &\leq \delta^*(\overline{\mathbf{v}}|\ \mathcal{P}^*) && \text{(by Thm. 1)} \\
&\leq \delta^*(\overline{\mathbf{v}}|\ \mathcal{P}_1 \cap \mathcal{P}_2) && \text{(by Eq. (6) and } \mathcal{P}^* \subseteq \mathcal{P}_1 \cap \mathcal{P}_2).
\end{aligned}$$

From Ben-Tal et al. (2015), $\delta^*(\overline{\mathbf{v}}|\ \mathcal{P}_1 \cap \mathcal{P}_2) = \min_{\mathbf{y}} \delta^*(\mathbf{y}|\ \mathcal{P}_1) + \delta^*(\overline{\mathbf{v}} - \mathbf{y}|\ \mathcal{P}_2)$. Since $\mathbf{y} \mapsto \lambda \mathbf{v}_1$ is feasible,

$$\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\overline{\mathbf{v}}) \leq \delta^*(\lambda \mathbf{v}_1|\ \mathcal{P}_1) + \delta^*(\overline{\mathbf{v}} - \lambda \mathbf{v}_1|\ \mathcal{P}_2) = \lambda \mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}_1) + (1-\lambda)\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}_2)$$

by Eq. (EC.5) and definition of $\overline{\mathbf{v}}$. This contradicts Eq. (EC.4).

□

The following lemma will prove useful in the remainder:

LEMMA EC.1. *Fix any $\mathcal{S}$ and suppose $VaR^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) \leq \delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}))$, $ri(\mathcal{P}(\mathcal{S}) \cap \Theta) \neq \emptyset$. Then, $VaR^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) \leq \delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}) \cap \Theta)$.*

Proof of Lemma. Recall that Θ is convex by assumption. From Ben-Tal et al. (2015), $\delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}) \cap \Theta) = \min_{\mathbf{y}} \delta^*(\mathbf{v} - \mathbf{y}|\ \mathcal{P}(\mathcal{S})) + \delta^*(\mathbf{y}|\ \Theta)$. Then, letting $\mathbf{y}^*$ be an optimizer of this minimization,

$$\begin{aligned}
\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}^T\tilde{\boldsymbol{\theta}} > \delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}) \cap \Theta) &= \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\mathbf{v}^T\tilde{\boldsymbol{\theta}} > \delta^*\left(\mathbf{v} - \mathbf{y}^*|\ \mathcal{P}(\mathcal{S})\right) + \delta^*\left(\mathbf{y}^*|\ \Theta\right)\right) \\
&\leq \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left((\mathbf{v} - \mathbf{y}^*)^T\tilde{\boldsymbol{\theta}} > \delta^*(\mathbf{v} - \mathbf{y}^*|\ \mathcal{P}(\mathcal{S}))\right) + \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\mathbf{y}^{*T}\tilde{\boldsymbol{\theta}} > \delta^*(\mathbf{y}^*|\ \Theta)\right) \\
&\leq \epsilon + 0,
\end{aligned}$$

where the last line follows because $\tilde{\boldsymbol{\theta}} \in \Theta$, $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}$-a.s.

Proof of Thm. 3. Let r be the affine dimension of Θ. Using Eq. (9),

$$\begin{aligned}
\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) &= \mathbf{v}_{r+1,d}^T\boldsymbol{\beta} + \mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}_{1,r}|\mathcal{S}}(\mathbf{v}_{1,r} + \mathbf{A}^T\mathbf{v}_{r+1,d}) \\
&\leq \mathbf{v}_{r+1,d}^T\boldsymbol{\beta} + (\mathbf{v}_{1,r} + \mathbf{A}^T\mathbf{v}_{r+1,d})^T\boldsymbol{\mu}_{1,r} + \sqrt{\frac{1}{\epsilon} - 1}\sqrt{(\mathbf{v}_{1,r} + \mathbf{A}^T\mathbf{v}_{r+1,d})^T\boldsymbol{\Sigma}_{1,r}^{-1}(\mathbf{v}_{1,r} + \mathbf{A}^T\mathbf{v}_{r+1,d})} \\
&= \mathbf{v}_{r+1,d}^T\boldsymbol{\beta} + \max_{\boldsymbol{\theta}_{1,r}:(\boldsymbol{\theta}_{1,r} - \boldsymbol{\mu}_{1,r})^T\boldsymbol{\Sigma}_{1,r}^{-1}(\boldsymbol{\theta}_{1,r} - \boldsymbol{\mu}_{1,r}) \leq \frac{1}{\epsilon} - 1} (\mathbf{v}_{1,r} + \mathbf{A}^T\mathbf{v}_{r+1,d})^T\boldsymbol{\theta}_{1,r},
\end{aligned}$$

where the inequality follows from Eq. (7) and the last equality follows from a standard formula for the support function of an ellipse. Next, this last optimization is equivalent to

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \mathbf{v}^T\boldsymbol{\theta} \\ \text{s.t.} \quad & \|\boldsymbol{\theta}-\boldsymbol{\mu}_N\|_{\boldsymbol{\Sigma}_N} \le \sqrt{\frac{1}{\epsilon}-1} \\ & \boldsymbol{\theta}_{r+1,d} = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\theta}_{1,r} \end{aligned}$$

by definition of $\boldsymbol{\Sigma}_N^{-1}$ in Eq. (10). Combining Lemma EC.1 with Thm. 1 proves the result. □

Proof of Thm. 4. For the first part of the theorem, we show first that $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ is convex. Note $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{0}) = \mathbf{0}$. Furthermore, since $\tilde{\theta}_1 + \tilde{\theta}_2 = 1$ almost surely, $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{e}) = 1$, $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(-\mathbf{e}) = -1$, and $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(v\mathbf{e}) = |v|\text{sign}(v) = v$ by positive homogeneity. Next, assuming $\mathbf{v} \neq \mathbf{0}$, $v_1 \neq v_2$, we have two cases.

First, suppose $v_1 > v_2$. Then, $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(v_1\tilde{\theta}_1 + v_2\tilde{\theta}_2 \le t) = \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\tilde{\theta}_1 \le \frac{t-v_2}{v_1-v_2})$ since $\tilde{\theta}_1 + \tilde{\theta}_2 = 1$. Since $\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}$ is Dirichlet with parameters (τ_1, τ_2), $\tilde{\theta}_1|\mathcal{S}$ follows a Beta distribution with parameter (τ_1, τ_2). Setting this probability equal to $1-\epsilon$ and solving for t yields $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) = v_2 + (v_1 - v_2)\beta_{1-\epsilon}(\tau_1, \tau_2)$.

Next, suppose $v_1 < v_2$. By symmetry, $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) = v_1 + (v_2 - v_1)\beta_{1-\epsilon}(\tau_2, \tau_1)$. We will rewrite this expression slightly so that it can easily be combined with the previous case. We claim that for $0 < \epsilon < 0.5$,

$$\beta_{1-\epsilon}(\tau_1, \tau_2) + \beta_{1-\epsilon}(\tau_2, \tau_1) \ge 1. \tag{EC.6}$$

Indeed, if this were not true, then, since $\tilde{\theta}_1 + \tilde{\theta}_2 = 1$,

$$1 \;=\; \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\tilde{\theta}_1 + \tilde{\theta}_2 > \beta_{1-\epsilon}(\tau_1, \tau_2) + \beta_{1-\epsilon}(\tau_2, \tau_1)\right) \;\le\; \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\tilde{\theta}_1 > \beta_{1-\epsilon}(\tau_1, \tau_2)\right) + \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\tilde{\theta}_2 > \beta_{1-\epsilon}(\tau_2, \tau_1)\right) \;\le\; 2\epsilon,$$

which is a contradiction since $\epsilon < 0.5$.

Multiplying Eq. (EC.6) by $(v_1 - v_2)$ and rearranging terms, we obtain

$$v_1 > v_2 \iff v_2 + (v_1 - v_2)\beta_{1-\epsilon}(\tau_1, \tau_2) > v_1 + (v_2 - v_1)\beta_{1-\epsilon}(\tau_2, \tau_1).$$

Comparing this to our above two expressions for $\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ shows that

$$\text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) = \max\left(\beta_{1-\epsilon}(\tau_1, \tau_2)v_1 + (1 - \beta_{1-\epsilon}(\tau_1, \tau_2))v_2,\ (1 - \beta_{1-\epsilon}(\tau_2, \tau_1))v_1 + \beta_{1-\epsilon}(\tau_2, \tau_1)v_2\right). \tag{EC.7}$$

This formula is also accurate when $\mathbf{v} = 0$ and $v_1 = v_2$ and is convex. To complete the first part of the proof, observe that the support function of the set defined in the theorem is Eq. (EC.7).

The second part of the theorem can be validated numerically. For most examples, one can observe non-convexity along the line $\gamma \mapsto \text{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(1-\gamma, 1+\gamma, 0, \ldots, 0)$ for γ slightly positive and slightly negative. We

prove this formally in the special case when $\mathcal{S}$ is such that $\boldsymbol{\tau} = (1,1,1,0,\ldots,0)$. Note that this posterior requires that we take an improper prior $\boldsymbol{\tau}' = \mathbf{0}$.

By the merging property of the Dirichlet distribution, $(\tilde{\theta}_1, \tilde{\theta}_2, \sum_{i=3}^d \tilde{\theta}_i)$ also has a Dirichlet distribution with parameter $(1,1,1)$, i.e., it is uniform over the simplex. By integrating,

$$\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(v_1\theta_1 + v_2\theta_2 \le t) = \frac{t^2 - 2tv_2 + v_1v_2}{v_2(v_1 - v_2)} \quad \text{if } 0 < v_1 < t < v_2.$$

By setting this probability equal to $1-\epsilon$ and solving for t, we obtain two roots, only the smaller of which satisfies $0 < v_1 < t < v_2$. Thus, we conclude that when $0 < v_1 < v_2$, $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(v_1, v_2, \mathbf{0}) = v_2 - \sqrt{\epsilon}\sqrt{v_2(v_2 - v_1)}$. This computation is symmetric in v_1, v_2, so that when $0 < v_2 < v_1$, we have $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(v_1, v_2, \mathbf{0}) = \mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(v_2, v_1, \mathbf{0})$. For γ sufficiently small, $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(1-\gamma, 1+\gamma, \mathbf{0}) = 1 + |\gamma| - \sqrt{2\epsilon}\sqrt{(1+|\gamma|)|\gamma|}$, which one can verify directly is non-convex at $\gamma = 0$. Thus, $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ is non-convex. □

Proof of Thm. 5. We require the following well-known result (see, e.g., Gelman et al. (2014)):

Let $\tilde{Y}_1, \ldots, \tilde{Y}_d$ be independent Gamma random variables with $\tilde{Y}_i \sim \mathrm{Gamma}(\tau_i, 1)$. Then, $(\tilde{Y}_1/\sum_{i=1}^d \tilde{Y}_i, \ldots, \tilde{Y}_d/\sum_{i=1}^d \tilde{Y}_i)$ follows a Dirichlet distribution with parameter $\boldsymbol{\tau}$.

We can now bound $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})$ using a technique similar to that of Nemirovski and Shapiro (2006):

$$\begin{aligned}
\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}^T\tilde{\boldsymbol{\theta}} \ge t) &= \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\sum_{i=1}^d v_i\tilde{Y}_i \ge t\sum_{i=1}^d \tilde{Y}_i\right) \\
&= \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\sum_{i=1}^d (v_i - t)\tilde{Y}_i \ge 0\right) \\
&\le \inf_{\lambda > \underline{\lambda}} \prod_{i=1}^d \mathbb{E}[e^{\frac{v_i - t}{\lambda}\tilde{Y}_i}] \\
&= \inf_{\lambda > \underline{\lambda}} \prod_{i=1}^d \left(1 - \frac{v_i - t}{\lambda}\right)^{-\tau_i}.
\end{aligned}$$

The inequality follows from Markov's inequality and the independence of the $\tilde{Y}_i$, and the last equality follows from the moment-generating function of a Gamma random variable. Throughout, $\underline{\lambda} \equiv \max_j (v_j - t)^+$.

