Near-Optimal 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). The key idea is to measure the relative size between a candidate ambiguity set and a specific, asymptotically optimal set. This asymptotically optimal set satisfies a particular robustness guarantee and is provably the smallest such set as the amount of data grows large; it is a subset of any other convex set that also satisfies the robustness guarantee. Surprisingly, we show that existing, popular ambiguity set proposals based on statistical confidence regions are significantly larger than the asymptotically optimal set – the ratio of their sizes scales with the square root of the dimension of the ambiguity – suggesting current DRO models are unnecessarily conservative. Consequently, we also propose a new class of ambiguity sets that are tractable, satisfy our Bayesian robustness guarantee with finite data and are only a small, constant factor larger than the asymptotically optimal set; we call them near-optimal. Finally, we leverage our framework to develop simple guidelines for practitioners with respect to selecting ambiguity sets in DRO. Computational evidence in portfolio allocation using real and simulated data confirms that our framework provides practical insight into the performance of DRO models, and that our near-optimal sets outperform their traditional confidence region variants in real applications.

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

History: This paper was first submitted in July 2015.

1. Introduction

Many applications in decision-making under uncertainty can be modeled as optimization problems where constraints may depend both on the decision variables $\mathbf{x}$ and also on the distribution $\mathbb{P}^*$ of some random variables $\tilde{\xi}$. For example, in inventory problems, constraints on the probability
of stock-outs depend both on the ordering policy ($x$) and the distribution of future demand ($P^*$). Generically, we can write such constraints as $g(P^*, x) \leq 0$ for some function $g$.

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

$$\sup_{P \in \mathcal{P}_\epsilon(S)} g(P, x) \leq 0,$$

which depends on $\mathcal{P}_\epsilon(S)$. In other words, DRO protects against worst-case behavior in $\mathcal{P}_\epsilon(S)$.

Despite its seeming complexity, this approach is known to be computationally tractable for many types of functions $g$ and classes of ambiguity sets $\mathcal{P}_\epsilon(S)$ (See Ben-Tal et al. (2015), Wiesemann et al. (2013) and references therein for recent results).

Since the seminal work of Scarf (1958) in inventory, DRO models with different choices of $\mathcal{P}_\epsilon(S)$ have been proposed in the operations management literature in supply-chain design, revenue management, finance and other applications (see, e.g., Klabjan et al. (2013), Wang and Zhang (2014), Lim and Shanthikumar (2007), Bertsimas and Popescu (2002), Postek et al. (2014)). Empirical evidence in these applications suggests the approach often offers significant benefits over naive methods which neglect the ambiguity in the unknown $P^*$. This combination of tractability and effectiveness has fueled the increasing popularity of DRO approaches in operations research. Unfortunately, empirical evidence also suggests that the performance of DRO models strongly depends on the choice of $\mathcal{P}_\epsilon(S)$.

This observation raises several practical questions: Is there a “best” possible ambiguity set? 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 that perform well in practice?

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:
(A1) The unknown \( P^* \) is defined by a finite-dimensional parameter, i.e., \( P^* = P_{\theta^*} \) for some \( \theta^* \in \Theta \subseteq \mathbb{R}^d \).

(A2) For any fixed \( x \), the function \( g \) is concave in this parameter \( \theta \) and finite-valued.

(A1) is sufficiently general to include a number of special cases of DRO previously studied in the literature and employed in practice, including when \( P_{\theta^*} \) belongs to a parametric class such as normal distributions, when \( P_{\theta^*} \) is non-parametric but has known, finite discrete support, or when \( P_{\theta^*} \) is a finite, mixture model with known components (cf., Secs. 2.0.1 to 2.0.3). Importantly, it allows us to rewrite Eq. (1) (by possibly redefining \( g \) and \( P_\epsilon(S) \)) more simply as

\[
g(\theta, x) \leq 0 \quad \forall \theta \in \mathcal{P}_\epsilon(S),
\]

where now \( \mathcal{P}_\epsilon(S) \subseteq \mathbb{R}^d \).

(A2) is also quite mild. Many constraints typically found in DRO problems are concave in \( \theta \).

(See Sec. 2.0.1 for examples.) This observation may not be so surprising. If \( g \) were not concave in \( \theta \), determining feasibility of a fixed \( x \) in Eq. (2) would require maximizing a non-concave objective over \( \mathcal{P}_\epsilon(S) \) which might be numerically challenging. We denote the set of functions satisfying (A2) by \( \mathcal{G} \).

Under these two assumptions, it is possible to meaningfully define a notion of “best” and quantify the relative strength of different proposals. The key idea of our framework is to identify the smallest convex ambiguity set which satisfies a particular Bayesian robustness property (to be defined) for all \( g \in \mathcal{G} \). By smallest, we mean that it is a subset of any other convex set which also satisfies this property. We then use this set as a benchmark to assess the relative size of other ambiguity sets. As we argue later, the size of the ambiguity set is strongly related to the performance of its DRO model.

We defer a formal statement of our robustness property until Sec. 2 but note that it is a Bayesian analogue of a standard property used to measure the robustness of sets in both the traditional and distributional robust optimization literature (see, e.g., Ben-Tal et al. (2009), Bertsimas et al. (2013, 2014)).
The use of size as a proxy for performance in our framework, however, is less standard and is motivated by Eq. (2). If one ambiguity set is a subset of another, the smaller set will always yield a solution with a better objective value in DRO models. This improvement comes with no loss in robustness if both sets satisfy the same robustness property. In this sense, the smallest set that satisfies this guarantee, if it exists, is optimal.

One of our main results is that although optimal sets need not exist for finite $N$, as $N \to \infty$, an optimal set does exist under very mild assumptions. For many popular proposals, we can calculate the relative size to this asymptotically optimal set explicitly as $N \to \infty$. Intuitively, this relative size provides a good metric for choosing between competing proposals when $N$ is large; we confirm this intuition through various numerical experiments.

A perhaps surprising result of our analysis is that most popular proposals for ambiguity sets in the data-driven DRO literature are very large. This includes, for example, the $\phi$-divergence sets of Ben-Tal et al. (2013) and the elliptical set of Zhu et al. (2014). Indeed, the relative ratio of their size to the asymptotically optimal set typically scales like $\Omega(\sqrt{d})$. By contrast, we can construct ambiguity sets that satisfy the Bayesian robustness property for all $g \in \mathcal{G}$, are generally tractable and are at most a small, constant factor (independent of $d$) larger than the asymptotically optimal set. We call such sets “near-optimal.”

This distinction in size has an important practical consequence: many popular DRO models are unnecessarily conservative. By replacing their ambiguity sets with our near-optimal variants, practitioners can improve performance of these models without sacrificing robustness.

The key intuition behind the distinction in size is that our near-optimal constructions exploit the concave structure of $g$ while most popular proposals for ambiguity sets are based on statistical confidence regions, and do not exploit any such structure (Bertsimas et al. 2014). While some authors have exploited concavity to prove DRO models over certain ambiguity sets are tractable (Postek et al. 2014), to our knowledge, we are the first to exploit this concavity in constructing ambiguity sets.
That said, these results parallel ideas in traditional robust optimization. In that context, it is well-known one can construct uncertainty sets whose size is dimension independent that satisfy a slightly different robustness guarantee for functions that are concave in the uncertain parameters, and that these sets are generally much smaller than sets which contain the uncertainty with probability $1 - \epsilon$. (See, e.g., Ben-Tal et al. (2009), Chen et al. (2007), Bertsimas et al. (2013)). There is, however, no notion of an “optimal” set, and no theoretical quantification of how much smaller these sets may be.

Strangely, to the best of our knowledge, this parallel has not been utilized in constructing ambiguity sets for distributionally robust optimization. One possible explanation is that it is mathematically challenging to apply techniques from traditional robust optimization to the frequentist framework for distributionally robust optimization. One of the contributions of our work is to show that the Bayesian viewpoint overcomes this difficulty, enabling us to leverage these techniques directly.

We summarize our contributions:

1. We prove that as $N \to \infty$, there exists a smallest, convex ambiguity set that satisfies a Bayesian analogue of a common robustness property for all $g \in \mathcal{G}$. We term this set *asymptotically optimal*. We also prove that such sets need not exist for finite $N$.

2. We propose new ambiguity sets which, for finite $N$, satisfy this Bayesian robustness property for all $g \in \mathcal{G}$, are generally tractable, and enjoy other asymptotic convergence properties. Most importantly, we prove these sets are at most a small, explicitly known, constant factor larger than the asymptotically optimal set as $N \to \infty$. We term such sets *near-optimal*.

3. By contrast, we prove that existing ambiguity set proposals in the literature motivated by statistical confidence regions are much larger than the asymptotically optimal set. Specifically, there exist directions in which these sets are at least $\Omega(\sqrt{d})$ times larger than the optimal set. Importantly, this suggests that in practice, existing proposals are unnecessarily conservative, even for moderate $d$. 

4. We strengthen the above results in the case where $P_{\theta^*}$ has known, finite, discrete support. In particular, we show that the class of $\phi$-divergence ambiguity sets is at least $\Omega(\sqrt{d})$ times larger than the asymptotically optimal set in every direction. With simple modifications, however, we can scale these $\phi$-divergence sets to be near-optimal.

5. Using the above results, we propose guidelines for practitioners for selecting ambiguity sets in applications, and heuristics for tuning them when provably good performance is not required.

6. We provide computational evidence in portfolio allocation using real and simulated data, confirming that these theoretical results give practical insight into the empirical performance of DRO models, and that our near-optimal sets can significantly outperform existing proposals, even in frequentist settings.

1.1. Notations

Ordinary lowercase letters (e.g., $p_i, \theta_i$) denote scalars, boldfaced lowercase letters (e.g., $p, \theta$) denote vectors, boldfaced capital letters denote matrices (e.g., $A$), and calligraphic capital letters (e.g., $\mathcal{X}, \mathcal{S}$) denote sets. A superscript tilde (e.g., $\tilde{\theta}_i, \tilde{\theta}$) denotes a random variable or random vector. Let $e_i$ denote the $i$-th coordinate vector and $e$ denote a vector of ones. We assume throughout that $\epsilon$ is fixed, with $0 < \epsilon < 0.5$.

For any $\mathcal{P} \subseteq \mathbb{R}^d$, let $\mathcal{P} + v \equiv \{p + v : p \in \mathcal{P}\}$ denote translation of $\mathcal{P}$, and, for $\alpha > 0$, $\alpha \mathcal{P} \equiv \{\alpha p : p \in \mathcal{P}\}$ denote dilation of $\mathcal{P}$. Finally, let $\text{ri}(\mathcal{P}) = \{\theta \in \mathcal{P} : \forall z \in \mathcal{P}, \exists \lambda > 1 \text{ s.t. } \lambda \theta + (1 - \lambda)z \in \mathcal{P}\}$ denote the relative interior of $\mathcal{P}$ (cf. [Bertsekas 1999]).

2. Model and Background

We present our general framework to study a single constraint of the form Eq. (2). (Distributional robustness in the objective can be studied similarly via an epigraphic formulation.) We will assume that $\mathcal{S} = \{\hat{\xi}_1, \ldots, \hat{\xi}_N\}$ denotes our data, drawn as $N$ i.i.d. realizations of a random vector $\tilde{\xi}$ with distribution $P_{\theta^*}$, and $P_{\theta^*}$ is parameterized by an unknown, finite dimensional parameter $\theta^* \in \Theta \subseteq \mathbb{R}^d$. Strictly speaking, these draws need not be independent to apply our general framework. All that
is required is that a certain posterior quantities can be computed easily and Thm. 6 holds. These conditions are frequently met by many models where the data are not necessarily independent realizations. (We will see examples shortly.) For expositional simplicity, however, we focus on the independent case. We let \( x \in \mathcal{X} \) denote the set of constraints in our DRO problem that do not depend on \( P_{\theta^*} \).

The most common way to formalize the robustness of an ambiguity set \( P_{\epsilon}(S) \) in DRO is to require the set to satisfy the following guarantee for a fixed \( \epsilon, 0 < \epsilon < 0.5 \):

\[
\text{\textit{(Frequentist Feasibility)}} \quad \text{If } \sup_{\theta \in P_{\epsilon}(S)} g(\theta, x) \leq 0 \text{ for some } x \in \mathcal{X}, \text{ then } P(g(\theta^*, x) \leq 0) \geq 1 - \epsilon.
\]

Eq. (3) is at the heart of the theoretical justification for the DRO approach; it ensures that any solution which is robust feasible with respect to \( P_{\epsilon}(S) \) will be feasible with respect to the unknown \( P_{\theta^*} \) with probability at least \( 1 - \epsilon \). For the avoidance of doubt, we emphasize that \( x \) depends on \( P_{\epsilon}(S) \) and, hence, indirectly, on the data \( S \) in Eq. (3). Thus, it is \( x = x(S) \) which is the random quantity in the probability statement.

Ideally, Eq. (3) should hold for a large class of functions \( g \). For example, Ben-Tal et al. (2013) shows that the class of \( \phi \)-divergence sets satisfies this property for all measurable \( g \) whenever \( P_{\theta^*} \) has known, finite, discrete support, while Delage and Ye (2010) shows that a particular ambiguity set based on the first two moments of a distribution satisfies this property for all measurable \( g \) whenever \( P_{\theta^*} \) has bounded support. Similarly, Bertsimas et al. (2014) present a variety of ambiguity sets based on hypothesis tests, each of which satisfies this property for different classes of \( g \) under various assumptions on \( P_{\theta^*} \), including all measurable functions, all separable functions, and certain polynomial functions.

As mentioned, we adopt a slightly different Bayesian viewpoint. Specifically we assume that \( \theta^* \) is a realization of a random variable \( \tilde{\theta} \) with a known prior distribution supported on \( \Theta \). Then, for
any set $A \subseteq \Theta$, we can define the prior probability $P(\tilde{\theta} \in A)$ and posterior probability $P(\tilde{\theta} \in A | S)$ of this set. We propose a Bayesian analogue of the above guarantee:

**Posterior Feasibility** If $\sup_{\theta \in P_{\epsilon}(S)} g(\theta, x) \leq 0$ for some $x \in \mathcal{X}$, then $P(g(\tilde{\theta}, x) \leq 0 | S) \geq 1 - \epsilon$.

(4)

For the avoidance of doubt, we emphasize that the probability statement is conditional on the data $S$. The only random quantity is $\tilde{\theta}$. We compare Eq. (3) to Eq. (4) more fully in Sec. 3.

Our results in what follow do not strongly depend on the choice of prior. To facilitate comparisons with existing frequentists methods, we will use uninformative priors wherever possible in our own experiments and examples. By design, these priors add no additional information to the model that was not available to their frequentist counterparts.

We remark that the tractability of Eq. (2) under $[A2]$ is well-studied. Specifically, Ben-Tal et al. (2015) prove that for non-empty, convex, compact $P_{\epsilon}(S)$ satisfying a mild regularity condition\footnote{An example of a sufficient regularity condition is that $\text{ri}(P_{\epsilon}(S)) \cap \text{ri}(\text{dom}(g(\cdot, x))) \neq \emptyset$, for all $x$.}, Eq. (2) is equivalent to

$$\exists v \in \mathbb{R}^d \text{ s.t. } \delta^*(v | P_{\epsilon}(S)) - g_*(v, x) \leq 0. \quad (5)$$

Here, $g_*$ denotes the partial concave conjugate of $g$, and $\delta^*(v | P)$ denotes the support function of $P$ defined respectively as

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

Recall that support functions are convex and positively homogenous by construction. Moreover, for any closed convex, positively homogenous function $\phi(v)$, there exists a closed convex set for which $\phi$ is its support function. We will use these properties later to analyze the posterior feasibility guarantee.

For many $g$, including bi-affine and conic quadratic representable functions, $g_*(v, x)$ admits a simple, computationally tractable description. (We refer the readers to Ben-Tal et al. (2015), Bertsimas et al. (2013), Postek et al. (2014) 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 separate over \( \{(v,t) : \delta^*(v|P,\mathcal{S}) \leq t\} \) tractably. This is a substantially simpler task since it involves only linear functions of \( \theta \). In what follows, we will simply say that \( P,\mathcal{S} \) is tractable whenever this is the case.

We next introduce some special cases of the above framework that have been widely studied in the DRO literature. We will return to these illustrative examples throughout the paper.

2.0.1. Finite, discrete support. Suppose \( \tilde{\xi} \) has known finite discrete support, i.e., \( \tilde{\xi} \in \{a^1, \ldots, a^d\} \) almost surely, but the probability \( \tilde{\xi} \) attains each of these values is uncertain. Ben-Tal et al. (2013), Klabjan et al. (2013), Postek et al. (2014), Bertsimas et al. (2014) study DRO problems involving these unknown probabilities with applications in portfolio allocation and inventory management, and propose various ambiguity sets \( P,\mathcal{S} \).

We cast this setting in our framework by letting \( \theta^*_j \equiv \mathbb{P}(\tilde{\xi} = a^j) \) for \( j = 1, \ldots, d \), and \( \Theta \equiv \Delta_d = \{\theta \in \mathbb{R}^d_+ : \mathbf{e}^T \theta = 1\} \). As observed by Postek et al. (2014), most common constraints involve \( g \in \mathcal{G} \):

**Expectation and Chance Constraints:** For any function \( v(\tilde{\xi},x) \), the constraint \( \mathbb{E}[v(\tilde{\xi},x)] \leq 0 \) is equivalent to \( \sum_{j=1}^d \theta^*_j v(a^j, x) \leq 0 \) which is concave, in fact, linear, in \( \theta \). Chance constraints are a special case of expectations obtained by setting \( v \) to an indicator function, i.e., \( \mathbb{P}(v(\tilde{\xi},x) \leq 0) = \mathbb{E}[\mathbb{I}(v(\tilde{\xi},x) \leq 0)] \).

**Conditional Value at Risk and Spectral Risk Measures:** For any function \( v(\tilde{\xi},x) \), the conditional value at risk of \( v(\xi, x) \) at level \( \lambda \) is defined by \( \text{CVaR}_\lambda(v(\tilde{\xi},x)) \equiv \min_{\beta} \left\{ \beta + \frac{1}{\lambda} \mathbb{E}[v(\tilde{\xi},x) - \beta]^+ \right\} \). Conditional value at risk is a popular risk measure in financial applications. Since expectations are linear in \( \theta \), \( \text{CVaR}_\lambda \) is the minimum of a set of linear functions and hence concave in \( \theta \). Spectral risk measures are generalizations \( \text{CVaR}_\lambda \). Under suitable regularity conditions, a spectral risk measure \( \rho(v(\tilde{\xi},x)) \) can be rewritten as

\[
\int_0^1 \text{CVaR}_\lambda(v(\tilde{\xi},x))\nu(d\lambda),
\]

for some measure \( \nu \) (Noyan and Rudolf 2014). As a positive combination of concave functions of \( \theta \), spectral risk measures are also concave functions of \( \theta \).
Mean Absolute Deviation: Finally, certain statistical measures are concave in $\theta$. For example, the mean absolute deviation from the median, $\mathbb{E}[|v(\tilde{\xi}, x) - \text{Median}(v(\tilde{\xi}, x))|]$, can be rewritten as $\min_\beta \mathbb{E}[|v(\tilde{\xi}, x) - \beta|]$, which, again, is the minimum of linear functions in $\theta$, and is, hence, concave.

There do exist examples of natural constraints in this setting which are not concave in $\theta$. For example, bounds on coefficient of variation are generally non-concave, although they can sometimes be reformulated to be concave. (See [Postek et al. (2014)].)

We will adopt a Dirichlet prior for $\tilde{\theta}$. Recall, $\tilde{\theta}$ follows a Dirichlet distribution with parameter $\tau$ if it admits the probability density

$$f_{\tilde{\theta}}(\theta) = B(\tau)^{-1} \prod_{i=1}^{d} \theta_i^{\tau_i-1},$$

where $\tau_i > 0$ for all $i$ and $B(\tau)$ is a normalizing constant. The Dirichlet distribution is conjugate prior in this setting, meaning the posterior distribution of $\tilde{\theta}|S$ is also Dirichlet with updated parameter $\alpha$, $\alpha_i = \tau_i + \sum_{j=1}^{N} \mathbb{I}(\hat{\xi}_j = a_i)$. When $\tau = e$, the Dirichlet distribution reduces to a uniform distribution on the simplex, a reasonable choice for an uninformative prior. Alternatively, it is common in the Bayesian literature to let $\tau = 0$. The resulting prior distribution is also uninformative, but improper (it is non-integrable). It yields a proper posterior as long as $\alpha > 0$. In some applications, this choice of prior may be preferred ([Gelman et al. 2014], pg. 52).

2.0.2. Finite mixtures of known distributions. As a second example suppose instead that $\tilde{\xi}$ follows a mixture distribution with a finite number of known components, but that the precise mixing weights are unknown. In other words, $\tilde{\xi} \sim \sum_{i=1}^{d} \theta_i^* F_i$ where the $F_i$ are known distribution functions, and $\theta^* \in \Delta_d$, but is otherwise unknown. [Zhu and Fukushima (2009), Zhu et al. (2014)] propose ambiguity sets for $\theta^*$ and formulate DRO problems for particular financial applications. In their applications, the $F_i$ represent the distribution of asset returns under a possible future market scenario $i$.

This example naturally generalizes that of Sec. 2.0.1 and similarly maps to our framework by taking $\Theta = \Delta_d$. The reader may check that all examples of concave constraints from Sec. 2.0.1 remain concave in this setting.
We will again use a Dirichlet prior for \( \tilde{\theta} \). In this setting, the posterior distribution is not Dirichlet and must be determined numerically. For our purposes, it will be sufficient to be able to sample from the posterior, which we can do efficiently using MCMC methods [Gelman et al. 2014, Chapt. 11-12]. Both open-source and commercial implementations of these methods are widely available.