Thus, $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) \le t$ if there exists $\lambda > \underline{\lambda}$ such that $\prod_{i=1}^d \left(1 - \frac{v_i - t}{\lambda}\right)^{-\tau_i} \le \epsilon$, or, equivalently,

$$\inf_{\lambda > \underline{\lambda}} \lambda \frac{\log(1/\epsilon)}{\tau_0} - \lambda \sum_{i=1}^d \mu_{N,i} \log\left(1 - \frac{v_i - t}{\lambda}\right) \le 0.$$

Using Theorem 1 of Ben-Tal et al. (2013), we recognize the lefthand side as $\delta^*(\mathbf{v} - t\mathbf{e} \mid \mathcal{Q}) \le 0$ for $\mathcal{Q} = \left\{\boldsymbol{\theta} \ge \mathbf{0} : \sum_{i=1}^d \mu_{N,i} \log\left(\frac{\mu_{N,i}}{\theta_i}\right) \le \frac{\log(1/\epsilon)}{\tau_0}\right\}$. Since $\boldsymbol{\theta} \ge 0$ for all $\boldsymbol{\theta} \in \mathcal{Q}$, by rescaling,

$$(\mathbf{v} - t\mathbf{e})^T\boldsymbol{\theta} \le 0 \ \ \forall \boldsymbol{\theta} \in \mathcal{Q} \iff (\mathbf{v} - t\mathbf{e})^T\boldsymbol{\theta} \le 0 \ \ \forall \boldsymbol{\theta} \in \mathcal{Q} \cap \{\boldsymbol{\theta} : \mathbf{e}^T\boldsymbol{\theta} = 1\}, \iff \mathbf{v}^T\boldsymbol{\theta} \le t \ \ \forall \boldsymbol{\theta} \in \mathcal{Q} \cap \{\boldsymbol{\theta} : \mathbf{e}^T\boldsymbol{\theta} = 1\}.$$

This last set is $\mathcal{P}^{KL}(\mathcal{S}, \sqrt{\frac{\log(1/\epsilon)}{\tau_0}})$. Since this inequality holds for arbitrary $(\mathbf{v}, t)$, we have shown $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) \leq \delta^*\left(\mathbf{v} |\ \mathcal{P}^{KL}(\mathcal{S}, \sqrt{\frac{\log(1/\epsilon)}{\tau_0}})\right)$, completing the proof. □

The following lemma will be used repeatedly in what follows. Let $\|\cdot\|_F$ denote the Frobenius norm.

LEMMA EC.2. *For any positive semidefinite matrices* $\mathbf{A}, \mathbf{B}$*, and any* $\mathbf{v}$*,*

$$|\|\mathbf{v}\|_{\mathbf{A}^{-1}} - \|\mathbf{v}\|_{\mathbf{B}^{-1}}| \leq \sqrt{|\mathbf{v}^T(\mathbf{A}-\mathbf{B})\mathbf{v}|} \leq \|\mathbf{v}\|_{\mathbf{A}^{-1}} \sqrt{\|\mathbf{A}-\mathbf{B}\|_F}.$$

Proof. For non-negative a, b, we have the identity $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a-b|}$. Thus,

$$|\|\mathbf{v}\|_{\mathbf{A}^{-1}} - \|\mathbf{v}\|_{\mathbf{B}^{-1}}| \leq \sqrt{|\mathbf{v}^T\mathbf{A}\mathbf{v} - \mathbf{v}^T\mathbf{B}\mathbf{v}|} = \sqrt{|\mathbf{v}^T(\mathbf{A}-\mathbf{B})\mathbf{v}|},$$

proving the first inequality. For the second, let $\mathbf{C}^T\mathbf{C} = \mathbf{A}$ be a Cholesky decomposition of $\mathbf{A}$ and $\mathbf{C}^{-1}$ be a corresponding pseudoinverse. Then,

$$\sqrt{|\mathbf{v}^T(\mathbf{A}-\mathbf{B})\mathbf{v}|} = \sqrt{|\mathbf{v}^T\mathbf{C}^T(\mathbf{C}^{-T}(\mathbf{A}-\mathbf{B})\mathbf{C}^{-1})\mathbf{C}\mathbf{v}|} \leq \sqrt{\mathbf{v}^T\mathbf{C}^T\mathbf{C}\mathbf{v} \cdot \|\mathbf{C}^{-T}(\mathbf{A}-\mathbf{B})\mathbf{C}^{-1}\|_2} = \|\mathbf{v}\|_{\mathbf{A}^{-1}}\sqrt{\|\mathbf{A}-\mathbf{B}\|_2}$$

because $\|\mathbf{C}^{-T}(\mathbf{A}-\mathbf{B})\mathbf{C}^{-1}\|_2 = \|\mathbf{A}-\mathbf{B}\|_2$ by definition of the spectral norm. Note that $\|\cdot\|_2 \leq \|\cdot\|_F$ to complete the proof. □

Proof of Thm. 7 We first prove Eq. (15). Let $R_N(\mathcal{S}) = \sup_{\mathcal{A}} \left|\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\sqrt{N}(\tilde{\boldsymbol{\theta}} - \mu_N) \in \mathcal{A}\right) - \mathbb{P}(\tilde{\boldsymbol{\zeta}} \in \mathcal{A})\right|$ denote the total variation distance from Thm. 6 for realization $\mathcal{S}$. For any $\mathbf{v} \in \mathbb{R}^d$ and any $\delta > 0$,

$$\begin{aligned}
1-\epsilon &\leq \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\mathbf{v}^T\tilde{\boldsymbol{\theta}} \leq \mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) + \delta\right) \\
&= \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}^T(\tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}_N)\sqrt{N} \leq (\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) + \delta - \mathbf{v}^T\boldsymbol{\mu}_N)\sqrt{N}) \\
&\leq \mathbb{P}(\mathbf{v}^T\tilde{\boldsymbol{\zeta}} \leq (\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) + \delta - \mathbf{v}^T\boldsymbol{\mu}_N)\sqrt{N}) + R_N(\mathcal{S}) \\
&= \Phi\left(\frac{(\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) + \delta - \mathbf{v}^T\boldsymbol{\mu}_N)\sqrt{N}}{\|\mathbf{v}\|_{\mathcal{I}(\boldsymbol{\theta}^*)}}\right) + R_N(\mathcal{S}),
\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Rearranging terms yields

$$\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) \geq -\delta + \mathbf{v}^T\boldsymbol{\mu}_N + \frac{z_{1-\epsilon-R_N(\mathcal{S})}}{\sqrt{N}}\|\mathbf{v}\|_{\mathcal{I}(\boldsymbol{\theta}^*)}. \tag{EC.8}$$

(If $1-\epsilon-R(\mathcal{S}) < 0$, interpret the righthand side as $-\infty$.) Taking the limit as $\delta \to 0$ yields a lower-bound. A similar argument starting from the identity $1-\epsilon \geq \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\mathbf{v}^T\tilde{\boldsymbol{\theta}} \leq \mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) - \delta\right)$ yields a corresponding upper-bound:

$$\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v}) \leq \mathbf{v}^T\boldsymbol{\mu}_N + \frac{z_{1-\epsilon+R_N(\mathcal{S})}}{\sqrt{N}}\|\mathbf{v}\|_{\mathcal{I}(\boldsymbol{\theta}^*)}. \tag{EC.9}$$

(If $1-\epsilon+R_N(\mathcal{S})>1$, interpret the righthand side as ∞.) Combining yields

$$\sup_{\mathbf{v}\in\mathbb{R}^d:\|\mathbf{v}\|=1}\sqrt{N}\left|\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\mathbf{v})-\mathbf{v}^T\boldsymbol{\mu}_N-\frac{z_{1-\epsilon}}{\sqrt{N}}\|\mathbf{v}\|_{\mathcal{I}(\boldsymbol{\theta}^*)}\right|\le\sup_{\mathbf{v}\in\mathbb{R}^d:\|\mathbf{v}\|=1}\|\mathbf{v}\|_{\mathcal{I}(\boldsymbol{\theta}^*)}\left(z_{1-\epsilon+R_N(\mathcal{S})}-z_{1-\epsilon-R_N(\mathcal{S})}\right)$$
$$\le\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)^{-1}\|_F}\left(z_{1-\epsilon+R_N(\mathcal{S})}-z_{1-\epsilon-R_N(\mathcal{S})}\right),$$

where the last line follows from the Cauchy-Schwarz inequality $\mathbf{v}^T\mathcal{I}(\boldsymbol{\theta}^*)^{-1}\mathbf{v}\le\|\mathbf{v}\mathbf{v}^T\|_F\|\mathcal{I}(\boldsymbol{\theta}^*)^{-1}\|_F$. Now reinterpret this last inequality for the random draw $\tilde{\mathcal{S}}$. Since the normal quantile is continuous in its argument and $R_N(\tilde{\mathcal{S}})\to_{\mathbb{P}_\mathcal{S}}0$ by A3, we conclude

$$\sup_{\mathbf{v}\in\mathbb{R}^d:\|\mathbf{v}\|=1}\sqrt{N}\left|\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\tilde{\mathcal{S}}}(\mathbf{v})-\mathbf{v}^T\tilde{\boldsymbol{\mu}}_N-\frac{z_{1-\epsilon}}{\sqrt{N}}\|\mathbf{v}\|_{\mathcal{I}(\boldsymbol{\theta}^*)}\right|\to_{\mathbb{P}_\mathcal{S}}0.$$

Thus, to prove Eq. (15), it remains to show that

$$\sup_{\mathbf{v}\in\mathbb{R}^d:\|\mathbf{v}\|=1}\left|\|\mathbf{v}\|_{\mathcal{I}(\boldsymbol{\theta}^*)}-\sqrt{N}\|\mathbf{v}\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}}\right|\to_{\mathbb{P}_\mathcal{S}}0. \tag{EC.10}$$

Note $\sqrt{N}\|\mathbf{v}\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}}=\|\mathbf{v}\|_{N^{-1}\tilde{\boldsymbol{\Sigma}}_N^{-1}}$. Then, by Lemma EC.2,

$$\sup_{\mathbf{v}\in\mathbb{R}^d:\|\mathbf{v}\|=1}\left|\|\mathbf{v}\|_{\mathcal{I}(\boldsymbol{\theta}^*)}-\sqrt{N}\|\mathbf{v}\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}}\right|\le\sup_{\mathbf{v}\in\mathbb{R}^d:\|\mathbf{v}\|=1}\sqrt{\left|\mathbf{v}^T(\mathcal{I}(\boldsymbol{\theta}^*)^{-1}-N\tilde{\boldsymbol{\Sigma}}_N)\mathbf{v}\right|}\le\sqrt{\left\|\mathcal{I}(\boldsymbol{\theta}^*)^{-1}-N\tilde{\boldsymbol{\Sigma}}_N\right\|_F},$$

where the last inequality follows from the Cauchy-Schwarz inequality. By A3, $\left\|\mathcal{I}(\boldsymbol{\theta}^*)^{-1}-N\tilde{\boldsymbol{\Sigma}}_N\right\|_F\to_{\mathbb{P}_\mathcal{S}}0$, which completes the proof of Eq. (15).

We next prove that $\mathcal{P}^*(\tilde{\mathcal{S}},(1+\kappa)z_{1-\epsilon}/\sqrt{N})$ satisfies the posterior guarantee with $\mathbb{P}_\mathcal{S}$-probability tending to 1. Since $\boldsymbol{\theta}^*\in\mathrm{ri}(\Theta)$ and $\tilde{\boldsymbol{\mu}}_N\to_{\mathbb{P}_\mathcal{S}}\boldsymbol{\theta}^*$ by A3, with $\mathbb{P}_\mathcal{S}$-probability tending to 1, $\mathcal{P}^*(\tilde{\mathcal{S}},z_{1-\epsilon}(1+\kappa)/\sqrt{N})\subset\mathrm{ri}(\Theta)$. It follows that for any $\mathbf{v}\in\mathbb{R}^d$, $\delta^*(\mathbf{v}|\ \mathcal{P}^*(\tilde{\mathcal{S}},z_{1-\epsilon}(1+\kappa)/\sqrt{N})\to_{\mathbb{P}_\mathcal{S}}\mathbf{v}^T\tilde{\boldsymbol{\mu}}_N+(1+\kappa)z_{1-\epsilon}\|\mathbf{v}\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}}$. Thus, from the previous part, with $\mathbb{P}_\mathcal{S}$-probability tending to 1, $\delta^*(\mathbf{v}|\ \mathcal{P}^*(\tilde{\mathcal{S}},z_{1-\epsilon}(1+\kappa)/\sqrt{N}))$ upper bounds $\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\tilde{\mathcal{S}}}(\mathbf{v})$ for all $\mathbf{v}\in\mathbb{R}^d:\|\mathbf{v}\|=1$, whereby from Eq. (5), $\mathcal{P}^*(\tilde{\mathcal{S}},z_{1-\epsilon}(1+\kappa)/\sqrt{N})$ satisfies the posterior guarantee.