### 2.0.3. Parametric models.

Finally, many parametric models can be cast in our framework, either exactly or approximately. Although parametric models are somewhat unpopular in the DRO literature because they require strong assumptions on the underlying system, they are used frequently in practice.

As an example, suppose \( \tilde{\xi} \sim \mathcal{N}(\mu, \Sigma) \) is normally distributed, but \( \mu, \Sigma \) are unknown and estimated from data. Distributionally robust linear chance constraints can be directly cast in our framework by noting that

\[
P(\mathbf{a}(\mathbf{x})^T \tilde{\xi} > \mathbf{b}(\mathbf{x})) \leq \kappa \iff \mathbf{a}(\mathbf{x})^T \mu + z_{1-\kappa} \sqrt{\mathbf{a}(\mathbf{x})^T \Sigma \mathbf{a}(\mathbf{x})} - \mathbf{b}(\mathbf{x}) \leq 0,
\]

which is concave in \( \theta = (\mu, \Sigma) \) for any functions \( \mathbf{a}(\mathbf{x}), \mathbf{b}(\mathbf{x}) \). Similarly, distributionally robust expected quadratic constraints can be cast in our framework by noting that

\[
\mathbb{E}[(\mathbf{a}(\mathbf{x})^T \tilde{\xi})^2] \leq \kappa \iff \mathbf{a}(\mathbf{x})^T (\Sigma + \mu \mu^T) \mathbf{a}(\mathbf{x}) + \mathbf{b}(\mathbf{x})^T \mu - \kappa \leq 0,
\]

which is concave in the transformed parameter \( \theta = (\mu, \Sigma + \mu \mu^T) \) for any \( \mathbf{a}(\mathbf{x}), \mathbf{b}(\mathbf{x}) \).

We note in passing that other constraints can often be well-approximated by Eq. (6) when the posterior covariance of \( \theta \) is small using a Taylor series. Generally speaking, for many well-behaved models, as \( N \to \infty \), the posterior covariance of \( \theta \to 0 \). (See also, Thm. [6] and surrounding discussion.) Consequently, such constraints can be approximately analyzed in our framework as \( N \to \infty \).

Although the assumption of multivariate normality may seem contrived, it occurs naturally in context of many forecasting applications. For example, consider a simple autoregressive process for \( \tilde{\xi} \), i.e.

\[
\tilde{\xi}_t = \sum_{i=1}^{p} \theta_i \tilde{\xi}_{t-i} + \eta_t, \quad t = 1, \ldots,
\]
where \( \eta_t \) represents a random shock at time \( t \) and the indices represent time steps. Autoregressive processes are ubiquitous in time-series modeling, and frequently used in practice to forecast future demands or sales. Typical statistical approaches to estimating \( \theta \), such as Kalman-Filtering, assume \( \eta_t \) is normally distributed with mean 0, variance \( \sigma^2 \) and independent and identically distributed across time. Under these assumptions, \( \tilde{\xi}_t | \tilde{\xi}_1, \ldots, \tilde{\xi}_{t-1} \sim N(\sum_{i=1}^{p} \theta_i \tilde{\xi}_{t-i}, \sigma^2) \). Similar results hold for vector autoregressive processes and some linear state models that occur frequently in econometrics, control and finance. Consequently, distributionally robust parametric analysis is a reasonable choice in these applications.

We remark that in the autoregressive case, the data \( S \) are not drawn independently; future realizations do depend on the past. Nonetheless, this case can be analyzed within our framework since Thm. [6] frequently holds [Chatfield 2013].

As a final example of casting a parametric model in our framework, we consider the popular multinomial logit choice model. This model posits that when presented with an assortment \( A \subseteq \{1, \ldots, d\} \) of items, an individual \( j \) assigns utility \( v_i + \eta_{ij} \) to item \( i \), and chooses the item with highest utility if that utility is positive, and otherwise chooses no product. Here \( \eta_{ij} \) are standard Gumbel random variables, independent across \( i \) and \( j \). It is well-known that the probability item \( i \) is selected if assortment \( A \) is presented to a random customer is

\[
\frac{\theta_i}{1 + \sum_{j \in A} \theta_j}, \quad \theta_i = e^{v_i}, i = 1, \ldots, d.
\]

Variants of the multinominal logit model have been used throughout the operations management literature to model customer behavior in optimization models. These optimization problems depend on the distribution of customer utilities through the parameters \( \theta \), which must be estimated from data, making them ideal candidates for DRO. For example, in the assortment optimization problem, we seek to solve \( \max_{A \subseteq \{1, \ldots, d\}} \sum_{i \in A} r_i \theta_i / \sum_{j \in A} \theta_j \), where \( r_i \) is the revenue of item \( i \). This problem can be rewritten epigraphically as

\[
\min_{t, A \subseteq \{1, \ldots, d\}} t \quad \text{s.t.} \quad -\sum_{i \in A} r_i \theta_i / \sum_{j \in A} \theta_j \leq -t.
\]
As written, this constraint is not concave in $\theta$, however, it is equivalent to $\sum_{j\in A}(t - r_j)\theta_j \leq 0$, which is concave in $\theta$, and hence falls within our framework.

3. Comparing Feasibility Guarantees and a General Construction

Despite the continued debate on philosophical differences between Bayesian and frequentist modeling, for our purposes, the Bayesian viewpoint is mostly a mathematical convenience in our analysis.

Numerical evidence suggests that ambiguity sets which satisfy Eq. (4) also approximately satisfy Eq. (3) and vice versa. For completeness, we provide a more thorough comparison in this section.

The frequentist guarantee describes the performance of a DRO model over repeated applications with different data sets drawn from the same $\theta^*$. By contrast, the posterior feasibility guarantee describes the anticipated performance of the model for this specific dataset in this particular instance. To build intuition regarding these comparisons, it may help to introduce a third, related guarantee.

(Prior Feasibility) If $\sup_{\theta \in \mathcal{P}(\mathcal{S})} g(\theta, x) \leq 0$ for some $x \in \mathcal{X}$, then $\mathbb{P}(g(\tilde{\theta}, x) \leq 0) \geq 1 - \epsilon$. (7)

The prior feasibility guarantee is simply the frequentist feasibility guarantee (cf. Eq. (3)) averaged against the prior distribution of $\tilde{\theta}$. Similarly, it is also the posterior feasibility guarantee (cf. Eq. (4)) averaged against the sampling distribution of $\mathcal{S}$. More formally,

**PROPOSITION 1.**

1. Fix $x \in \mathcal{X}$. If $\mathcal{P}_e(\mathcal{S})$ satisfies the frequentist feasibility guarantee for $g(\cdot, x)$ for any $\theta^* \in \Theta$,
   then $\mathcal{P}_e(\mathcal{S})$ satisfies the prior feasibility guarantee for $g(\cdot, x)$ and any choice of prior.

2. Fix a prior for $\tilde{\theta}$ and $x \in \mathcal{X}$. If $\mathcal{P}_e(\mathcal{S})$ satisfies the posterior feasibility guarantee for $g(\cdot, x)$,
   then $\mathcal{P}_e(\mathcal{S})$ satisfies the prior feasibility guarantee for $g(\cdot, x)$.

Thus, both guarantees can be viewed as sufficient conditions for the prior feasibility guarantee.

An ambiguity set may satisfy the prior feasibility guarantee and significantly fail the frequentist guarantee only if the prior distribution assigns very different weights to different realizations of $\tilde{\theta}$. Similarly, an ambiguity may satisfy the prior feasibility guarantee and significantly fail the posterior
guarantee only if \( \mathbb{P}(g(\tilde{\theta}, x) \leq 0 | S) \) varies greatly for different realizations \( S \). Both possibilities seem rather unlikely in practice, so we expect in typical applications that good ambiguity sets should approximately satisfy all three guarantees. Indeed, this intuition is confirmed in our numerical experiments in Sec. 6.

The next theorem uses \((A2)\) to simplify these guarantees. The second portion of the theorem was proven in a different context in Bertsimas et al. (2013).

**Theorem 1.** Suppose \( \mathcal{P}_\epsilon(S) \) is closed and convex, and \( \text{ri}(\mathcal{P}_\epsilon(S)) \neq \emptyset \). Then,

1. \( \mathcal{P}_\epsilon(S) \) satisfies the frequentist feasibility guarantee for all \( g \in \mathcal{G} \) at level \( \epsilon \) if and only if

\[
\mathbb{P}(v(S)^T \theta^* \leq \delta^*(v(S)|\mathcal{P}_\epsilon(S))) \geq 1 - \epsilon,
\]

for all functions \( v(\cdot) \).

2. \( \mathcal{P}_\epsilon(S) \) satisfies the posterior feasibility guarantee for all \( g \in \mathcal{G} \) at level \( \epsilon \) if and only if

\[
\mathbb{P}(v^T \tilde{\theta} \leq \delta^*(v|\mathcal{P}_\epsilon(S)) \mid S) \geq 1 - \epsilon \quad \forall \ v \in \mathbb{R}^d.
\]

Thm. 1 illustrates the aforementioned mathematical convenience of the Bayesian perspective. Specifically, Eq. (8) is mathematically challenging because both \( v(S) \) and \( \mathcal{P}_\epsilon(S) \) are random (depending on \( S \)), and \( v(\cdot) \) is an arbitrary function. By contrast, Eq. (9) is simpler. Since we condition on \( S \), the only random variable is \( \tilde{\theta} \), which occurs linearly. This simpler structure gives rise to a natural technique for constructing ambiguity sets discussed in the next section.

Because of the complexity of Eq. (8), most proofs that a given convex \( \mathcal{P}_\epsilon(S) \) satisfies the frequentist feasibility guarantee do not attack Eq. (8) directly. Rather, they prove \( \mathcal{P}_\epsilon(S) \) satisfies the stronger property that for all \( \theta^* \in \Theta \),

\[
\mathbb{P}(\theta^* \in \mathcal{P}_\epsilon(S)) \geq 1 - \epsilon \iff \mathbb{P}(v^T \theta^* \leq \delta^*(v|\mathcal{P}_\epsilon(S)) \forall v \in \mathbb{R}^d) \geq 1 - \epsilon.
\]

Sets satisfying Eq. (10) are called confidence regions in statistics (see also Bertsimas et al. 2014). By construction confidence regions satisfy the frequentist feasibility guarantee for all measurable \( g \), not just \( g \in \mathcal{G} \). Comparing the righthand side of Eq. (10) with the righthand side of Eq. (9) suggests Eq. (9) is a much weaker requirement on \( \mathcal{P}_\epsilon(S) \). The “for all” occurs outside the probability. We prove in Sec. 4 sets satisfying (9) can be much smaller than those satisfying Eq. (10).
3.1. A General Construction

Thm. 1 suggests a schema for constructing ambiguity sets in DRO. This schema essentially adapts the schema of Bertsimas et al. (2013) for constructing uncertainty sets in traditional robust optimization to a Bayesian framework.

Define

\[ \text{VaR}_{\tilde{\theta}|S}(\mathbf{v}) \equiv \inf \{ t : \mathbb{P}(\mathbf{v}^T \tilde{\theta} \leq t \mid S) \geq 1 - \epsilon \}, \]

to be the posterior value at risk of \( \mathbf{v}^T \tilde{\theta} \). From Eq. (9), \( \mathcal{P}_\epsilon(S) \) satisfies the posterior feasibility guarantee for all \( g \in \mathcal{G} \) at level \( \epsilon \) if and only if

\[ \text{VaR}_{\tilde{\theta}|S}(\mathbf{v}) \leq \delta^*(\mathbf{v} \mid \mathcal{P}_\epsilon(S)) \quad \forall \mathbf{v} \in \mathbb{R}^d \text{ s.t. } \|\mathbf{v}\| = 1. \]  

(11)

Thus, to construct an ambiguity set that satisfies the posterior feasibility guarantee, it suffices to, first, compute a closed convex, positively homogenous upperbound \( \phi(\mathbf{v}) \) to \( \text{VaR}_{\tilde{\theta}|S}(\mathbf{v}) \), and, second, use standard techniques from convex analysis to identify the ambiguity set for which it is the support function. We will apply this schema later to construct \( \mathcal{P}_\epsilon(S) \) that satisfy the posterior guarantee.

Although this schema is tailored to a single constraint, we can extend to the case of multiple constraints in the same way as Bertsimas et al. (2013). In the terminology of that work, ambiguity sets satisfying Eq. (11) simultaneously satisfy the posterior guarantee for all \( 0 < \epsilon < .5 \), and, hence, we can apply their approach to optimize the division of \( \epsilon \) across multiple constraints.

Recall for any two sets, \( \mathcal{P}_1 \subseteq \mathcal{P}_2 \iff \delta^*(\mathbf{v} \mid \mathcal{P}_1) \leq \delta^*(\mathbf{v} \mid \mathcal{P}_2) \) for all \( \mathbf{v} \in \mathbb{R}^d \). Thus, tighter upper bounds in Eq. (11) yield smaller ambiguity sets. A “tightest” upperbound would yield an “optimal” set.

**Definition 1.** We say that a convex set \( \mathcal{P}_\epsilon(S) \) that satisfies the posterior feasibility guarantee for all \( g \in \mathcal{G} \) is *optimal* if \( \mathcal{P}_\epsilon(S) \) is a subset of any other ambiguity set that satisfies the posterior feasibility guarantee for all \( g \in \mathcal{G} \).

\(^2\)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).
Theorem 2. An optimal ambiguity set exists if and only if $\text{VaR}_{\tilde{\theta}|S}(v)$ is convex. When it exists, this set is unique and satisfies Eq. (11) with equality.

Although it is possible to use Thm. 2 to develop sufficient conditions on $F_\theta$ for the existence of an optimal set, in practice, these conditions are too restrictive to be useful. In the typical case, $\text{VaR}_{\tilde{\theta}|S}(v)$ is non-convex.

3.2. A General Purpose Ambiguity Set

Fortunately, there is a rich literature on upper-bounding $\text{VaR}_{\tilde{\theta}|S}(v)$ when it is non-convex. As observed in Bertsimas et al. (2013), 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 Ghaoui et al. (2003). In our Bayesian context, the bound yields

$$\text{VaR}_{\tilde{\theta}|S}^v(v) \leq v^T \mu_N + \sqrt{\frac{1-\epsilon}{\epsilon} v^T \Sigma_N v} \quad \forall v \in \mathbb{R}^d,$$

where $\mu_N, \Sigma_N$ are the posterior mean and covariance of $\tilde{\theta}|S$. When $\Sigma_N$ is invertible, we can define for any $\Gamma > 0$,

$$\mathcal{P}(\mu_N, \Sigma_N, \Gamma) \equiv \{ \theta \in \Theta : (\theta - \mu_N)^T \Sigma_N^{-1}(\theta - \mu_N) \leq \Gamma^2 \}.$$

When $\Sigma_N$ is not invertible, say $\text{rank}(\Sigma_N) = r < d$, this set is not well-defined. To remedy this, note that the random variable $\tilde{\theta}|S$ belongs to an $r$ dimensional affine subspace, almost surely. By possibly permuting the indices, we assume without loss of generality that $\hat{\theta}_1|S, \ldots, \hat{\theta}_r|S$ span this space, i.e., there exists $\beta \in \mathbb{R}^{d-r}$, $A \in \mathbb{R}^{(d-r) \times r}$ such that

$$\tilde{\theta}_{r+1,d} = \beta + A \hat{\theta}_{1,r}, \quad \text{a.s. under } \mathbb{P}(\cdot|S),$$

where $\hat{\theta}_{1,r}$ is the first $r$ components of $\hat{\theta}$ and $\tilde{\theta}_{r+1,d}$ are the remaining components. Define the approximate inverse

$$\Sigma_N^{-1} \equiv \begin{pmatrix} \Sigma_{1,r}^{-1} & 0 \\ 0^T & 0 \end{pmatrix},$$

3 For example, one can show that an optimal set exists if $F_\theta$ belongs to an exponential family and is log-concave and symmetric in $\theta$.
where $\Sigma_{1,r}$ is the restriction of $\Sigma_N$ to its first $r$ rows and columns. By construction, $\Sigma_N^{-1}$ inverts $\Sigma_N$ on the space spanned by the first $r$ components. When $\Sigma_N$ is not invertible, we interpret Eq. (13) via this approximate inverse. Then, using Eq. (12) in the previous schema yields

**Theorem 3.** $P(\mu_N, \Sigma_N, \sqrt{1/\epsilon - 1})$ satisfies the posterior feasibility guarantee at level $\epsilon$ and for all $g \in G$ almost surely.

**Remark 1.** $P(\mu_N, \Sigma_N, \Gamma)$ will be tractable for any $\Gamma > 0$ whenever we can separate over $\Theta$ tractably (Ghaoui et al. 2003). For example, when $\Theta$ is a polyhedron or SOCP representable, $\delta^*(v|P(\mu_N, \Sigma_N, \Gamma))$ is also SOCP representable. In the special case when $\Theta = \mathbb{R}^d$, $\delta^*(v|P(\mu_N, \Sigma_N, \Gamma))$ is given explicitly by Eq. (12).

**Remark 2.** When $\Sigma_N$ is singular, our definition of $\Sigma_N^{-1}$ involved the seemingly arbitrary choice of basis $\tilde{\theta}_1|S, \ldots, \tilde{\theta}_r|S$. Using other bases in the proof yields alternate, equivalent representations of the same geometric set. We prefer our representation as it simplifies some results of the next subsection.

Of course, Eq. (12) is only one of many possible upper bounds for $\text{VaR}^\varepsilon_{\tilde{\theta}|S}(v)$ that could be used to create an ambiguity set. Stronger bounds leveraging the support $\Theta$ are possible, e.g., using techniques from (Ghaoui et al. 2003). A computational benefit of $P(\mu_N, \Sigma_N, \Gamma)$ is that it only depends on the posterior mean and covariance, not the full posterior distribution. In Sec. 4 we assess the size of $P(\mu_N, \Sigma_N, \sqrt{1/\epsilon - 1})$ and show that despite its simplicity, it is nearly smallest possible.

### 3.2.1. Ambiguity sets under finite discrete support.

Recall our example from Sec. 2.0.1 where $\tilde{\theta}|S$ follows a Dirichlet distribution with parameter $\alpha$. We next specialize and strengthen Thms. 1 and 3 to this example assuming $\alpha > 0$, i.e., that the posterior distribution is proper.

**Theorem 4.**

1. For $d = 2$, $\text{VaR}^\varepsilon_{\tilde{\theta}|S}(v)$ is convex for all $S$. The optimal ambiguity set is

$$
P^\varepsilon_c(\alpha) = \left\{ \lambda \begin{pmatrix} \beta_{1-\epsilon}(\alpha_1, \alpha_2) \\ 1 - \beta_{1-\epsilon}(\alpha_2, \alpha_1) \end{pmatrix} + (1 - \lambda) \begin{pmatrix} 1 - \beta_{1-\epsilon}(\alpha_1, \alpha_2) \\ \beta_{1-\epsilon}(\alpha_2, \alpha_1) \end{pmatrix} : 0 \leq \lambda \leq 1 \right\}, \quad a.s.,$$
where $\beta_{1-\epsilon}(\alpha_1, \alpha_2)$ is the $1-\epsilon$-quantile of a Beta distribution with parameters $\alpha_1, \alpha_2$.

2. For $d \geq 3$, there exist $S$ such that $\text{VaR}^\epsilon_{\tilde{\theta}, S}(v)$ is non-convex. Consequently, there may not exist an optimal ambiguity set.

Since $\text{VaR}^\epsilon_{\tilde{\theta}}(v)$ may be non-convex, we seek convex upperbounds. Define $\alpha_0 \equiv \sum_{i=1}^d \alpha_i$. Then, it is known that

$$\mu_{N,i} \equiv \frac{\alpha_i}{\alpha_0}, \quad \Sigma_N \equiv \frac{1}{\alpha_0 + 1} \left( \text{diag}(\mu_N) - \mu_N \mu_N^T \right).$$

Notice that $\Sigma_N$ is singular, i.e., $\Sigma_N \mathbf{e} = \mathbf{0}$, corresponding to the fact that $\mathbf{e}^T \tilde{\theta} = 1$ almost surely.

Eq. (15) simplifies to

$$\Sigma_N^{-1} \equiv (1 + \alpha_0) \begin{pmatrix} \text{diag}(\mu_{N,-})^{-1} + \mu_{N,d}^{-1} \mathbf{e} \mathbf{e}^T & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix},$$

(16)

where $\mu_{N,-}$ is the restriction of $\mu_N$ to its first $d-1$ components. Substituting this formula into Thm. 3 yields

**COROLLARY 1.** Define

$$\mathcal{P}_{\chi^2}(\mu_N, \Gamma) \equiv \left\{ \theta \in \Delta_d : \sum_{j=1}^d \frac{(\theta_j - \mu_{N,i})^2}{\mu_{N,i}} \leq \Gamma^2 \right\}.$$

Then $\mathcal{P}_{\chi^2}(\mu_N, \sqrt{\frac{1/\epsilon - 1}{\alpha_0 + 1}})$ satisfies the posterior feasibility guarantee for all $S$ and for all $g \in \mathcal{G}$ a.s.

**REMARK 3.** The proposed set resembles the $\chi^2$-ambiguity set from Klabjan et al. (2013), Ben-Tal et al. (2013), Bertsimas et al. (2013), etc. An important difference is the radius of the set: $\sqrt{\frac{1/\epsilon - 1}{\alpha_0 + 1}}$. In each of the previous works, the proposed radius is $\sqrt{\chi_{d-1, 1-\epsilon}^2}$. Here $\chi_{d-1, 1-\epsilon}^2$ is the $1-\epsilon$ quantile of a chi-square random variable with $d-1$ degrees of freedom. Our ambiguity set can be much smaller than this existing proposal, especially for large $d$, and still satisfies a posterior feasibility guarantee. (See also Fig. 1a) This is a first example of the general phenomenon we discuss in detail in Sec. 4.1.