Finally, for the last statement, notice that $\delta^*(\mathbf{v}|\ \mathcal{P}^*(\tilde{\mathcal{S}},z_{1-\epsilon}(1-\kappa)/\sqrt{N}))\le\mathbf{v}^T\tilde{\boldsymbol{\mu}}_N+(1-\kappa)z_{1-\epsilon}\|\mathbf{v}\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}}$. Consequently, with $\mathbb{P}_\mathcal{S}$-probability tending to 1, $\delta^*(\mathbf{v}|\ \mathcal{P}^*(\tilde{\mathcal{S}},z_{1-\epsilon}(1-\kappa)/\sqrt{N}))\le\mathrm{VaR}^{1-\epsilon}_{\tilde{\boldsymbol{\theta}}|\tilde{\mathcal{S}}}(\mathbf{v})$. Using (6), this implies $\mathcal{P}^*(\tilde{\mathcal{S}},z_{1-\epsilon}(1-\kappa)/\sqrt{N})$ is a subset of any ambiguity set that satisfies the posterior guarantee. □

Proof of Thm. 8 The proof is immediate from the definitions. □

Proof of Thm. 9. To show that $\mathcal{P}^{KL}\left(\tilde{\mathcal{S}},\sqrt{\frac{\log(1/\epsilon)}{\tau_0+2}}\right)$ is $\left(\frac{\sqrt{2\log(1/\epsilon)}}{z_{1-\epsilon}}\right)$ near-optimal, it suffices to show that, with $\mathbb{P}_\mathcal{S}$-probability tending to 1, any $\boldsymbol{\theta}\in\mathcal{P}^{KL}\left(\tilde{\mathcal{S}},\sqrt{\frac{\log(1/\epsilon)}{\tau_0+2}}\right)$ also satisfies $\boldsymbol{\theta}-\tilde{\boldsymbol{\mu}}_N\in\left(\frac{\sqrt{2\log(1/\epsilon)}}{z_{1-\epsilon}}\right)\left(\mathcal{P}^*(\tilde{\mathcal{S}},z_{1-\epsilon}/\sqrt{N})-\tilde{\boldsymbol{\mu}}_N\right)$. This last requirement is equivalent to

$$\sum_{i=1}^d\frac{(\tilde{\mu}_{N,i}-\theta_i)^2}{2\tilde{\mu}_{N,i}}\le\frac{1}{2}\frac{z_{1-\epsilon}^2}{\tau_0+1}\cdot\frac{2\log(1/\epsilon)}{z_{1-\epsilon}^2}=\frac{\log(1/\epsilon)}{\tau_0+1}. \tag{EC.11}$$

To this end, fix some $\boldsymbol{\theta} \in \mathcal{P}^{KL}\left(\tilde{\mathcal{S}}, \sqrt{\frac{\log(1/\epsilon)}{\tau_0+2}}\right)$. Observe,

$$\begin{aligned}
\frac{\log(1/\epsilon)}{\tau_0+2} &\geq \sum_{i=1}^{d} \tilde{\mu}_{N,i} \log(\tilde{\mu}_{N,i}/\theta_i) && \text{(definition of } \mathcal{P}^{KL}) \\
&= \sum_{i=1}^{d} \tilde{\mu}_{N,i} \log(\tilde{\mu}_{N,i}/\theta_i) - \tilde{\mu}_{N,i} + \theta_i && \text{(since } \sum_{i=1}^{d} \tilde{\mu}_{N,i} = \sum_{i=1}^{d} \theta_i = 1) \\
&= \sum_{i=1}^{d} \frac{(\tilde{\mu}_{N,i} - \theta_i)^2}{2\tilde{\mu}_{N,i}} - \frac{(\tilde{\mu}_{N,i} - q_i)^3}{3\tilde{\mu}_{N,i}^2}, && \text{(EC.12)}
\end{aligned}$$

where the last line follows the multivariate mean-value theorem with $\mathbf{q} = \lambda\tilde{\boldsymbol{\mu}}_N + (1-\lambda)\boldsymbol{\theta}$ for some λ, $0 \leq \lambda \leq 1$. Since $\mathcal{P}^{KL}\left(\tilde{\mathcal{S}}, \sqrt{\frac{\log(1/\epsilon)}{\tau_0+2}}\right)$ is convex and $\tilde{\boldsymbol{\mu}}_N \in \mathcal{P}^{KL}\left(\tilde{\mathcal{S}}, \sqrt{\frac{\log(1/\epsilon)}{\tau_0+2}}\right)$, $\mathbf{q} \in \mathcal{P}^{KL}\left(\tilde{\mathcal{S}}, \sqrt{\frac{\log(1/\epsilon)}{\tau_0+2}}\right)$. Consequently,

$$\begin{aligned}
\|\tilde{\boldsymbol{\mu}}_N - \mathbf{q}\|_3 &\leq \|\tilde{\boldsymbol{\mu}}_N - \mathbf{q}\|_1 && \text{(Monotonicity of norms)} \\
&\leq \sqrt{\sum_{i=1}^{d} \tilde{\mu}_{N,i} \log(\tilde{\mu}_{N,i}/q_i) - \tilde{\mu}_{N,i} + q_i} && \text{(Plinsker's inequality)} \\
&\leq \sqrt{\frac{\log(1/\epsilon)}{\tau_0+2}}, && \left(\text{since } \mathbf{q} \in \mathcal{P}^{KL}\left(\tilde{\mathcal{S}}, \sqrt{\frac{\log(1/\epsilon)}{\tau_0+2}}\right)\right). \quad \text{(EC.13)}
\end{aligned}$$

Combining this last inequality with Eq. (EC.12), we have thus shown that

$$\sum_{i=1}^{d} \frac{(\tilde{\mu}_{N,i} - \theta_i)^2}{2\tilde{\mu}_{N,i}} \leq \frac{\log(1/\epsilon)}{\tau_0+2} + \frac{1}{3\tilde{\mu}_{N,min}^2}\left(\frac{\log(1/\epsilon)}{\tau_0+2}\right)^{3/2}, \quad \text{(EC.14)}$$

where $\tilde{\mu}_{N,min} \equiv \min_{i=1,\ldots,d} \tilde{\mu}_{N,i}$.

Comparing to Eq. (EC.11), it thus suffices to show that with $\mathbb{P}_{\mathcal{S}}$-probability tending to 1,

$$\frac{\log(1/\epsilon)}{\tau_0+2} + \frac{1}{3\tilde{\mu}_{N,min}}\left(\frac{\log(1/\epsilon)}{\tau_0+2}\right)^{3/2} \leq \frac{\log(1/\epsilon)}{\tau_0+1} \quad \iff \quad 3\sqrt{\frac{\tau_0+2}{\log(1/\epsilon)}}\left(\frac{\tau_0+2}{\tau_0+1} - 1\right) \geq \frac{1}{\tilde{\mu}_{N,min}}.$$

By the Law of Large Numbers, $\tilde{\boldsymbol{\mu}}_N \to_{\mathbb{P}_{\mathcal{S}}} \boldsymbol{\theta}^* > \mathbf{0}$ since $\boldsymbol{\theta}^* \in ri(\Theta)$. Note that $\tau_0 = O(N)$ and take limits to complete the proof. □

Proof of Thm. 10 Consider any $\mathcal{S}$ such that $\mathcal{P}(\mathcal{S}) - \boldsymbol{\mu}_N \subseteq \alpha\left(\mathcal{P}^*(\mathcal{S}, \frac{z_{1-\epsilon}}{\sqrt{N}}) - \boldsymbol{\mu}_N\right)$. Since $\mathcal{P}(\mathcal{S})$ is a credible region,

$$\begin{aligned}
1-\epsilon &\leq \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}_N \in \mathcal{P}(\mathcal{S}) - \boldsymbol{\mu}_N) \\
&\leq \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}_N \in \alpha\left(\mathcal{P}^*\left(\mathcal{S}, \frac{z_{1-\epsilon}}{\sqrt{N}}\right) - \boldsymbol{\mu}_N\right)\right) \\
&= \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(N^{-1}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}_N)^T \boldsymbol{\Sigma}_N^{-1}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}_N) \leq \frac{\alpha^2 z_{1-\epsilon}^2}{N}\right).
\end{aligned}$$

Since $\alpha < \frac{\sqrt{\chi^2_{r,1-\epsilon}}}{z_{1-\epsilon}}$, there exists $\delta > 0$ such that $\alpha^2 z^2_{1-\epsilon} \le \chi^2_{r,1-\epsilon-\delta} - \delta$. Fix such a δ. Then,

$$\begin{aligned}
\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(N^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N)^T\boldsymbol{\Sigma}_N^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N) \le \frac{\alpha^2 z^2_{1-\epsilon}}{N}\right) &\le \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(N^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N)^T\boldsymbol{\Sigma}_N^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N) \le \frac{\chi^2_{r,1-\epsilon-\delta}-\delta}{N}\right)\\
&\le \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left((\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N)^T\mathcal{I}(\boldsymbol{\theta}^*)(\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N) \le \frac{\chi^2_{r,1-\epsilon-\delta}}{N}\right)\\
&\quad + \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left((\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N)^T(N^{-1}\boldsymbol{\Sigma}_N^{-1}-\mathcal{I}(\boldsymbol{\theta}^*))(\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N) \le -\frac{\delta}{N}\right)\\
&\le \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}(\|\sqrt{N}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N)\|^2_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} \le \chi^2_{r,1-\epsilon-\delta}) + Z(\mathcal{S})\\
&\le \mathbb{P}(\|\tilde{\boldsymbol{\zeta}}\|^2_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} \le \chi^2_{r,1-\epsilon-\delta}) + R_N(\mathcal{S}) + Z(\mathcal{S}),
\end{aligned}$$

where $\tilde{\boldsymbol{\zeta}} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$, $R_N(\mathcal{S}) = \sup_{\mathcal{A}} \left|\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\sqrt{N}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N) \in \mathcal{A}\right) - \mathbb{P}(\tilde{\boldsymbol{\zeta}} \in \mathcal{A})\right|$ denotes the total variation distance from Thm. 6 for the realization $\mathcal{S}$ and $Z(\mathcal{S}) \equiv \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\left|(\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N)^T(N^{-1}\boldsymbol{\Sigma}_N^{-1}-\mathcal{I}(\boldsymbol{\theta}^*))(\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N)\right| > \frac{\delta}{N}\right)$.

The first probability is at most $1-\epsilon-\delta$. Thus, for any $\mathcal{S}$ such that Eq. (16) does not hold, $R_N(\mathcal{S}) + Z(\mathcal{S}) > \delta$.

Fix $t > 0$ such that $\mathbb{P}(\|\tilde{\boldsymbol{\zeta}}\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} > t) \le \delta/2$. From the *second* inequality in Lemma EC.2,

$$\begin{aligned}
Z(\mathcal{S}) &\le \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} \cdot \sqrt{\|N^{-1}\boldsymbol{\Sigma}_N^{-1}-\mathcal{I}(\boldsymbol{\theta}^*)\|_F} > \sqrt{\frac{\delta}{N}}\right)\\
&\le \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} > \frac{t}{\sqrt{N}}\right) + \mathbb{I}\left(\sqrt{\|N^{-1}\boldsymbol{\Sigma}_N^{-1}-\mathcal{I}(\boldsymbol{\theta}^*)\|_F} > \sqrt{\delta}/t\right)\\
&\le \mathbb{P}\left(\|\tilde{\boldsymbol{\zeta}}\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} > t\right) + R(\mathcal{S}) + \mathbb{I}\left(\sqrt{\|N^{-1}\boldsymbol{\Sigma}_N^{-1}-\mathcal{I}(\boldsymbol{\theta}^*)\|_F} > \sqrt{\delta}/t\right)\\
&\le \delta/2 + R_N(\mathcal{S}) + \mathbb{I}\left(\sqrt{\|N^{-1}\boldsymbol{\Sigma}_N^{-1}-\mathcal{I}(\boldsymbol{\theta}^*)\|_F} > \sqrt{\delta}/t\right).
\end{aligned}$$

Thus, for a randomly drawn $\tilde{\mathcal{S}}$,

$$\begin{aligned}
\mathbb{P}_{\mathcal{S}}(\text{Eq. (16) does not hold}) &\le \mathbb{P}_{\mathcal{S}}\left(R_N(\tilde{\mathcal{S}}) + Z(\tilde{\mathcal{S}}) > \delta\right)\\
&\le \mathbb{P}_{\mathcal{S}}\left(2R_N(\tilde{\mathcal{S}}) + \mathbb{I}\left(\sqrt{\|N^{-1}\tilde{\boldsymbol{\Sigma}}_N^{-1}-\mathcal{I}(\boldsymbol{\theta}^*)\|_F} > \sqrt{\delta}/t\right) > \delta/2\right)\\
&\le \mathbb{P}_{\mathcal{S}}\left(2R_N(\tilde{\mathcal{S}}) > \delta/4\right) + \mathbb{P}_{\mathcal{S}}\left(\mathbb{I}\left(\sqrt{\|N^{-1}\tilde{\boldsymbol{\Sigma}}_N^{-1}-\mathcal{I}(\boldsymbol{\theta}^*)\|_F} > \sqrt{\delta}/t\right) > \delta/4\right).
\end{aligned}$$

By A3, the first probability tends to zero, and since $N\tilde{\boldsymbol{\Sigma}}_N \to_{\mathbb{P}_{\mathcal{S}}} \mathcal{I}(\boldsymbol{\theta}^*)^{-1}$, the second probability tends to zero as well. This proves the first statement.