**REMARK 4.** Ben-Tal et al. (2013) shows $\{(v, t) : \delta^*(v | \mathcal{P}_{\chi^2}(\mu_N, \Gamma)) \leq t \}$ is second-order cone representable, and, hence, $\mathcal{P}_{\chi^2}(\mu_N, \Gamma)$ is tractable.
**Remark 5.** Bertsimas et al. (2014) show that \( P_{\chi^2}(\mu_N, \sqrt{\frac{\chi^2}{N}}) \) enjoys a strong asymptotic convergence property. Namely, under mild conditions, the optimal solution and objective value of a DRO problem using this set, converges to the full-information optimal solution and optimal solution had \( P_{\theta^*} \) be known exactly. Their proof can be readily adapted to show that \( P_{\chi^2}(\mu_N, \sqrt{\frac{1/\epsilon - 1}{\alpha_0 + 1}}) \) enjoys the same property.

\( P_{\chi^2}(\mu_N, \sqrt{\frac{1/\epsilon - 1}{\alpha_0 + 1}}) \) utilizes the general purpose Thm. 3. By exploiting specific properties of the Dirichlet distribution, we can construct a much smaller ambiguity set that also satisfies the posterior feasibility guarantee.

**Theorem 5.** Define

\[
P_{KL}(\mu_N, \Gamma) \equiv \left\{ \theta \in \Delta_d : \sum_{i=1}^d \mu_{N,i} \log \left( \frac{\mu_{N,i}}{\theta_i} \right) \leq \Gamma^2 \right\}.
\]

Then \( P_{KL}(\mu_N, \sqrt{\frac{\log(1/\epsilon)}{\alpha_0}}) \) satisfies the posterior guarantee for all \( S \) and for all \( g \in \mathcal{G} \) a.s.

**Remark 6.** Again, this set resembles the popular relative entropy set of Ben-Tal et al. (2013), Bertsimas et al. (2014) and others, but enjoys a much smaller radius: \( \sqrt{\frac{\log(1/\epsilon)}{\alpha_0}} \). (See also Fig. 1a). In previous works, the proposed radius is \( \sqrt{\frac{\chi^2}{N} - 1} \). This is a second example of the aforementioned general phenomenon.

**Remark 7.** Ben-Tal et al. (2013) proves that the set \( \{(v, t) : \delta^*(v | P_{KL}(\mu_N, \Gamma))\} \) is exponential cone-representable and admits a self-concordant barrier. Thus, \( P_{KL}(\mu_N, \Gamma) \) is tractable.

**Remark 8.** Again, Bertsimas et al. (2014) prove \( P_{KL}(\mu_N, \sqrt{\frac{\chi^2}{N}}) \) enjoys the asymptotic convergence property of Remark 5. Their proof implies our near-optimal variant does also.

**Remark 9.** Although theoretically tractable, the exponential cone can be numerically challenging. Bertsimas et al. (2013) observe that if \( \theta \in P_{KL}(\mu_N, \Gamma) \) with \( \Gamma = O(1/\sqrt{N}) \), then \( \theta \in \mathcal{P}_{\chi^2}(\mu_N, \Gamma + O(\sqrt{dN^{-3/2}})) \) for \( N \) sufficiently large, almost surely. This motivates heuristically replacing \( P_{KL}(\mu_N, \sqrt{\frac{\log 1/\epsilon}{\alpha_0}}) \) with \( \mathcal{P}_{\chi^2}(\mu_N, \sqrt{\frac{\log 1/\epsilon}{\alpha_0}}) \) in applications, since we can optimize over the latter as a second-order cone optimization. Strictly speaking, this latter set may not satisfy a posterior feasibility guarantee, but for large \( N \), the difference will be small.
4. Asymptotics and Relative Size

Although optimal sets need not in general exist for finite $N$, asymptotically, an essentially optimal set does exist. The key idea is based on the classical Bernstein von Mises Theorem.

**Theorem 6 (Bernstein von Mises Theorem).** Suppose $\theta^* \in ri(\Theta)$, and, for any $\kappa > 0$, $\mathbb{P}(\|\tilde{\theta} - \theta^*\| \leq \kappa) > 0$. Then under mild regularity conditions, as $N \to \infty$,

$$\sqrt{N}(\tilde{\theta} - \theta^*)|S \to_{TV} \mathcal{N}(0, \mathcal{I}(\theta^*)) \quad a.s.,$$

where the convergence is in total variation, and $\mathcal{I}(\theta^*)$ denotes the Fisher information matrix.

**Remark 10.** Explicit formulas for $\mathcal{I}(\theta)$ exist in terms of the log-likelihood of $\tilde{\xi}$ given $\tilde{\theta}$. We will not need these formulas in what follows, and, hence, omit them.

**Remark 11.** Thm. 6 only requires the prior assign positive probability to a neighborhood of the true $\theta^*$, which we can achieve by choosing a prior that admits a density over $\Theta$. More importantly, the limiting distribution does not depend on the particular prior. Thus, asymptotic results based on Thm. 6 will be robust to the choice of prior.

Thm. 6 is sometimes called the “Bayesian Central Limit Theorem.” Like the traditional Central Limit Theorem, it holds under a host of very general, but sometimes technical, assumptions. (See Chen (1985) for a rigorous treatment and references.) The result is sufficiently general, in fact, that some authors such as Gelman et al. (2014) suggest that unless the model belongs to one of a few well-known pathological cases, it is safe to simply assume asymptotic normality holds rather than try to validate the regularity conditions in practice. To avoid unnecessary technicalities in our exposition, we will simply assume Thm. 6 holds for $\tilde{\theta}|S$. For the special cases considered in Sec. 2.0.1 and Sec. 2.0.2, proofs of this claim can be found in Geyer et al. (2013), McLachlan and Peel (2004). For time-series models, see Chatfield (2013), and for the multinomial logit model, see Gelman et al. (2014).

Thm. 6 allows us to characterize the asymptotics of ambiguity sets which satisfy the posterior feasibility guarantee.
**Theorem 7.** Assume the conditions of Thm. 6 hold. Then, as $N \to \infty$,

$$
\sup_{v \in \mathbb{R}^d : \|v\|=1} \left| \text{VaR}_{\hat{\theta}}(S)(v) - v^T \mu_N - z_{1-\epsilon} \sqrt{v^T \Sigma_N v} \right| \to 0 \quad \text{a.s.,} \tag{19}
$$

where $z_{1-\epsilon}$ is the $1-\epsilon$ quantile of a standard normal distribution. Consequently, for any $0 < \kappa < 1$, with probability 1, for all $N$ sufficiently large,

1. $\mathcal{P}(\mu_N, \Sigma_N, (1 + \kappa)z_{1-\epsilon})$ satisfies the posterior feasibility guarantee at level $\epsilon$ for all $g \in \mathcal{G}$.
2. $\mathcal{P}(\mu_N, \Sigma_N, (1 - \kappa)z_{1-\epsilon})$ is a subset of any other convex ambiguity set $\mathcal{P}_\epsilon(S)$ that satisfies posterior feasibility guarantee at level $\epsilon$ for all $g \in \mathcal{G}$.

In words, Thm. 7 asserts that as $N \to \infty$, $\mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon})$ is essentially an optimal set. Any other set which satisfies the posterior feasibility guarantee for all $g \in \mathcal{G}$ eventually contains a small contraction of $\mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon})$. Any small inflation of $\mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon})$ eventually satisfies the posterior feasibility guarantee for all $g \in \mathcal{G}$.

We stress that the theorem makes no claim regarding finite $N$. Indeed, $\mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon})$ will generally not satisfy the posterior feasibility guarantee for finite $N$. Nonetheless, we can use $\mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon})$ as a benchmark to measure the relative size of other ambiguity set proposals which do satisfy the posterior guarantee for finite $N$.

**Definition 2.** We say $\mathcal{P}_\epsilon(S)$ is $\alpha$-near optimal, or simply near optimal, if there exists $\alpha = \alpha(\epsilon)$ (not depending on $N$ or $d$) such that for any $\kappa > 0$, there exists $N(\kappa)$ such that for all $N \geq N(\kappa)$,

$$
\mathcal{P}_\epsilon(S) - \mu_N \subseteq (\alpha + \kappa) \left( \mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon}) - \mu_N \right) \text{ a.s.}.
$$

**Remark 12.** From Thm. 7, an $\alpha$-near optimal set is, asymptotically, no more than $\alpha$ times larger than any other set which satisfies a posterior guarantee. This justifies our terminology “near optimal.” Def. 2 mirrors the definition of a performance guarantee for an approximation algorithm.

Perhaps surprisingly, our general purpose ambiguity set from the previous section is near optimal.

**Theorem 8.** Assume the conditions of Thm. 6 hold. The set $\mathcal{P}(\mu_N, \Sigma_N, \sqrt{1/\epsilon - 1})$ is $\sqrt{1/\epsilon - 1}$-near-optimal a.s.
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
$\epsilon$ & $\sqrt{\frac{1}{\epsilon - 1}}$ & $\frac{\sqrt{2 \log(1/\epsilon)}}{z_{1-\epsilon}}$ & $\frac{\sqrt{2 \log(1/\epsilon)}}{z_{1-\epsilon}}$ & d=3 & d=5 & d=10 & d=20 \\
\hline
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 & \\
\hline
\end{tabular}
\caption{The lefthand table shows inflation factors relating common uncertainty sets to the optimal set. The righthand graph plots these inflation factors for varying $\epsilon$ for $P_{KL}(\mu_N, \sqrt{\frac{1}{\epsilon - 1}})$ (denoted “KL”), $P_{\chi^2}(\mu_N, \sqrt{\frac{1}{\epsilon - 1}})$ (denoted “$\chi^2$”) and $P_{\phi}(S)$ (denoted “$\phi$-Div”) for $d=5, 10, 20$.}
\end{table}

**Figure 1** The lefthand table shows inflation factors relating common uncertainty sets to the optimal set. The righthand graph plots these inflation factors for varying $\epsilon$ for $P_{KL}(\mu_N, \sqrt{\frac{1}{\epsilon - 1}})$ (denoted “KL”), $P_{\chi^2}(\mu_N, \sqrt{\frac{1}{\epsilon - 1}})$ (denoted “$\chi^2$”) and $P_{\phi}(S)$ (denoted “$\phi$-Div”) for $d=5, 10, 20$.

Fig. 1a shows the value of the constant $\sqrt{\frac{1}{\epsilon - 1}}$ for some typical values of $\epsilon$.