For the second statement, a standard Chernoff bound shows that $\frac{\sqrt{\chi^2_{r,1-\epsilon}}}{z_{1-\epsilon}} = \Omega(\sqrt{r})$ as $r \to \infty$. Thus, the size of $\mathcal{P}(\cdot)$ relative to $\mathcal{P}^*(\cdot, z_{1-\epsilon}/\sqrt{N})$ must grow with d, and $\mathcal{P}(\cdot)$ is not near-optimal. □

Proof of Thm. 11. Consider $\boldsymbol{\mu}_N + \mathbf{y} \in \mathcal{P}^*(\mathcal{S}, z_{1-\epsilon}/\sqrt{N})$. For $\alpha < \frac{\sqrt{\chi^2_{d-1,1-\epsilon}}}{z_{1-\epsilon}}$, we must show that $\boldsymbol{\mu}_N + \alpha\mathbf{y} \in \mathcal{P}^\phi\left(\mathcal{S}, \sqrt{\frac{\phi''(1)\chi^2_{d-1,1-\epsilon}}{2N}}\right)$ for sufficiently large N. Note that $\boldsymbol{\mu}_N + \mathbf{y} \in \mathcal{P}^*(\mathcal{S}, z_{1-\epsilon}/\sqrt{N})$ implies that

$$\frac{z^2_{1-\epsilon}}{\tau_0+1} \ge \sum_{i=1}^d \frac{y_i^2}{\mu_{N,i}} = \sum_{i=1}^d \left(\frac{y_i}{\sqrt{\mu_{N,i}}}\right)^2 \ge \max_{i=1,\ldots,d}\left(\frac{y_i}{\sqrt{\mu_{N,i}}}\right)^2. \tag{EC.15}$$

By Taylor's Theorem, there exists a function $r(t)$ such that $\lim_{t\to 0} r(t) = 0$ and $\phi(1+t) = \phi(1) + \phi'(1)t + \frac{1}{2}\phi''(1)t^2 + r(t)t^2$. Recall $\phi(1) = 0$, and write

$$\begin{aligned}\sum_{i=1}^{d} \mu_{N,i}\phi\left(\frac{\mu_{N,i}+\alpha y_i}{\mu_{N,i}}\right) &= \sum_{i=1}^{d} \mu_{N,i}\phi\left(1+\frac{\alpha y_i}{\mu_{N,i}}\right) \\ &= \sum_{i=1}^{d} \mu_{N,i}\cdot\phi'(1)\cdot\frac{\alpha y_i}{\mu_{N,i}} + \sum_{i=1}^{d}\mu_{N,i}\cdot\frac{1}{2}\phi''(1)\frac{\alpha^2 y_i^2}{\mu_{N,i}^2} + \sum_{i=1}^{d}\mu_{N,i}\cdot r\left(\frac{\alpha y_i}{\mu_{N,i}}\right)\cdot\frac{\alpha^2 y_i^2}{\mu_{N,i}^2} \\ &= \alpha\phi'(1)\sum_{i=1}^{d} y_i + \frac{\alpha^2}{2}\phi''(1)\sum_{i=1}^{d}\frac{y_i^2}{\mu_{N,i}} + \alpha^2\sum_{i=1}^{d} r\left(\frac{\alpha y_i}{\mu_{N,i}}\right)\frac{y_i^2}{\mu_{N,i}}.\end{aligned}$$

The first summation disappears since $\boldsymbol{\mu}_N + \mathbf{y} \in \mathcal{P}^*(\mathcal{S}, z_{1-\epsilon}/\sqrt{N})$ implies that $\mathbf{e}^T\mathbf{y} = 0$. We use the first inequality of Eq. (EC.15) to bound the second summation and the last inequality of Eq. (EC.15) to bound each term of the third summation, yielding

$$\sum_{i=1}^{d} \mu_{N,i}\phi\left(\frac{\mu_{N,i}+\alpha y_i}{\mu_{N,i}}\right) \;\le\; \frac{\alpha^2}{2}\phi''(1)\frac{z_{1-\epsilon}^2}{\tau_0+1} + \alpha^2\frac{z_{1-\epsilon}^2}{\tau_0+1}\sum_{i=1}^{d} r\left(\frac{\alpha y_i}{\mu_{N,i}}\right) \;=\; \frac{\phi''(1)\alpha^2 z_{1-\epsilon}^2}{2(\tau_0+1)}\left(1+\frac{2}{\phi''(1)}\sum_{i=1}^{d} r\left(\frac{\alpha y_i}{\mu_{N,i}}\right)\right).$$

Thus, to complete the proof, it remains to show that for N sufficiently large,

$$\frac{\phi''(1)\alpha^2 z_{1-\epsilon}^2}{2(\tau_0+1)}\left(1+\frac{2}{\phi''(1)}\sum_{i=1}^{d} r\left(\frac{\alpha y_i}{\mu_{N,i}}\right)\right) \le \frac{\phi''(1)\chi^2_{d-1,1-\epsilon}}{2N} \iff \frac{N}{(\tau_0+1)}\left(1+\frac{2}{\phi''(1)}\sum_{i=1}^{d} r\left(\frac{\alpha y_i}{\mu_{N,i}}\right)\right) \le \frac{\chi^2_{d-1,1-\epsilon}}{\alpha^2 z_{1-\epsilon}^2}.$$

As $N\to\infty$, $\frac{\alpha y_i}{\mu_{N,i}} \to 0$ for each i by Eq. (EC.15), and $\tau_0 \ge N$. Consequently, $\frac{N}{(\tau_0+1)}\left(1+\frac{2}{\phi''(1)}\sum_{i=1}^{d} r\left(\frac{\alpha y_i}{\mu_{N,i}}\right)\right)$ is less than 1 for N sufficiently large, and the result follows from the condition on α. □

Proof of Thm. 12 Suppose $\boldsymbol{\mu}_N + \mathbf{y} \in \mathcal{P}^*(\mathcal{S}, z_{1-\epsilon}/\sqrt{N})$. To prove the first statement, it suffices to prove that $\boldsymbol{\mu}_N + \lambda\mathbf{y} \in \Delta_d$ for any λ with $0 \le \lambda \le \frac{\sqrt{\tau_0+1}}{z_{1-\epsilon}}\min_i\sqrt{\mu_{N,i}}$. Note $\boldsymbol{\mu}_N + \mathbf{y} \in \mathcal{P}^*(\mathcal{S}, z_{1-\epsilon}/\sqrt{N})$ implies $\mathbf{e}^T\mathbf{y} = 0$, so it remains to prove $\boldsymbol{\mu}_N + \lambda\mathbf{y} \ge \mathbf{0}$. For any i such that $y_i \ge 0$, this is immediate. For i such that $y_i < 0$, note that as in the proof of Thm. 11, from the last inequality of Eq. (EC.15), $|y_i| \le \frac{z_{1-\epsilon}\sqrt{\mu_{N,i}}}{\sqrt{\tau_0+1}}$. Then, from the definition of λ,

$$\mu_{N,i} + \lambda y_i \ge \mu_{N,i} - |y_i|\frac{\sqrt{\tau_0+1}}{z_{1-\epsilon}}\min_j\sqrt{\mu_{N,j}} \ge \mu_{N,i}\left(1 - \frac{z_{1-\epsilon}}{\sqrt{\tau_0+1}}\frac{\sqrt{\tau_0+1}}{z_{1-\epsilon}}\cdot\frac{\min_j\sqrt{\mu_{N,j}}}{\sqrt{\mu_{N,i}}}\right) \ge \mu_{N,i}\left(1 - \frac{\min_j\sqrt{\mu_{N,j}}}{\sqrt{\mu_{N,i}}}\right) \ge 0,$$

proving the first statement. The second follows since $\tilde{\boldsymbol{\mu}}_N \to_{\mathbb{P}_\mathcal{S}} \boldsymbol{\theta}^* > \mathbf{0}$ implies that $\min_i\sqrt{\tilde{\boldsymbol{\mu}}_{N,i}} > 0$ with $\mathbb{P}_\mathcal{S}$-probability tending to 1. □

We present the following lemma, which is interesting in its own right. It establishes that the worst-case performance over certain ambiguity sets converges to the full-information performance *uniformly* over $\mathbf{x}$. Uniform approximation is the key property for establishing asymptotic consistency of optimization solutions.

LEMMA EC.3. *Suppose $g \in \mathcal{G}$ is equicontinuous in $\boldsymbol{\theta}$ for $\mathbf{x} \in \mathcal{C}$. Fix any sample path of $\mathcal{S}$ such that $(\boldsymbol{\mu}_N, N\boldsymbol{\Sigma}_N) \to (\boldsymbol{\theta}^*, I(\boldsymbol{\theta}^*)^{-1})$ as $N \to \infty$. Let α_N such that $\mathcal{P}(\mathcal{S}) - \boldsymbol{\mu}_N \subseteq \alpha_N \left(\mathcal{P}^*(\mathcal{S}, z_{1-\epsilon}/\sqrt{N}) - \boldsymbol{\mu}_N\right)$ and $\alpha_N = o(\sqrt{N})$. Then,*

$$\sup_{\mathbf{x}\in\mathcal{C}} \left| \sup_{\boldsymbol{\theta}\in\mathcal{P}(\mathcal{S})} g(\boldsymbol{\theta}, \mathbf{x}) - g(\boldsymbol{\theta}^*, \mathbf{x}) \right| \to 0.$$

Proof of Lemma. By definition of the supremum, for any $\epsilon > 0$, there exists a $\boldsymbol{\theta}_N \in \mathcal{P}(\mathcal{S})$ such that

$$\sup_{\mathbf{x}\in\mathcal{C}} \left| \sup_{\boldsymbol{\theta}\in\mathcal{P}(\mathcal{S})} g(\boldsymbol{\theta}, \mathbf{x}) - g(\boldsymbol{\theta}^*, \mathbf{x}) \right| \leq \delta + \sup_{\mathbf{x}\in\mathcal{C}} |g(\boldsymbol{\theta}_N, \mathbf{x}) - g(\boldsymbol{\theta}^*, \mathbf{x})|. \tag{EC.16}$$

Next, $\|\boldsymbol{\theta}_N - \boldsymbol{\theta}^*\| \leq \|\boldsymbol{\theta}_N - \boldsymbol{\mu}_N\| + \|\boldsymbol{\mu}_N - \boldsymbol{\theta}^*\|$. Using $\boldsymbol{\theta}_N \in \mathcal{P}(\mathcal{S})$ and the assumption on α_N, $N^{-1/2}\|\boldsymbol{\theta}_N - \boldsymbol{\mu}_N\|_{\boldsymbol{\Sigma}_N} \to 0$, and since $N\boldsymbol{\Sigma}_N \to \mathcal{I}(\boldsymbol{\theta}^*)$, this implies that $\|\boldsymbol{\theta}_N - \boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} \to 0$. Since, $\boldsymbol{\mu}_N \to \boldsymbol{\theta}^*$, we conclude that $\boldsymbol{\theta}_N \to \boldsymbol{\theta}^*$. It follows from the equicontinuity of g that for N sufficiently large, the supremum on the righthand side of Eq. (EC.16) is at most δ, and the requisite supremum is at most 2δ. Taking the limit as $\delta \to 0$ completes the proof. □

Proof of Thm. 13. We use Thms. 5.3 and 5.5 of Shapiro et al. (2014). Although these two theorems are proven for SAA, a careful reading shows that their proofs do not leverage the particular structure of SAA, but rather just the explicit conditions listed in the two theorems. In other words, it suffices to show that $\mathbf{P}$ and $\mathbf{P_N}$ satisfy the conditions of these two theorems to prove the result.

For Thm. 5.3, first note that $\mathcal{C}$ is compact by assumption and necessarily contains $\mathcal{O}^*$ and $\tilde{\mathcal{O}}_N$. The objective of $\mathbf{P}$ is finite since $g_0(\boldsymbol{\theta}^*, \cdot)$ is continuous in $\mathbf{x}$, and $\mathcal{C}$ is compact. That $\sup_{\theta\in\mathcal{P}(\mathcal{S})} g_0(\boldsymbol{\theta}, \mathbf{x})$ converges to $g_0(\boldsymbol{\theta}^*, \mathbf{x})$ uniformly in $\mathbf{x}$ follows from Lemma EC.3. We next show that $\tilde{\mathcal{O}}_N$ is non-empty. By assumption, there exists $\mathbf{x}_0 \in \{\mathbf{x} \in \mathcal{C} : g_l(\boldsymbol{\theta}^*, \mathbf{x}) < 0,\ l = 1, \ldots, L\}$. By Lemma EC.3, for N sufficiently large, $\mathbf{x}_0$ will be feasible in $\mathbf{P_N}$, whereby $\tilde{\mathcal{O}}_N$ is non-empty. All the conditions of Thm. 5.3 are met.

For the remaining conditions for Thm 5.5, fix a sample path where $(\boldsymbol{\mu}_N, N\boldsymbol{\Sigma}_N) \to (\boldsymbol{\theta}^*, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$. We argue along this path. For Thm 5.5 condition a), let $\mathbf{x}_N$ be feasible in $\mathbf{P_N}$ and suppose $\mathbf{x}_N \to \mathbf{x}_\infty$. By compactness, $\mathbf{x}_\infty \in \mathcal{C}$. We claim $\mathbf{x}_\infty$ is feasible in $\mathbf{P}$. Fix any $\delta > 0$. For N sufficiently large, continuity of g_l in $\mathbf{x}$ implies $|g_l(\boldsymbol{\theta}^*, \mathbf{x}_\infty) - g_l(\boldsymbol{\theta}^*, \mathbf{x}_N)| \leq \delta$. Then,

$$\begin{aligned} g_l(\boldsymbol{\theta}^*, \mathbf{x}_\infty) &= (g_l(\boldsymbol{\theta}^*, \mathbf{x}_\infty) - g_l(\boldsymbol{\theta}^*, \mathbf{x}_N)) + \left(g_l(\boldsymbol{\theta}^*, \mathbf{x}_N) - \sup_{\boldsymbol{\theta}\in\mathcal{P}(\tilde{\mathcal{S}})} g(\boldsymbol{\theta}, \mathbf{x}_N) \right) + \sup_{\boldsymbol{\theta}\in\mathcal{P}(\mathcal{S})} g(\boldsymbol{\theta}, \mathbf{x}_N) \\ &\leq \delta + \sup_{\mathbf{x}\in\mathcal{C}} \left| g_l(\boldsymbol{\theta}^*, \mathbf{x}) - \sup_{\boldsymbol{\theta}\in\mathcal{P}(\mathcal{S})} g(\boldsymbol{\theta}, \mathbf{x}) \right|, \end{aligned}$$

where we have bound the last term by 0 since $\mathbf{x}_N$ is feasible in $\mathbf{P_N}$. By Lemma EC.3, for N sufficiently large, the supremum is at most δ. Taking the limit as $\delta \to 0$ proves $\mathbf{x}_\infty$ is feasible in $\mathbf{P}$.

For Thm 5.5 condition b), let $\mathbf{x}^*$ be optimal for $\mathbf{P}$. Suppose by contradiction that there exists $\epsilon > 0$ such that every convergent sequence of robust feasible solutions converges to a point more than ϵ far away from $\mathbf{x}^*$. By Condition iii) (from the statement of this theorem) there exists an $\mathbf{x}_0$ such that $\|\mathbf{x}_0 - \mathbf{x}^*\| < \epsilon$ and $\mathbf{x}_0 \in \{\mathbf{x} \in \mathcal{C} : g_l(\boldsymbol{\theta}^*, \mathbf{x}) < 0,\ l = 1, \ldots, L\}$. As above, for all N sufficiently large, $\mathbf{x}_0$ is feasible in $\mathbf{P_N}$. This yields a contradiction.

The result now follows from Thm 5.5 of Shapiro et al. (2014).

□

Proof of Thm. 14. The "only if" direction is immediate since confidence regions satisfy the frequentist guarantee for all measurable g. For the "if" direction, note that for this choice of g, the set $\{(\mathbf{v}, \delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}))) : \mathbf{v} \in \mathbb{R}^d\} \subseteq \mathcal{X}(\mathcal{P}(\mathcal{S}))$ for all $\mathcal{S}$, whereby from the frequentist guarantee, $\mathbb{P}_{\mathcal{S}}(\mathbf{v}^T\boldsymbol{\theta}^* \le \delta^*(\mathbf{v}|\ \mathcal{P}(\tilde{\mathcal{S}}))\ \forall \mathbf{v} \in \mathbb{R}^d) \ge 1 - \epsilon$. Meanwhile, since for any $\mathcal{S}$, $\mathcal{P}(\mathcal{S})$ is closed and convex, $\mathbf{v}^T\boldsymbol{\theta}^* \le \delta^*(\mathbf{v}|\ \mathcal{P}(\mathcal{S}), \forall \mathbf{v} \in \mathbb{R}^d$ if and only if $\boldsymbol{\theta}^* \in \mathcal{P}(\mathcal{S})$. Combining yields the theorem. □

Proof of Thm. 15. Let η denote a chi-squared random variable with r degrees of freedom. From the condition on α, $\mathbb{P}(\sqrt{\eta} \le \alpha z_{1-\epsilon}) < 1 - \epsilon$. Thus, there exists $\delta > 0$, such that $\mathbb{P}(\sqrt{\eta} \le \alpha z_{1-\epsilon} + \delta) \le 1 - \epsilon - \delta$. Fix such a δ.

For any $\mathcal{S}$ such that $\mathcal{P}(\mathcal{S}) - \boldsymbol{\mu}_N \subseteq \alpha(\mathcal{P}^*(\mathcal{S}, z_{1-\epsilon}/\sqrt{N}) - \boldsymbol{\mu}_N)$, $\boldsymbol{\theta}^* \in \mathcal{P}(\mathcal{S})$ implies

$$\alpha z_{1-\epsilon} \ge \|\boldsymbol{\theta}^* - \boldsymbol{\mu}_N\|_{\boldsymbol{\Sigma}_N} \ge \|\boldsymbol{\theta}^* - \boldsymbol{\theta}^{MLE}\|_{\boldsymbol{\Sigma}_N} - \|\boldsymbol{\theta}^{MLE} - \boldsymbol{\mu}_N\|_{\boldsymbol{\Sigma}_N}$$

by the triangle inequality. Hence,

$$\begin{aligned}\mathbb{P}_{\mathcal{S}}\left(\boldsymbol{\theta}^* \in \mathcal{P}(\tilde{\mathcal{S}}) \text{ and } \mathcal{P}(\tilde{\mathcal{S}}) - \tilde{\boldsymbol{\mu}}_N \subseteq \alpha(\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}) - \tilde{\boldsymbol{\mu}}_N)\right) &\le \mathbb{P}_{\mathcal{S}}\left(\|\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}^{MLE}\|_{\tilde{\boldsymbol{\Sigma}}_N} - \|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} \le \alpha z_{1-\epsilon}\right) \\ &\le \mathbb{P}_{\mathcal{S}}\left(\|\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}^{MLE}\|_{\tilde{\boldsymbol{\Sigma}}_N} \le \alpha z_{1-\epsilon} + \delta\right) + \mathbb{P}_{\mathcal{S}}(\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} > \delta).\end{aligned}$$

By A3, $N\tilde{\boldsymbol{\Sigma}}_N \to_{\mathbb{P}_{\mathcal{S}}} \mathcal{I}(\boldsymbol{\theta}^*)^{-1}$, and by Eq. (20), $\sqrt{N}(\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}^{MLE})$ converges in distribution to $\mathcal{N}(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$. By the continuous mapping theorem,

$$\mathbb{P}_{\mathcal{S}}\left(\|\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}^{MLE}\|_{\tilde{\boldsymbol{\Sigma}}_N} \le \alpha z_{1-\epsilon} + \delta\right) \to \mathbb{P}(\sqrt{\eta} \le \alpha z_{1-\epsilon} + \delta) \le 1 - \epsilon - \delta.$$

Thus,

$$\mathbb{P}_{\mathcal{S}}\left(\boldsymbol{\theta}^* \in \mathcal{P}(\tilde{\mathcal{S}}) \text{ and } \mathcal{P}(\tilde{\mathcal{S}}) - \tilde{\boldsymbol{\mu}}_N \subseteq \alpha(\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}) - \tilde{\boldsymbol{\mu}}_N)\right) \quad \le 1 - \epsilon - \delta + \mathbb{P}_{\mathcal{S}}(\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} > \delta).$$

Now, since $\mathcal{P}(\cdot)$ is a confidence region,

$$
\begin{aligned}
1-\epsilon &\leq \mathbb{P}_{\mathcal{S}}(\boldsymbol{\theta}^* \in \mathcal{P}(\tilde{\mathcal{S}})) \\
&= \mathbb{P}_{\mathcal{S}}(\boldsymbol{\theta}^* \in \mathcal{P}(\tilde{\mathcal{S}}),\ \text{and } \mathcal{P}(\tilde{\mathcal{S}}) - \tilde{\boldsymbol{\mu}}_N \subseteq \alpha(\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}) - \tilde{\boldsymbol{\mu}}_N)) \\
&\quad + \mathbb{P}_{\mathcal{S}}(\boldsymbol{\theta}^* \in \mathcal{P}(\tilde{\mathcal{S}}) \text{ and } \mathcal{P}(\tilde{\mathcal{S}}) - \tilde{\boldsymbol{\mu}}_N \not\subseteq \alpha(\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}) - \tilde{\boldsymbol{\mu}}_N) \\
&\leq 1-\epsilon-\delta + \mathbb{P}_{\mathcal{S}}(\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} > \delta) + \mathbb{P}_{\mathcal{S}}\left(\boldsymbol{\theta}^* \in \mathcal{P}(\tilde{\mathcal{S}}),\ \mathcal{P}(\tilde{\mathcal{S}}) - \tilde{\boldsymbol{\mu}}_N \not\subseteq \alpha(\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}) - \tilde{\boldsymbol{\mu}}_N)\right). \\
&\leq 1-\epsilon-\delta + \mathbb{P}_{\mathcal{S}}(\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} > \delta) + \mathbb{P}_{\mathcal{S}}\left(\mathcal{P}(\tilde{\mathcal{S}}) - \tilde{\boldsymbol{\mu}}_N \not\subseteq \alpha(\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}) - \tilde{\boldsymbol{\mu}}_N)\right).
\end{aligned}
$$

Rearranging shows that

$$
\mathbb{P}_{\mathcal{S}}\left(\mathcal{P}(\tilde{\mathcal{S}}) - \tilde{\boldsymbol{\mu}}_N \not\subseteq \alpha(\mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}/\sqrt{N}) - \tilde{\boldsymbol{\mu}}_N)\right) \geq \delta - \mathbb{P}_{\mathcal{S}}(\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} > \delta).
$$

By assumption, $\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} \to_{\mathbb{P}_{\mathcal{S}}} 0$. Thus, taking the limit as $N \to \infty$ proves the theorem. □

Proof of Thm. 16 Define $\mathbf{x}_N \equiv \sup_{\boldsymbol{\theta} \in \mathcal{P}(\mathcal{S})} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}$. Consider any $\mathcal{S}$ such that

$$
\mathcal{P}(\mathcal{S}) - \boldsymbol{\mu}_N \subseteq \alpha\left(\mathcal{P}^*\left(\mathcal{S}, z_{1-\epsilon}/\sqrt{N}\right) - \boldsymbol{\mu}_N\right). \tag{EC.17}
$$

We will bound $\mathbf{x}_N$ for such $\mathcal{S}$. In particular, for $\boldsymbol{\theta} \in \mathcal{P}(\mathcal{S})$,

$$
\begin{aligned}
\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} &\leq \|\boldsymbol{\theta} - \boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} + \|\boldsymbol{\mu}_N - \boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} \\
&= \|\boldsymbol{\theta} - \boldsymbol{\mu}_N\|_{N\boldsymbol{\Sigma}_N} + \left(\|\boldsymbol{\theta} - \boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} - \|\boldsymbol{\theta} - \boldsymbol{\mu}_N\|_{N\boldsymbol{\Sigma}_N}\right) + \|\boldsymbol{\mu}_N - \boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} \\
&\leq \|\boldsymbol{\theta} - \boldsymbol{\mu}_N\|_{N\boldsymbol{\Sigma}_N} + \|\boldsymbol{\theta} - \boldsymbol{\mu}_N\|_{N\boldsymbol{\Sigma}_N} \sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*) - N^{-1}\boldsymbol{\Sigma}_N^{-1}\|_F} + \|\boldsymbol{\mu}_N - \boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}},
\end{aligned}
$$

where we have used Lemma EC.2. Note that

$$
\|\boldsymbol{\theta} - \boldsymbol{\mu}_N\|_{N\boldsymbol{\Sigma}_N} = N^{-1/2}\|\boldsymbol{\theta} - \boldsymbol{\mu}_N\|_{\boldsymbol{\Sigma}_N} \leq N^{-1/2}\alpha z_{1-\epsilon}
$$

by Eq. (EC.17). Combining with the above yields,

$$
\mathbf{x}_N = \sup_{\boldsymbol{\theta} \in \mathcal{P}(\mathcal{S})} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} \leq \frac{\alpha z_{1-\epsilon}}{\sqrt{N}} \left(1 + \sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*) - N^{-1}\boldsymbol{\Sigma}_N^{-1}\|_F}\right) + \|\boldsymbol{\mu}_N - \boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}
$$

whenever Eq. (EC.17) holds.