**Remark 13.** Since $P(\mu_N, \Sigma_N, \sqrt{1/\epsilon - 1})$ and $P(\mu_N, \Sigma_N, z_{1-\epsilon})$ only differ in size, it is reasonable to consider a set of the form $P(\mu_N, \Sigma_N, \lambda(N))$ where $\lambda(1) = \sqrt{1/\epsilon - 1}$, $\lambda(N) \rightarrow z_{1-\epsilon}$ as $N \rightarrow \infty$ and $\lambda(N)$ is tuned to ensure that this set still satisfies the posterior guarantee for finite $N$. Computing such a $\lambda(N)$ requires a bounding the rate of convergence in Thm. 6. Unfortunately, this rate depends strongly on the true value of $\theta^*$, which is unknown. Thus, it seems difficult to make this construction precise. Nonetheless, it does motivate a useful heuristic in applications, where one can tune the precise size using cross-validation (see, e.g., Friedman et al. (2001)).

**4.0.2. Near optimality of $P_{KL}(\mu_N, \sqrt{\frac{\log(1/\epsilon)}{\alpha_0}})$ under finite, discrete support.** Under the model of Sec. 2.0.1 Thm. 8 directly implies that $P_{\chi^2}(\mu_N, \sqrt{\frac{1}{\epsilon - 1}})$ is also $\sqrt{\frac{1}{\epsilon - 1}}$ near-optimal. We prove an analogous result for $P_{KL}(\mu_N, \sqrt{\frac{\log(1/\epsilon)}{\alpha_0}})$.

**Theorem 9.** Assume the conditions of Thm. 6 hold. Then $P_{KL}(\mu_N, \sqrt{\frac{\log(1/\epsilon)}{\alpha_0}})$ is $\sqrt{2 \log(1/\epsilon)}$ near-optimal almost surely.
For comparison to Thm. 9 we include the constant of Thm. 9 in Fig. 1a. Neither constant uniformly dominates the other. For \( \epsilon < 0.219 \), \( P^{KL}(\mu_N, \sqrt{\frac{\log(1/\epsilon)}{\alpha_0}}) \) is asymptotically smaller than \( P^{\chi^2}(\mu_N, \sqrt{\frac{1/\epsilon-1}{\alpha_0+1}}) \). (See also Fig. 1b).

### 4.1. Sub-optimality of Credible and Confidence Regions

Recall that most existing proposals for ambiguity sets are confidence regions, i.e., they satisfy Eq. (10). The Bayesian analogue of a confidence region is a credible region, i.e., \( P_\epsilon(S) \) such that

\[
\mathbb{P}(\hat{\theta} \in P_\epsilon(S) | S) \geq 1 - \epsilon.
\]

Credible regions cannot be near-optimal.

**Theorem 10.** Suppose Thm. 6 holds, and \( P_\epsilon(S) \) is a credible region. Fix any \( \kappa > 0 \) and let \( r = \text{rank}(I(\theta^*)) \). Then, for \( N \) sufficiently large,

\[
P_\epsilon(S) - \mu_N \not\subseteq (1 - \kappa) \sqrt{\frac{1}{z_{1-\epsilon}}} \left( P(\mu_N, \Sigma_N, z_{1-\epsilon}) - \mu_N \right), \quad \text{a.s.}
\]

**Remark 14.** For most models of interest, \( \text{rank}(I(\theta^*)) \) equals the affine-dimension of \( \Theta \), which is usually \( d \). An exception is our finite, discrete support model where \( \Theta = \Delta_{d} \) has dimension \( d - 1 \).

**Remark 15.** Even for relatively small \( r \), the constant in Thm. 10 can be quite large. Fig. 1a provides some typical values. Using a standard Chernoff bound for a \( \chi^2 \) random variable, one can prove that \( \sqrt{\frac{1}{z_{1-\epsilon}}} = \Omega(\sqrt{r}) \) as \( r \to \infty \).
Remark 16. Thm. 10 only asserts that there exists a direction $v$ in which the credible region is large, preventing it from being near-optimal. For specific credible regions, we can strengthen this claim to all directions $v \in \mathbb{R}^d$ simultaneously. (See Sec. 4.1.1 and Sec. 4.1.2).

The key intuition behind Thm. 10 is illustrated in Fig. 2 in the case that $g(\theta, (v, t)) \equiv v^T \theta - t$ is a linear function. The left panel shows a credible region $\mathcal{P}_\epsilon(S)$ and a robust feasible pair $(\hat{v}, \hat{t})$, i.e., $\sup_{\theta \in \mathcal{P}_\epsilon(S)} \hat{v}^T \theta \leq \hat{t}$. The shaded trapezoid represented the sub-level set $\{\theta \in \Theta : g(\theta, (\hat{v}, \hat{t})) \leq 0\}$, which contains $\mathcal{P}_\epsilon(S)$ and some additional volume. Consequently, $\mathbb{P}(g(\tilde{\theta}, (\hat{v}, \hat{t})) \leq 0 \mid S) > \mathbb{P}(\tilde{\theta} \in \mathcal{P}_\epsilon(S) \mid S) = 1 - \epsilon$, and this inequality can be very loose depending on how much mass lies in the shaded region outside $\mathcal{P}_\epsilon(S)$. By contrast, the right panel shows a near-optimal set $\mathcal{P}_\epsilon^*(S)$, with robust feasible pair $(\hat{v}, \hat{s})$. Note, $\hat{s} < \hat{t}$. By construction, $\mathbb{P}(g(\tilde{\theta}, (\hat{v}, \hat{s})) \leq 0 \mid S)$ is very close to $1 - \epsilon$, and, consequently, $\mathcal{P}_\epsilon^*(S)$ is much smaller than $\mathcal{P}_\epsilon(S)$. While Fig. 2 illustrates the case where $g(\theta, x)$ is linear, the concave case is quite similar.

Our comments so far concern credible regions. The situation with confidence regions is more subtle. We can prove Thm. 10 holds for several, special cases of confidence regions. The key idea in these instances is that these confidence regions are also credible regions for a well-chosen prior. Such priors are called probability matching in the Bayesian literature and are known to exist for a number of popular models. They remain an active area of research (see Datta and Sweeting 2005 for a survey). To the best of our knowledge, simple general conditions for the existence of a probability matching prior ensuring that a given confidence region is also a credible region are not known. Thus, it is difficult to make a general claim like Thm. 10 for all confidence regions.

Nonetheless, we can prove a weaker corollary in the general setting:

Corollary 2. If for any $\theta^* \in \Theta \mathcal{P}_\epsilon(S)$ is a confidence region, then there exists data realizations $S$ such that $\mathcal{P}_\epsilon(S)$ is not near-optimal.

We next specialize and strengthen Thm. 10 to analyze some ambiguity sets based on confidence regions previously proposed in the literature.
4.1.1. Sub-optimality of $\phi$-divergence confidence regions. We return to our example of Sec. 2.0.1. The most popular class of ambiguity sets in this case are based on $\phi$-divergences, treated extensively in Ben-Tal et al. (2013) and utilized by many other authors. Given a function $\phi(t)$ such that $\phi(t)$ is convex for $t \geq 0$ and $\phi(1) = 0$, the $\phi$-divergence between two vectors $p, q$ is defined as $\sum_{i=1}^{d} q_i \phi \left( \frac{p_i}{q_i} \right)$. $\phi$-divergences resemble distance metrics. The most common way to create an ambiguity set from a $\phi$-divergence is to bound the distance to the posterior mean:

$$\mathcal{P}^{\phi}(\mu_N, \Gamma) \equiv \left\{ \theta \in \Delta_d : \sum_{i=1}^{d} \mu_{N,i} \phi \left( \frac{\theta_i}{\mu_{N,i}} \right) \leq \Gamma^2 \right\}.$$

This formulation generalizes many other popular ambiguity sets. For example, with $\phi(t) = (t-1)^2$, $\mathcal{P}^{\phi}(\mu_N, \Gamma) \mapsto \mathcal{P}^{\chi^2}(\mu_N, \Gamma)$ and when $\phi(t) = t \log t - t + 1$, $\mathcal{P}^{\phi}(\mu_N, \Gamma) \mapsto \mathcal{P}^{KL}(\mu_N, \Gamma)$.

Leveraging a known result in statistics, Ben-Tal et al. (2013) observe that for the prior $\tau = 0$, if $\phi''(t)$ exists in a neighborhood of 1, $\mathbb{P}(\theta^* \in \mathcal{P}^{\phi}(\mu_N, \sqrt{\frac{\phi''(1)}{2N} \chi^2_{d-1,1-\epsilon}})) \to 1 - \epsilon$. In other words, this set is a confidence region, asymptotically. They also show that $\phi$-divergence sets are tractable, with computational complexity depending on the choice of $\phi$. These two features have made this choice of ambiguity set very popular.

Unfortunately, $\phi$-divergence sets with the proposed radius are not near-optimal. This is in sharp contrast to the sets $\mathcal{P}^{\chi^2}(\mu_N, \sqrt{\frac{1/\epsilon - 1}{\alpha_0 + 1}})$ and $\mathcal{P}^{KL}(\mu_N, \sqrt{\frac{\log(1/\epsilon)}{\alpha_0}})$. Although one can prove this claim by showing that $\tau = 0$ is asymptotically a probability matching prior, we take a different approach to prove a stronger claim:

**Theorem 11.** Suppose $\phi''(t)$ exists in a neighborhood of 1 and Thm. 6 holds. Fix any $\kappa > 0$. Then with probability 1, for $N$ sufficiently large,

$$(1 - \kappa) \sqrt{\frac{X^2_{d-1,1-\epsilon}}{z_{1-\epsilon}}} \left( \mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon}) - \mu_N \right) \subseteq \mathcal{P}^{\phi} \left( \mu_N, \sqrt{\frac{\phi''(1) X^2_{d-1,1-\epsilon}}{2N}} \right) - \mu_N.$$

In other words, $\mathcal{P}^{\phi}(\mu_N, \sqrt{\frac{\phi''(1) X^2_{d-1,1-\epsilon}}{2N}})$ is not near-optimal.

**Remark 17.** Thm. 11 strengthens Thm. 10 since it shows that $\phi$-divergence sets with the proposed radius are large in every direction $v$ simultaneously.
4.1.2. Sub-optimality of the sets of ambiguity sets of Zhu and Fukushima (2009), Zhu et al. (2014). We return to our example from Sec. 2.0.2. The authors Zhu and Fukushima (2009), Zhu et al. (2014) both propose ambiguity sets for this example in slightly different applications. The first suggests in a non-data-driven scenario to use $P = \Delta_d$ to upperbound worst-case conditional value at risk. As $N \to \infty$, this set does not shrink. Hence, we expect it to perform arbitrarily badly when data is available:

**Theorem 12.** Assume Thm. 6 holds. Fix any $\kappa > 0$. For $N$ sufficiently large,

$$(\min_i \theta_i^*) (1 - \kappa) \sqrt{N} P(\mu_N, \Sigma_N, z_{1-\epsilon}) \subseteq (\Delta_d - \mu_N) \quad a.s.$$  

By contrast, Zhu et al. (2014) suggest the set $P_{\chi^2}(\mu_N, \sqrt{\chi^2_d}, \frac{1 - \epsilon}{N})$ and argue that it is asymptotically a credible region under some regularity conditions. A sub-optimality bound for $P_{\chi^2}(\mu_N, \sqrt{\chi^2_d}, \frac{1 - \epsilon}{N})$ follows directly from Thm. 11. Indeed, the proof of Thm. 11 only relies on the geometry of the sets, not the fact that $\tilde{\xi}$ has finite discrete support. Consequently, it applies readily to case of a finite number of mixture distributions as well.