By construction, $\mathbf{x}_N$ is robust feasible and, hence, satisfies the posterior guarantee for any $\mathcal{S}$. By the triangle inequality, $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} \geq \|\boldsymbol{\theta} - \boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} - \|\boldsymbol{\mu}_N - \boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}$. Substituting this inequality and our bound $\mathbf{x}_N$ into the posterior guarantee yields for any $\mathcal{S}$ satisfying Eq. (EC.17),

$$
1-\epsilon \leq \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}} \leq \mathbf{x}_N\right)
$$

$$\leq \mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}-\|\boldsymbol{\mu}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\leq\frac{\alpha z_{1-\epsilon}}{\sqrt{N}}\left(1+\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)-N^{-1}\boldsymbol{\Sigma}_N^{-1}\|_F}\right)+\|\boldsymbol{\mu}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\right)$$

$$=\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\sqrt{N}\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\leq\alpha z_{1-\epsilon}\left(1+\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)-N^{-1}\boldsymbol{\Sigma}_N^{-1}\|_F}+\frac{2\sqrt{N}}{\alpha z_{1-\epsilon}}\|\boldsymbol{\mu}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\right)\right).$$

Since $\alpha z_{1-\epsilon}<\sqrt{\chi^2_{r,1-\epsilon}}$, there exists $\delta>0$ such that $(1+2\delta)\alpha z_{1-\epsilon}\leq\sqrt{\chi^2_{r,1-\epsilon-\delta}}$. Fix such a δ. Then, write

$$\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\sqrt{N}\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\leq\alpha z_{1-\epsilon}\left(1+\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)-N^{-1}\boldsymbol{\Sigma}_N^{-1}\|_F}+\frac{2\sqrt{N}}{\alpha z_{1-\epsilon}}\|\boldsymbol{\mu}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\right)\right)$$

$$\leq\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\sqrt{N}\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\leq\sqrt{\chi^2_{r,1-\epsilon-\delta}}\right)+\mathbb{I}\left(1+\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)-N^{-1}\boldsymbol{\Sigma}_N^{-1}\|_F}+\frac{2\sqrt{N}}{\alpha z_{1-\epsilon}}\|\boldsymbol{\mu}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\geq\frac{\sqrt{\chi^2_{r,1-\epsilon-\delta}}}{\alpha z_{1-\epsilon}}\right)$$

$$\leq\mathbb{P}_{\tilde{\boldsymbol{\theta}}|\mathcal{S}}\left(\sqrt{N}\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\mu}_N\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\leq\sqrt{\chi^2_{r,1-\epsilon-\delta}}\right)+\mathbb{I}\left(\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)-N^{-1}\boldsymbol{\Sigma}_N^{-1}\|_F}+\frac{2\sqrt{N}}{\alpha z_{1-\epsilon}}\|\boldsymbol{\mu}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\geq 2\delta\right)$$

$$\leq\mathbb{P}(\|\tilde{\boldsymbol{\zeta}}\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\leq\sqrt{\chi^2_{r,1-\epsilon-\delta}})+R(\mathcal{S})+\mathbb{I}\left(\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)-N^{-1}\boldsymbol{\Sigma}_N^{-1}\|_F}\geq\delta\right)+\mathbb{I}\left(\frac{2\sqrt{N}}{\alpha z_{1-\epsilon}}\|\boldsymbol{\mu}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\geq\delta\right)$$

$$\leq 1-\epsilon-\delta+R(\mathcal{S})+\mathbb{I}\left(\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)-N^{-1}\boldsymbol{\Sigma}_N^{-1}\|_F}\geq\delta\right)+\mathbb{I}\left(\sqrt{N}\|\boldsymbol{\mu}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\geq\frac{\delta\alpha z_{1-\epsilon}}{2}\right),$$

where $\tilde{\boldsymbol{\zeta}}\sim\mathcal{N}(\mathbf{0},\mathcal{I}(\boldsymbol{\theta}^*)^{-1})$ and $R(\mathcal{S})$ is the total variation distance from Thm. 6. Thus, we have shown that for any $\mathcal{S}$ that satisfies Eq. (EC.17),

$$\delta\leq R(\mathcal{S})+\mathbb{I}\left(\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)-N^{-1}\boldsymbol{\Sigma}_N^{-1}\|_F}\geq\delta\right)+\mathbb{I}\left(\sqrt{N}\|\boldsymbol{\mu}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\geq\frac{\delta\alpha z_{1-\epsilon}}{2}\right).$$

This implies for a random $\tilde{\mathcal{S}}$,

$$\mathbb{P}_{\mathcal{S}}\left(\tilde{\mathcal{S}}\text{ satisfies Eq. (EC.17) }\right)$$

$$\leq\mathbb{P}_{\mathcal{S}}\left(R(\tilde{\mathcal{S}})+\mathbb{I}\left(\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)-N^{-1}\tilde{\boldsymbol{\Sigma}}_N^{-1}\|_F}\geq\delta\right)+\mathbb{I}\left(\sqrt{N}\|\tilde{\boldsymbol{\mu}}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\geq\frac{\delta\alpha z_{1-\epsilon}}{2}\right)\geq\delta\right)$$

$$\leq\mathbb{P}_{\mathcal{S}}\left(R(\tilde{\mathcal{S}})>\frac{\delta}{2}\right)+\mathbb{P}_{\mathcal{S}}\left(\sqrt{\|\mathcal{I}(\boldsymbol{\theta}^*)-N^{-1}\tilde{\boldsymbol{\Sigma}}_N^{-1}\|_F}\geq\delta\right)+\mathbb{P}_{\mathcal{S}}\left(\sqrt{N}\|\tilde{\boldsymbol{\mu}}_N-\boldsymbol{\theta}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)^{-1}}\geq\frac{\delta\alpha z_{1-\epsilon}}{2}\right).$$

By A3, the first two probabilities tend to zero, and the third probability tends to zero by assumption. This completes the proof. □

Proof of Thm. 17 For the first part, note that $\tilde{\mathbf{x}}\in\mathcal{X}(\mathcal{P}(\tilde{\mathcal{S}}))$, $\mathbb{P}$-a.s. Fix any $\mathbf{x}\in\mathcal{X}(\mathcal{P}(\mathcal{S}))$. Then,

$$\begin{aligned}\mathbb{P}\left(g(\tilde{\boldsymbol{\theta}},\tilde{\mathbf{x}})\leq 0\mid\tilde{\mathcal{S}}=\mathcal{S},\ \tilde{\mathbf{x}}=\mathbf{x}\right)&=\mathbb{P}\left(g(\tilde{\boldsymbol{\theta}},\mathbf{x})\leq 0\mid\tilde{\mathcal{S}}=\mathcal{S},\ \tilde{\mathbf{x}}=\mathbf{x}\right)\\&=\mathbb{P}\left(g(\tilde{\boldsymbol{\theta}},\mathbf{x})\leq 0\mid\tilde{\mathcal{S}}=\mathcal{S}\right)\qquad\text{(since }\tilde{\mathbf{x}}\perp\!\!\!\perp\tilde{\boldsymbol{\theta}}|\tilde{\mathcal{S}})\\&\geq 1-\epsilon,\end{aligned}$$

where the last line follows because $\mathbf{x}\in\mathcal{X}(\mathcal{P}(\mathcal{S}))$ and $\mathcal{P}(\cdot)$ satisfies the posterior feasibility guarantee. Now take expectations of both sides, first with respect to $\tilde{\mathbf{x}}$ and then with respect to $\tilde{\mathcal{S}}$ to prove the claim.

For the second part, we construct an instance explicitly. Suppose that the prior distribution of $\tilde{\theta}$ is $\mathcal{N}(0,1/2)$, and we observe a single observation $\tilde{\mathcal{S}} = \{\tilde{\xi}\}$ where $\tilde{\xi}|\tilde{\theta} \sim \mathcal{N}(\tilde{\theta}, 1/2)$. A standard computation shows that under the posterior distribution $\mathbb{P}_{\tilde{\theta}|\tilde{\mathcal{S}}}$, $\tilde{\theta}|\tilde{\mathcal{S}} \sim \mathcal{N}(\tilde{\xi}/2, 1)$. In particular, $\mathrm{VaR}^{1-\epsilon}_{\mathbb{P}_{\tilde{\theta}|\tilde{\mathcal{S}}}}(v) = v\tilde{\xi}/2 + z_{1-\epsilon}|v|$ is convex and the set $\mathcal{P}^*(\mathcal{S}, z_{1-\epsilon}) = \{\tilde{\theta} : |\tilde{\theta} - \tilde{\xi}/2| \le z_{1-\epsilon}\}$ is the optimal Bayesian set, for any finite N. In particular, $\delta^*\left(v|\mathcal{P}^*(\mathcal{S}, z_{1-\epsilon})\right) = v\tilde{\xi}/2 + z_{1-\epsilon}|v|$.

Now let $g(\theta,(v,t)) = v\theta - t$. Consider the instance of $\mathbf{P_N}$ for data $\tilde{\mathcal{S}}$,

$$\begin{aligned} \min_{v,t} \quad & 0 \\ \text{s.t.} \quad & g(\theta,(v,t)) \le 0, \quad \forall \theta \in \mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon}) \\ & -1 \le v \le 1. \end{aligned}$$

Notice this problem admits multiple solutions. Let $\tilde{v} = \mathrm{sgn}(\tilde{\theta} - \tilde{\xi}/2)$, and consider the solution $\tilde{x} = \left(\tilde{v}, \delta^*\left(\tilde{v} \mid \mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon})\right)\right)$, which by construction is feasible, and hence optimal for the above problem. Notice, $\tilde{x} \not\perp\!\!\!\perp \tilde{\theta} \mid \tilde{\mathcal{S}}$ since it explicitly depends on the value of $\tilde{\theta}$.

Finally, compute directly,

$$\begin{aligned} \mathbb{P}\left(g(\tilde{\theta}, \tilde{x}) \le 0\right) &= \mathbb{P}\left(\tilde{v}\tilde{\theta} \le \tilde{v}\tilde{\xi}/2 + z_{1-\epsilon}|\tilde{v}|\right) \\ &= \mathbb{P}\left(|\tilde{\theta} - \tilde{\xi}/2| \le z_{1-\epsilon}\right) && \text{(by definition of } |\tilde{v}|) \\ &= \mathbb{E}\left[\mathbb{P}\left(|\tilde{\theta} - \tilde{\xi}/2| \le z_{1-\epsilon} \mid \tilde{\xi}\right)\right] \\ &= 1 - 2\epsilon, \end{aligned}$$

where we have used the fact that $\tilde{\theta}|\tilde{\xi} \sim \mathcal{N}(\tilde{\xi}/2, 1)$ to evaluate the inner probability. Note this quantity is strictly less than $1-\epsilon$ to complete the proof.

□

Proof of Thm. 18. Restrict attention to the l^{th} constraint, and, without loss of generality, write $g_l(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{v}^T\boldsymbol{\theta} - t$, where $(\mathbf{v}, t)$ are deterministic functions of $\mathbf{x}$. Let $\mathbf{v}_N, t_N$ be these deterministic functions evaluated at $\mathbf{x}_N$ and $\tilde{\mathbf{v}}_N, \tilde{t}_N, \tilde{\mathbf{x}}_N$ be their random counterparts . Thus, we must bound $\limsup \mathbb{P}_{\mathcal{S}}(\tilde{\mathbf{v}}_N^T\boldsymbol{\theta}^* > \tilde{t}_N)$.

Intuitively, since $\mathcal{P}(\mathcal{S})$ satisfies a posterior guarantee, it *eventually* contains a small contraction of the asymptotically optimal set. To this end, fix any $\delta > 0$, let $\epsilon' = \epsilon + \delta$, and consider the contraction $\mathcal{P}^*(\mathcal{S}, z_{1-\epsilon'}/\sqrt{N})$. Decompose

$$\mathbb{P}_{\mathcal{S}}(\tilde{\mathbf{v}}_N^T\boldsymbol{\theta}^* > \tilde{t}_N) = \mathbb{P}_{\mathcal{S}}\left(\tilde{\mathbf{v}}_N^T\boldsymbol{\theta}^* > \tilde{t}_N,\ \mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon'}/\sqrt{N}) \not\subseteq \mathcal{P}(\tilde{\mathcal{S}})\right) + \mathbb{P}_{\mathcal{S}}\left(\tilde{\mathbf{v}}_N^T\boldsymbol{\theta}^* > \tilde{t}_N,\ \mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon'}/\sqrt{N}) \subseteq \mathcal{P}(\tilde{\mathcal{S}})\right).$$

By Thm. 7, the first probability tends to zero since $\mathcal{P}(\cdot)$ satisfies a posterior feasibility guarantee. Thus, we focus on the second probability.