4.2. Necessity of (A2)

Thus far, we have constructed sets that satisfy the posterior feasibility guarantee for all $g \in G$ for finite $N$ which are roughly $\Omega(\sqrt{d})$ times smaller than the ambiguity sets currently used in practice. We next prove that if we expand $G$ to include convex functions, sets which satisfy the posterior guarantee are similarly sized to credible regions asymptotically. This highlights the strong role of concavity in our results. We discuss some modeling implications in Sec. 5. We focus on the case where $I(\theta^*)$ is invertible. The singular case is similar.

**Theorem 13.** Suppose Thm. 6 holds with $I(\theta^*)$ invertible, and suppose $P_\epsilon(S)$ satisfies a posterior feasibility guarantee for all $g(\theta, x)$ which are convex quadratic functions of $\theta$ for a fixed $x$. Fix any $\kappa > 0$. Then, with probability 1, for $N$ sufficiently large,

$$P_\epsilon(S) \subseteq (1 - \kappa) P(\mu_N, I(\theta^*), \sqrt{\frac{\chi^2_d}{N}}) .$$

**Remark 18.** By Thm. 6 as $N \to \infty$, the set on the right is nearly a $1 - \epsilon$ credible region. For large $d$, it has radius $\Omega(\sqrt{d})$. 
5. Guidelines for Practitioners

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

First, in applications which require a provable feasibility guarantee at level \( \epsilon \), constants like those in Thms. 8 to 12 can provide guidelines as to which sets to use when \( N \) is large. For example, in the finite, discrete model of Sec. 2.0.1, Fig. 1a suggests that \( \mathcal{P}_{KL}(\mu_N, \sqrt{\frac{\log(1/\epsilon)}{\alpha_0}}) \) should be preferred to \( \mathcal{P}_{\phi}(\mu_N, \sqrt{\frac{\phi''(1)\chi^2_{d-1,1-\epsilon}}{2N}}) \) for any \( \phi \)-divergence, especially when \( d \) is large. If \( \epsilon \) is not too small, \( \mathcal{P}_{\chi^2}(\mu_N, \sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{N}}) \) is a good, computationally simpler alternative. Developing similar constants for other models is straightforward via Thm. 7. Moreover, developing new, ambiguity sets for custom applications is possible by directly applying the schema of Sec. 3.1.

Moreover, whenever \( d \) is large, these constants also suggest that DRO models based on existing ambiguity sets are likely to be unnecessarily conservative. We can use Thm. 6 to quantify this conservatism for large \( N \). Namely, for each of our sets \( \mathcal{P}_\epsilon(S) \), we can compare the desired robustness level \( \epsilon \) to the achieved robustness level \( \epsilon' \) under the asymptotic distribution of Thm. 6. We illustrate the key idea in the case of our sets for finite, discrete support.

For example, the set \( \mathcal{P}_{\chi^2}(\mu_N, \sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{N}}) \) actually satisfies the posterior guarantee at level \( \epsilon' = 1 - \Phi(\sqrt{1/\epsilon' - 1}) \) as \( N \to \infty \). Similar computations can be made for each of our sets. We plot these asymptotically achieved robustness levels in the right hand panel of Fig. 3. The differences between the desired robustness and actual robustness can be striking, especially for large \( d \). This plot also highlights the fact that \( \mathcal{P}_{\chi^2}(\mu_N, \sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{N}}) \) is unsuitable for small \( \epsilon \).

For small \( N \), these constants are less informative. The left panel of Fig. 3 shows the ratios \( \delta^*(v|\mathcal{P})/\text{VaR}_{\theta^*}(v) \) for randomly chosen directions \( v \), on a single sample path varying \( N \), and sets

\[
\mathcal{P}_{\chi^2}(\mu_N, \sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{N}}), \mathcal{P}_{KL}(\mu_N, \sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{2N}}).
\]

Notice, these last two sets are confidence regions as \( N \to \infty \). We have set \( \theta^* = \frac{1}{10} e \), \( \epsilon = .1 \), and \( \tau = 0 \). We draw attention to the following features:

- The finite ratio can be above or below the asymptotic ratio.
Figure 3  The lefthand panel shows the ratio $\delta^*(v|\mathcal{P})/\text{VaR}_{\theta|S}(v)$ for $\mathcal{P}$ equal to $\mathcal{P}^{KL}(\mu_N, \sqrt{\log(1/\alpha_0)})$ (denoted “$KL$”), $\mathcal{P}^{\chi^2}(\mu_N, \sqrt{\frac{1/\epsilon - 1}{\frac{N}{\alpha_0} + 1}})$ (denoted “$\chi^2$”), $\mathcal{P}^{KL}(\mu_N, \sqrt{\frac{N}{2} - \frac{1/\epsilon - 1}{2N}})$ (denoted “$KL_C$”), $\mathcal{P}^{\chi^2}(\mu_N, \sqrt{\frac{N}{2} - \frac{1/\epsilon - 1}{2N}})$ (denoted “$\chi^2_C$”). We take $\theta = \frac{1}{15} e$, $d = 15$, and $\epsilon = .1$. The righthand panel compares the desired robustness level $\epsilon$ with the asymptotically achieved robustness level $\epsilon'$ for $\mathcal{P}^{KL}(\mu_N, \sqrt{\frac{\log(1/\alpha_0)}{\alpha_0}})$, $\mathcal{P}^{\chi^2}(\mu_N, \sqrt{\frac{1/\epsilon - 1}{\frac{N}{\alpha_0} + 1}})$, and several $\phi$-divergence sets for $d = 5, 7, 10$. We include the dotted line $\epsilon = \epsilon'$ for reference.

- For small $N$, the ordering between sets may change. For example, although $\mathcal{P}^{KL}(\mu_N, \sqrt{\frac{\log(1/\alpha_0)}{\alpha_0}})$ is asymptotically smaller than $\mathcal{P}^{\chi^2}(\mu_N, \sqrt{\frac{N}{2} - \frac{1/\epsilon - 1}{2N}})$, for $N < 20$, it yields a larger set in this example.

- The empirical rate of convergence differs by set and by its size. For example, $\mathcal{P}^{\chi^2}(\mu_N, \sqrt{\frac{N}{2} - \frac{1/\epsilon - 1}{2N}})$, converges almost immediately; $\mathcal{P}^{KL}(\mu_N, \sqrt{\frac{\log(1/\alpha_0)}{\alpha_0}})$ converges more slowly, and $\mathcal{P}^{KL}(\mu_N, \sqrt{\frac{N}{2} - \frac{1/\epsilon - 1}{2N}})$ converges even more slowly.

In applications where a provable guarantee at level $\epsilon$ is not required, the previous results strongly suggest tuning the radius of the ambiguity set for the application at hand, for example, with cross-validation. Indeed, in the case of finite, discrete support, the dominant difference between near-optimal sets and existing proposals is size, not shape. This observations suggests in other
applications, one may be able to shrink the size of the ambiguity set while retaining many robustness properties. A key insight is that a good radius of order $O\left(\frac{\log(1/\epsilon)}{\sqrt{N}}\right)$ likely exists, independently of $d$.

Finally, Sec. 4.2 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 $g(\theta, x)$ is concave in $\theta$, 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.

6. Computational Experiments

We present a series of numerical experiments based on synthetic and real data. We are interested in the following questions: Do near-optimal sets, developed in a Bayesian framework, still exhibit good frequentist properties? Does our theoretical analysis of size yield useful insight into the performance of optimization models? How sensitive are our results to misspecification of the Bayesian model? In real applications, do our near-optimal sets offer a benefit over traditional DRO sets?

We explore these questions with respect to a particular portfolio allocation problem. Portfolio allocation has been widely studied in the data-driven DRO literature, especially, because it is well-known that simple methods that neglect ambiguity in $\theta^*$ can perform poorly [Lim et al. (2011), DeMiguel and Nogales (2009)]. For concreteness, we focus on the optimization problem

$$
\max_{x \in \mathbb{R}_+^n} \min_{\mathbb{P}_\theta \in \mathcal{P}_\epsilon(S)} \mathbb{E}[x^T \tilde{\xi}]
$$

s.t. $\text{CVaR}_{\mathbb{P}_\theta}(x^T \tilde{\xi}) \leq \Gamma, \quad \forall \mathbb{P}_\theta \in \mathcal{P}_\epsilon(S),$

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

To facilitate comparisons to existing methods, we will assume the setup of Sec. 2.0.1, i.e., that $\tilde{\xi}$ has known, finite, discrete support. In particular, we will consider the sets $\mathcal{P}^{KL}(\mu_N, \sqrt{\log(1/\epsilon)/\alpha_0})$ (denoted “KL”), $\mathcal{P}^{KL}(\mu_N, \sqrt{\chi^2_{d,1-\epsilon}/(2N)})$ (denoted “KL$^C$”), $\mathcal{P}^{\chi^2}(\mu_N, \sqrt{1-\epsilon/(\alpha_0+1)}\epsilon)$ (denoted “$\chi^2$”), and $\mathcal{P}^{\chi^2}(\mu_N, \sqrt{\chi^2_{d,1-\epsilon}/N})$ (denoted “$\chi^2_C$”). In each case the subscript $C$ indicates the confidence
Table 1  Summary statistics for the individual industry portfolios.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std</td>
</tr>
<tr>
<td>Business Equipment</td>
<td>1.74</td>
<td>4.74</td>
</tr>
<tr>
<td>Chemicals</td>
<td>1.38</td>
<td>4.41</td>
</tr>
<tr>
<td>Consumer Durables</td>
<td>2.29</td>
<td>8.80</td>
</tr>
<tr>
<td>Energy</td>
<td>0.78</td>
<td>5.62</td>
</tr>
<tr>
<td>Healthcare</td>
<td>1.68</td>
<td>3.80</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>1.73</td>
<td>6.07</td>
</tr>
<tr>
<td>Finance</td>
<td>1.38</td>
<td>6.33</td>
</tr>
<tr>
<td>Consumer NonDurables</td>
<td>1.48</td>
<td>3.41</td>
</tr>
<tr>
<td>Other</td>
<td>1.53</td>
<td>5.48</td>
</tr>
<tr>
<td>Wholesale/Retail</td>
<td>1.70</td>
<td>3.99</td>
</tr>
<tr>
<td>Telecom</td>
<td>1.70</td>
<td>4.22</td>
</tr>
<tr>
<td>Utilities</td>
<td>1.14</td>
<td>3.57</td>
</tr>
</tbody>
</table>

region variant of the set, instead of the near-optimal one. Unless otherwise specified, we adopt the uninformative prior $\tau = e$. For comparison, we also consider the sample average approximation of Eq. (20), which replaces $P_\epsilon$ with the singleton set containing the empirical distribution.