If $\mathcal{P}^*(\mathcal{S}, z_{1-\epsilon'}/\sqrt{N}) \subseteq \mathcal{P}(\mathcal{S})$ and $\mathbf{x}_N$ is robust feasible, then

$$t_N \geq \sup_{\boldsymbol{\theta} \in \mathcal{P}(\mathcal{S})} \mathbf{v}_N^T \boldsymbol{\theta} \quad \geq \quad \sup_{\boldsymbol{\theta} \in \mathcal{P}^*(\mathcal{S}, z_{1-\epsilon'}/\sqrt{N})} \mathbf{v}_N^T \boldsymbol{\theta} \quad = \quad \boldsymbol{\mu}_N^T \mathbf{v}_N + z_{1-\epsilon'} \|\mathbf{v}_N\|_{\boldsymbol{\Sigma}_N^{-1}}.$$

In particular,

$$\mathbf{v}_N^T \boldsymbol{\theta}^* > t_N \implies \mathbf{v}_N^T \boldsymbol{\mu}_N + \mathbf{v}_N^T (\boldsymbol{\theta}^* - \boldsymbol{\mu}_N) > t_N \implies \mathbf{v}_N^T (\boldsymbol{\theta}^* - \boldsymbol{\mu}_N) > z_{1-\epsilon'} \|\mathbf{v}_N\|_{\boldsymbol{\Sigma}_N^{-1}}.$$

Thus, $\mathbb{P}_{\mathcal{S}} \left(\tilde{\mathbf{v}}_N^T \boldsymbol{\theta}^* > \tilde{t}_N,\ \mathcal{P}^*(\tilde{\mathcal{S}}, z_{1-\epsilon'}/\sqrt{N}) \subseteq \mathcal{P}(\tilde{\mathcal{S}}) \right) \leq \mathbb{P}_{\mathcal{S}}(\tilde{\mathbf{v}}_N^T (\boldsymbol{\theta}^* - \tilde{\boldsymbol{\mu}}_N) > z_{1-\epsilon'} \|\tilde{\mathbf{v}}_N\|_{\boldsymbol{\Sigma}_N^{-1}})$. Decomposing $\boldsymbol{\theta}^* - \tilde{\boldsymbol{\mu}}_N = \boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}^{MLE} + \tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N$, write, for $\delta > 0$,

$$\mathbb{P}_{\mathcal{S}}(\tilde{\mathbf{v}}_N^T (\boldsymbol{\theta}^* - \tilde{\boldsymbol{\mu}}_N) > z_{1-\epsilon'} \|\tilde{\mathbf{v}}_N\|_{\boldsymbol{\Sigma}_N^{-1}}) \leq \mathbb{P}_{\mathcal{S}} \left(\frac{\tilde{\mathbf{v}}_N^T (\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N)}{\|\tilde{\mathbf{v}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}}} > \delta \right) + \mathbb{P}_{\mathcal{S}} \left(\frac{\tilde{\mathbf{v}}_N^T (\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}^{MLE})}{\|\tilde{\mathbf{v}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}}} > z_{1-\epsilon'} - \delta \right). \tag{EC.18}$$

By the Cauchy-Schwarz inequality, the first probability is bounded by $\mathbb{P}_{\mathcal{S}}(\|\tilde{\boldsymbol{\theta}}^{MLE} - \tilde{\boldsymbol{\mu}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N} > \delta)$, which tends to zero by Condition iii) of the theorem.

Finally, for the remaining probability, by Thm. 13 and Condition v), $(\tilde{\mathbf{v}}_N, \tilde{t}_N) \to_{\mathbb{P}_{\mathcal{S}}} (\mathbf{v}^*, t^*)$ for some constants $(\mathbf{v}^*, t^*)$. Moreover, $N\tilde{\boldsymbol{\Sigma}}_N \to_{\mathbb{P}_{\mathcal{S}}} \mathcal{I}(\boldsymbol{\theta}^*)^{-1}$. Thus, $\frac{\tilde{\mathbf{v}}_N}{\sqrt{N}\|\tilde{\mathbf{v}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}}} \to_{\mathbb{P}_{\mathcal{S}}} \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|_{\mathcal{I}(\boldsymbol{\theta}^*)}}$ by the continuous mapping theorem. Combine this observation with Eq. (20) to conclude that

$$\frac{\tilde{\mathbf{v}}_N^T (\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}^{MLE})}{\|\tilde{\mathbf{v}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}}} = \frac{\tilde{\mathbf{v}}_N^T}{\sqrt{N}\|\tilde{\mathbf{v}}_N\|_{\tilde{\boldsymbol{\Sigma}}_N^{-1}}} \sqrt{N} (\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}^{MLE})$$

converges in distribution to a $\mathcal{N}(0,1)$ random variable.

In summary, we have shown that for N sufficiently large and $\tilde{y}$ a standard normal random variable,

$$\mathbb{P}_{\mathcal{S}}(\tilde{\mathbf{v}}_N^T \boldsymbol{\theta}^* > \tilde{t}_N) \leq \mathbb{P}(\tilde{y} > z_{1-\epsilon-\delta} - \delta) + 3\delta.$$

Taking $\delta \to 0$, the righthand side can be made arbitrarily close to $\mathbb{P}(\tilde{y} > z_{1-\epsilon}) = \epsilon$, proving the theorem. □

Appendix E: Additional Tables and Plots

E.1. Non-Uniform $\boldsymbol{\theta}^*$

In this appendix we study the performance of our sets when the true distribution is non-uniform. Specifically, we cluster the 10 most recent years of data (Jan. 2005 to Dec. 2014, 120 observations) into 36 clusters using hierarchical clustering (Friedman et al. 2001). We take the true distribution to be supported on the centroid of each cluster with probability equal to the proportion of observations in that cluster. Intuitively, these

clusters represent typical market environments over this time frame. Since some clusters are much larger than others, the resulting probabilities are non-uniform, ranging from 0.1 to 0.008 (see left panel of Fig. EC.2). Many of the small clusters correspond to large returns (or losses) in some of the asset classes (see right panel of Fig. EC.2). These rare scenarios with large losses/gains create a particularly challenging environment for data-driven portfolio allocation.

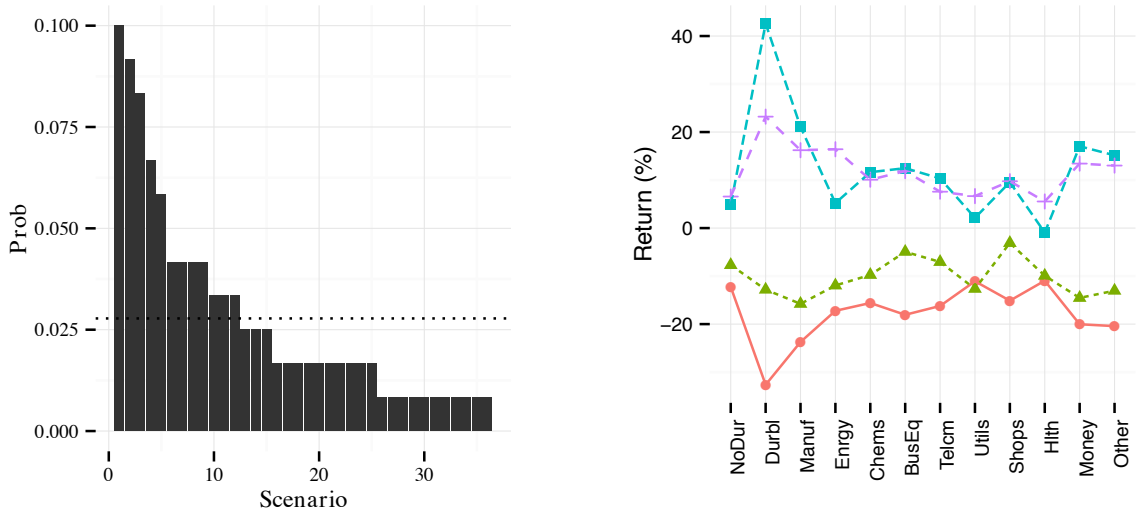

Figure EC.2 The left panel shows the proportion of observations in each cluster, i.e., θ^*. The dotted line corresponds to $1/36$, i.e., the uniform distribution. The right panel shows the centroids of some of the smallest clusters.

We repeat the experiment of Sec. 6.1 with this distribution, i.e., we repeatedly draw N data points from this distribution, form each of our portfolios from these data, and then record the performance with respect to the true distribution. Fig. EC.3 displays the results, which are largely similar to those of Sec. 6.1.

E.2. Auxiliary Results from Secs. 6.1 and 6.3 to 6.5

This section contains a number of additional plots for experiments run in Sections Secs. 6.1 and 6.3 to 6.5. Unless otherwise specified, experimental conditions are as described in the corresponding section.

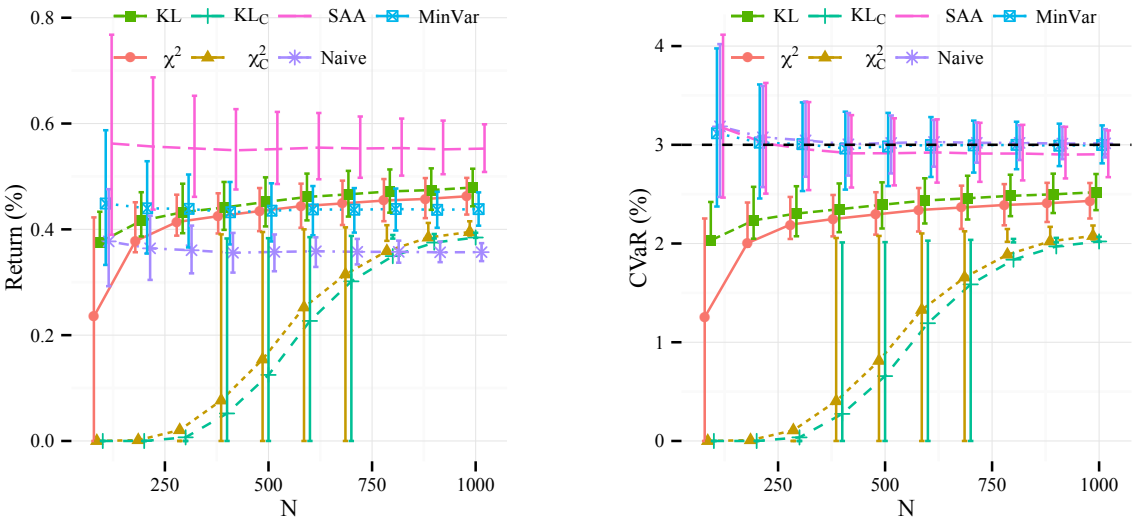

Figure EC.3 **The return and risk for portfolios corresponding to various ambiguity sets from Sec. 6.2.**

	$\|\mathbf{x}\| = 0$		St. Dev. of $\|\mathbf{x}\|$						
N	χ^2_C	KL_C	χ^2	χ^2_C	KL	KL_C	MinVar	Naive	SAA
100	1.00	1.00	0.04	0.00	0.04	0.00	0.10	0.02	0.12
200	0.91	0.99	0.05	0.08	0.05	0.03	0.07	0.01	0.10
300	0.51	0.81	0.03	0.13	0.03	0.11	0.04	0.01	0.06
400	0.13	0.27	0.04	0.10	0.04	0.13	0.04	0.01	0.06
500	0.03	0.06	0.04	0.06	0.04	0.08	0.04	0.01	0.06
600	0.00	0.01	0.04	0.04	0.04	0.05	0.03	0.01	0.05
700	0.00	0.00	0.03	0.03	0.04	0.03	0.03	0.01	0.05
800	0.00	0.00	0.04	0.03	0.04	0.03	0.03	0.01	0.06
900	0.00	0.00	0.04	0.03	0.04	0.03	0.03	0.01	0.05
1000	0.00	0.00	0.04	0.03	0.04	0.03	0.02	0.01	0.04

Table EC.1 **Fraction of runs that return the zero portfolio, i.e., $\|\mathbf{x}\| = 0$ by method and variability of portfolios by method. Methods that never return the zero portfolio are omitted. Based on the experiment in Sec. 6.1.**

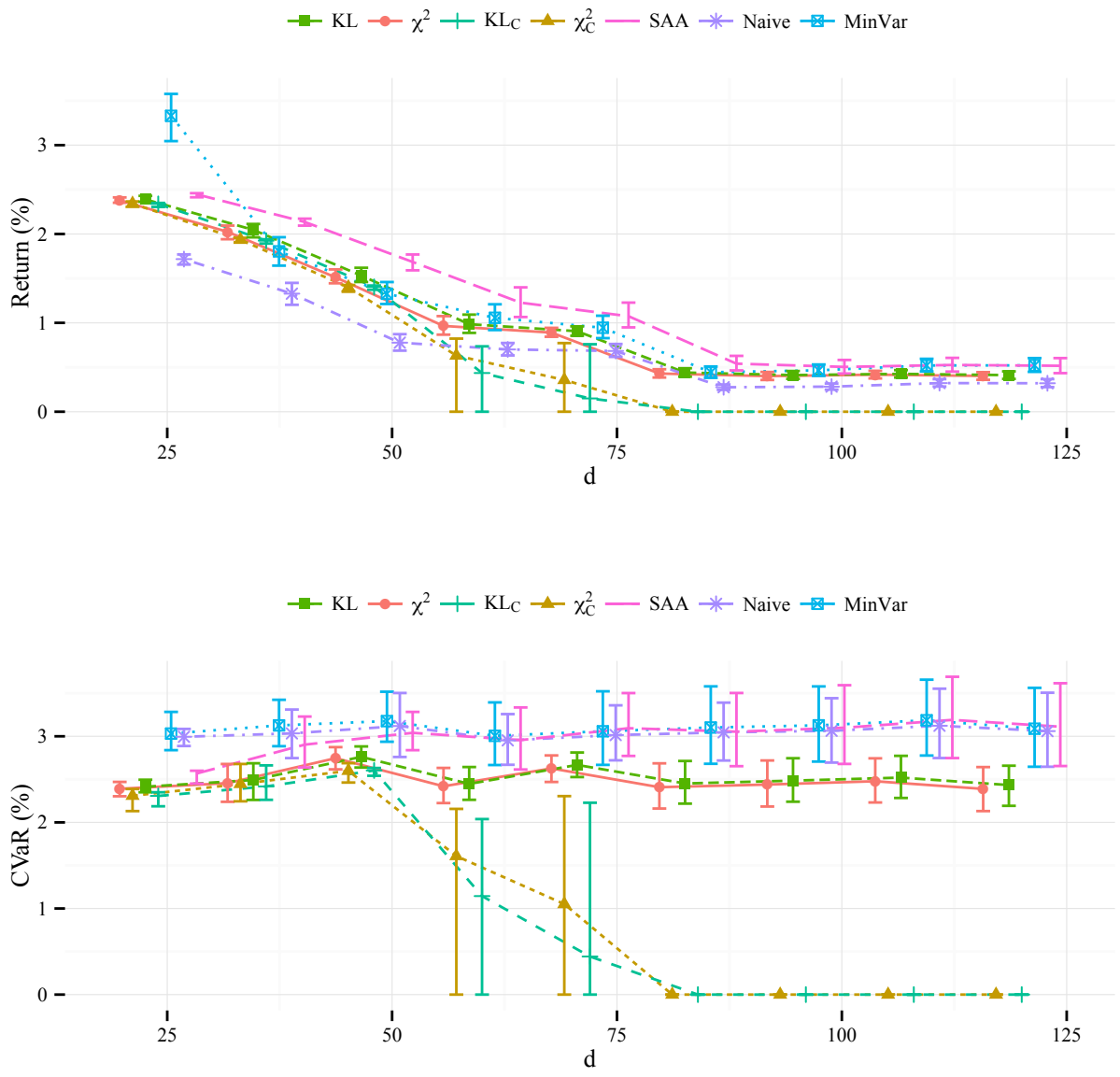

Figure EC.4 **The return and risk for all tested portfolios from Sec. 6.3.**

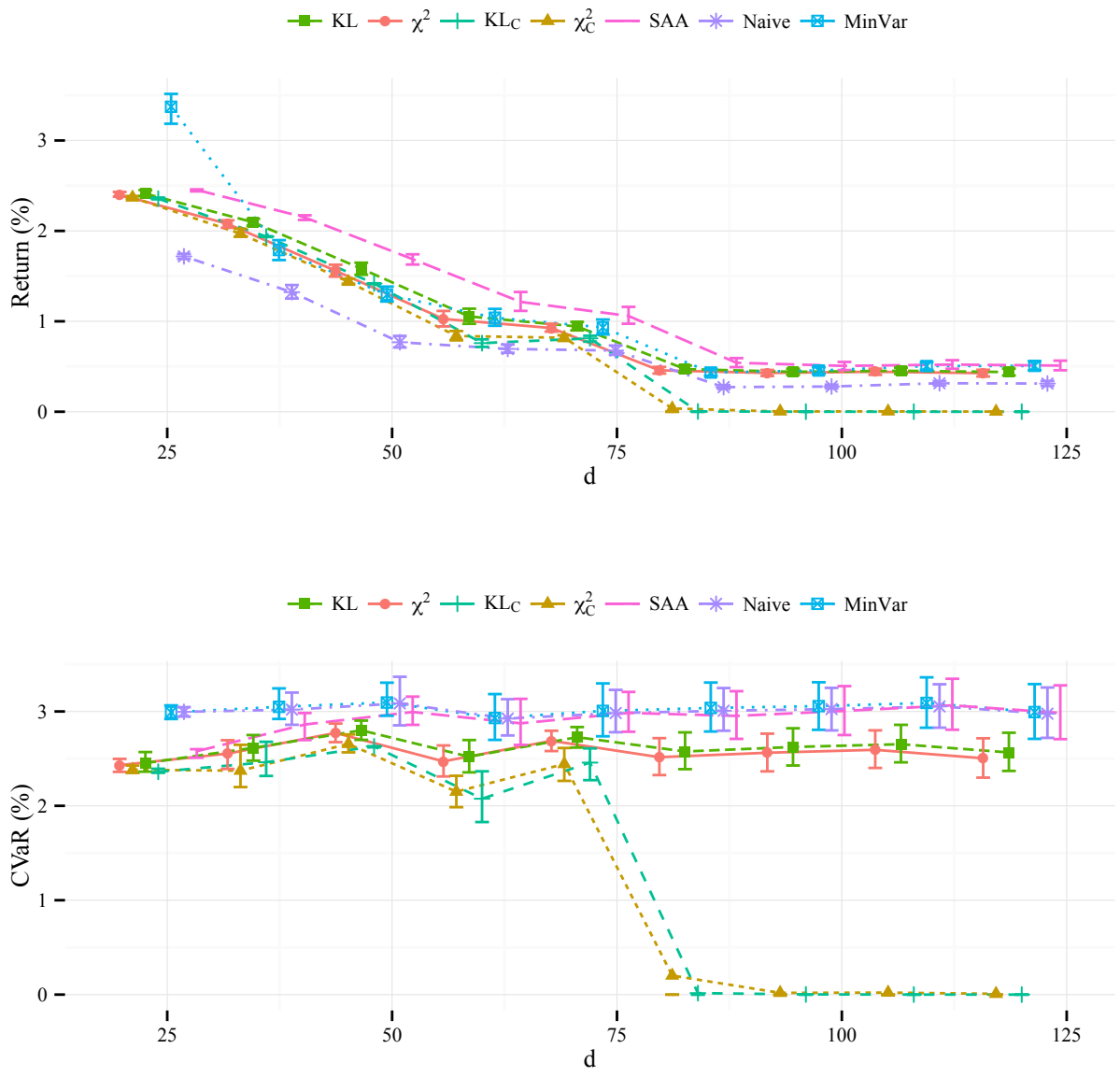

Figure EC.5 **The return and risk for all tested portfolios from Sec. 6.3 but with $N = 700$.**

Table EC.2 Summary statistics of out-of-sample portfolio performance for various randomly generated priors from Sec. 6.4. The columns 10% and 90% refer the sample quantiles of corresponding statistics.

		Return			CVaR		
Method	Strength (%)	Avg.	10%	90%	Avg.	10%	90%
ChiSq	10	0.90	0.85	0.96	2.65	2.51	2.79
	25	0.91	0.85	0.97	2.66	2.52	2.81
	50	0.92	0.86	0.99	2.69	2.54	2.83
	75	0.92	0.85	0.99	2.71	2.56	2.88
	100	0.94	0.86	1.02	2.76	2.57	2.96
	125	0.93	0.85	1.03	2.75	2.57	2.95
	150	0.94	0.85	1.03	2.78	2.59	2.98
KL	10	0.92	0.87	0.98	2.69	2.55	2.83
	25	0.93	0.87	0.99	2.70	2.56	2.85
	50	0.94	0.87	1.01	2.73	2.58	2.88
	75	0.94	0.86	1.01	2.75	2.59	2.93
	100	0.96	0.87	1.05	2.81	2.61	3.01
	125	0.95	0.86	1.06	2.80	2.61	3.01
	150	0.96	0.86	1.06	2.83	2.62	3.05

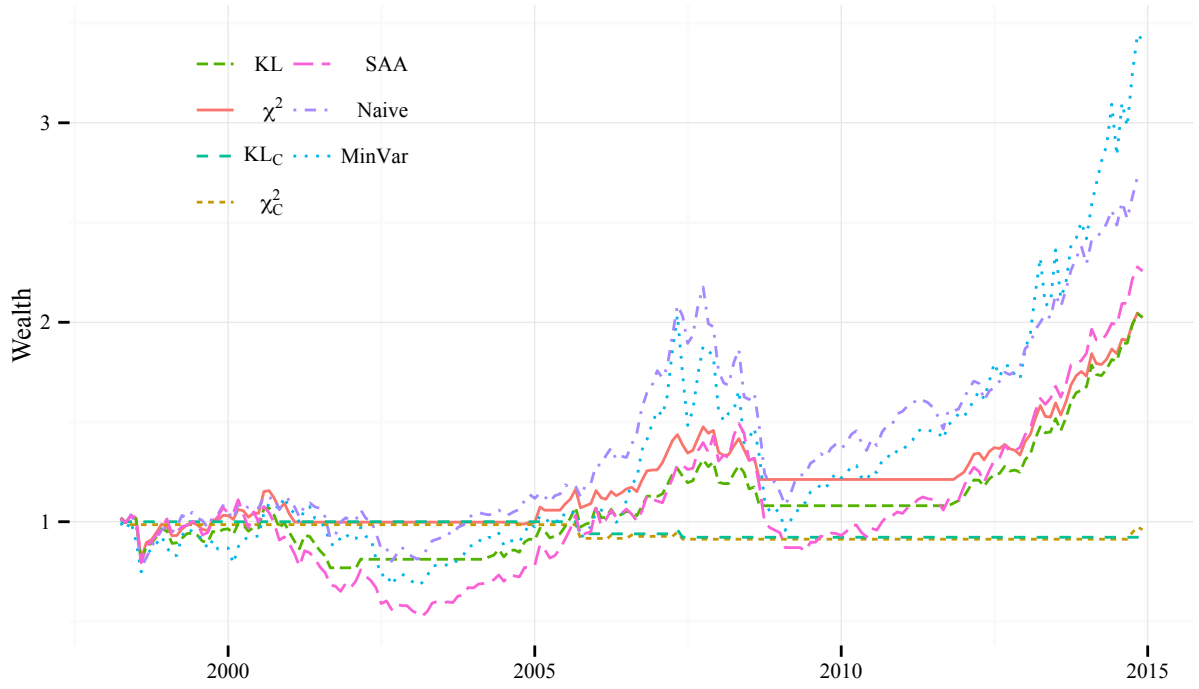

Figure EC.6 Realized performance of each portfolio from Sec. 6.5 from Mar. 1998 through Dec. 2014.

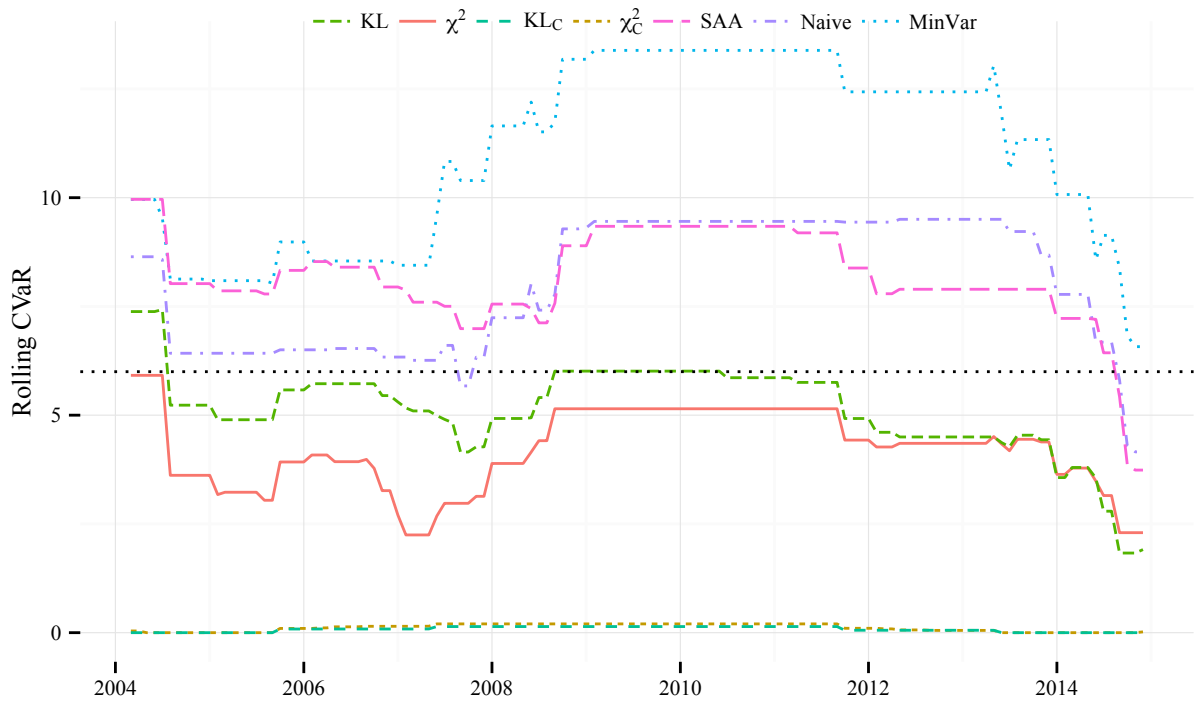

Figure EC.7 **Rolling CVaR of each portfolio from Sec. 6.5 using a trailing 72-month window from Mar. 2004 through Dec. 2014.**