Our data is based upon the historical returns of 12 industry portfolios available at 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 based on these and other experiments:

- Under frequentist assumptions, i.e. repeated sampling, our near-optimal sets approximately satisfy a frequentist feasibility guarantee. 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 near-optimal sets significantly outperform their confidence region variants in this application for both synthetic and real data.
These observations remain generally true under prior misspecification. For large $N$, moderate errors in the prior do not significantly affect the performance. For small $N$, however, very large errors in the prior can yield poor performance.

6.1. Dependence on $N$

We begin by studying the performance of our sets as $N \rightarrow \infty$ 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. (20) with $\Gamma = 3\%$. We then use the true distribution to compute the actual expected return and CVaR of this portfolio. Notice this repeated sampling set-up accords precisely with the frequentist viewpoint.

Fig. 4 displays the expected returns and CVaRs. We draw attention to several features:

- The SAA solution frequently incurs more than allocated 3% risk. Indeed, even for very large $N$, the error bars are almost symmetric around this value. For smaller $N$, the returns are also highly unstable; the error bars are very large. These are well-documented drawbacks of SAA (see, e.g., Bertsimas et al. (2014)).
On the other hand, 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 $N$ near 250 occur because, for some data realizations, the only portfolio that the model can safely guarantee will be feasible is $x = 0$, i.e., to not invest.

Finally, our near-optimal sets perform much stronger. They safely maintain a risk below 3%, but the error bars are fairly close to the budget; they are not overly conservative. As a consequence, their expected return is also much higher, reasonably close to the SAA return. Unlike, SAA, however, the returns are very stable; the error bars are very small.

These findings to support the idea that our Bayesian sets have good frequentist properties.

6.2. Dependence on $d$

An important implication of our theoretical results is that DRO models with confidence region based uncertainty sets may not scale well with $d$. Consequently, we next study the relative performance of these methods as a function of $d$ for synthetic data. Specifically, we take the true distribution to be supported on the most recent $d$ monthly returns of our 12 indices, and then repeatedly sample $N = 300$ data points from this distribution. We then solve our previous optimization problems with these points, and collect results for varying $d$. See Fig. 5.

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, again, converge to investing in the extremely conservative $x = 0$ portfolio. Similarly, although the SAA portfolio maintains a reasonably good return, as $d$ grows, it starts violating the risk bound. By contrast, our near-optimal ambiguity sets maintain a return fairly close to the SAA return, but safely maintain a risk below 3%.

This experiment confirms our theoretically predicted behavior and is a strong argument for using near-optimal sets when $d$ is moderately large.
6.3. Incorrect Priors

A substantive concern with Bayesian modeling is sensitivity to the choice of prior. Thm. 3 ensures that as $N \to \infty$ this choice becomes irrelevant, but it is less clear what the effect is for finite $N$. Consequently, we next study the performance of our methods under incorrectly specified priors. Specifically, we assume the true distribution is as described in Sec. 6.1. We then solve Eq. (20) using $P_{KL}(\mu_N, \sqrt{\log(1/\epsilon)/\alpha_0})$ under three, distinct priors:

- $\tau^1_i = 0.0076i$, $i = 1, \ldots, d$
- $\tau^2_i = 0.161e^{i/72}$, $i = 1, \ldots, d$,
- $\tau^3_i = 0.0005e^{5i/72}$, $i = 1, \ldots, d$.

Intuitively, the first prior increases linearly in $i$, while the second two increase exponentially in $i$, and all three are rescaled to sum to approximately 20. As can be seen in Fig. 6, the differences between these portfolios is minor. The behavior for $P_{\chi^2}(\mu_N, \sqrt{\frac{1-\epsilon}{(\alpha_0+1)\epsilon}})$ is similar.

As a more stringent test, we next fix $N = 300$. We consider various priors of the form $(\tau_0, 1, \ldots, 1)$ as $\tau_0$ increases. (The inclusion of the 1 terms ensure that the posterior is proper.) Notice this prior
Figure 6  The return and risk of portfolios built from the $P^{KL}(\mu_N, \sqrt{\log(1/\epsilon)}/\alpha_0)$ with various prior specifications for varying $N$ from Sec. 6.3.

Figure 7  The return and risk of portfolios built from the $P^{KL}(\mu_N, \sqrt{\log(1/\epsilon)}/\alpha_0)$ with an increasingly strong, incorrect belief Sec. 6.3 is not uninformative. It is a highly informative, but incorrect; the true distribution has relatively small probability of occurring under this prior. Larger $\tau_0$ correspond to a stronger belief in the prior. Fig. 7 shows the results for $P^{KL}(\mu_N, \sqrt{\log(1/\epsilon)}/\alpha_0)$. 
Table 2  Realized performance of each portfolio from Mar. 1998 - Dec. 2014 from Sec. 6.4. Target CVaR is 4%.

<table>
<thead>
<tr>
<th></th>
<th>Avg. Return</th>
<th>Std. Dev</th>
<th>VaR</th>
<th>CVaR</th>
<th>Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>.313</td>
<td>2.510</td>
<td>2.693</td>
<td>4.177</td>
<td>.19</td>
</tr>
<tr>
<td>$\chi^2_C$</td>
<td>(.016)</td>
<td>.750</td>
<td>-</td>
<td>.838</td>
<td>.06</td>
</tr>
<tr>
<td>KL</td>
<td>.345</td>
<td>2.851</td>
<td>377.7</td>
<td>4.957</td>
<td>.21</td>
</tr>
<tr>
<td>$KL_C$</td>
<td>(.039)</td>
<td>.529</td>
<td>-</td>
<td>.475</td>
<td>.03</td>
</tr>
<tr>
<td>SAA</td>
<td>.426</td>
<td>3.664</td>
<td>527.1</td>
<td>6.593</td>
<td>.30</td>
</tr>
</tbody>
</table>

Clearly, as $\tau_0$ increases, and our confidence in the incorrect prior increases, the performance suffers. We note, however, that it is not until $\tau_0 = 175$ that the portfolio begins to incur more than 3% risk. To build intuition, a value of $\tau_0 = 175$ should be interpreted as having $(175 + 71)/300 \approx 82\%$ as much confidence in our prior distribution as we do in our data. Similarly, even at $\tau_0 = 300$ the average return exceeds the average return of the corresponding confidence-region based set (cf. Fig. 4). A value of $\tau_0 = 300$ should be interpreted as having $(300 + 71)/300 \approx 125\%$ as much confidence in our prior distribution as we do in our data. In both cases, these are extremely strong, incorrect beliefs. Consequently, we feel justified in asserting that for uninformative priors, or well-constructed informative priors, our Bayesian sets outperform existing ambiguity sets based on confidence regions.

6.4. Historical Case Study

Finally, we consider historically back-testing portfolios built from our sets on real data from Mar. 1998 - Dec. 2014. At each point in time, we assume that the true distribution is supported on the most recent 36 monthly returns, and that these past 36 months represent i.i.d. draws from this distribution. We then form portfolios using each of our sets, and record the realized return from these portfolios using this upcoming month’s return. We set the budget $\Gamma = 4\%$, since lower values cause the confidence region sets to uniformly invest in the portfolio $x = 0$.

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 the performance of our methods under model misspecification. Table 2 shows summary statistics for the performance of each method.

Notice that as with our synthetic data, although SAA yields the highest return, it exceeds the threshold on CVaR significantly. Our near-optimal sets also exceed this threshold (due to the model
inaccuracy) but much less so. The confidence region-based sets perform much more poorly. They frequently do not invest at all, yielding an overall negative average return over the period. These observations accord well with our synthetic data.

Fig. 8 plots the cumulative monthly return for our near-optimal sets and the SAA set. (We omit the confidence region sets for clarity.) This figure summarizes the performance of our near-optimal sets well compared to SAA. In down-markets, such as between 2002 and 2004 or around 2009, our near-optimal sets recognize the potential risk and choose not to invest at all (the large flat regions around these points.) Consequently, they outperform SAA. In very strong up-markets, they are not as aggressive, and, hence, underperform relative to SAA (the peaks in the graph).

Finally, although our optimization problem Eq. (20) does not explicitly control for multi-period transaction costs, we note the average monthly turnover for the near-optimal sets is much smaller than the SAA portfolio in Table 2. Practically, this would correspond to smaller 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. Generally speaking, we find that existing proposals
based on confidence regions are unnecessarily large, causing them to be over conservative in applications. We propose a new class of ambiguity sets that enjoy the usual tractability, feasibility and asymptotic convergence guarantees, but are $\Omega(\sqrt{d})$ smaller than existing proposals. These results have important implications for using DRO models in practice, including the fact that existing models are likely to be unnecessarily conservative when $d$ is large.

**References**


Postek, Krzysztof, Dick Den Hertog, Bertrand Melenberg. 2014. Tractable counterparts of distributionally robust constraints on risk measures.


Wang, Xuan, Jiawei Zhang. 2014. Process flexibility: A distribution-free bound on the performance of k-chain. *Available at SSRN 2311268*.


Appendix. Omitted Proofs

Proof Thm. 1. First suppose that $P_r(S)$ satisfies the frequentist feasibility guarantee for all $g \in \mathcal{G}$. Define $g(\theta, (v, t)) = v^T \theta - t$, which is concave, in fact, linear, in $\theta$. Notice that for any $v(S)$, the pair $(v, \delta^*(v(S)|P_r(S)))$ satisfies

$$\sup_{\theta \in P_r(S)} g\left(\theta, (v(S), \delta^*(v(S)|P_r(S)))\right) \leq 0,$$

and, hence, by definition of the feasibility guarantee,

$$1 - \epsilon \leq \mathbb{P}\left(g\left(\theta^*, (v(S), \delta^*(v(S)|P_r(S)))\right) \leq 0\right) = \mathbb{P}\left(v(S)^T \theta^* \leq \delta^*(v(S)|P_r(S))\right)$$

For the converse, consider an arbitrary $g \in \mathcal{G}$ and, for any $S$, let $x(S)$ be a feasible solution satisfying $\sup_{\theta \in P_r(S)} g(\theta, x(S)) \leq 0$. By Theorem 2 of Ben-Tal et al. (2015), (cf. Eq. (5))

$$\exists v(S) \text{ s.t. } \delta^*(v(S)|P_r(S)) \leq g_*(v(S), x(S)) \, \forall S.$$ (The necessary regularity condition is met since $P_r(S)$ is non-empty and $g \in \mathcal{G}$.) We stress that $v(S)$ depends on $S$ through $P_r(S)$. By assumption, $\mathbb{P}\left(v(S)^T \theta^* \leq \delta^*(v(S)|P_r(S))\right) \geq 1 - \epsilon$, which implies that

$$1 - \epsilon \leq \mathbb{P}(v(S)^T \theta^* \leq g_*(v(S), x))$$

$$\leq \mathbb{P}(\inf_{v \in v(S)} \{v^T \theta^* - g_*(v, x)\} \leq 0)$$

$$= \mathbb{P}(g(\theta^*, x) \leq 0),$$

where the last line follows because $g_*(\cdot, x) = g_*(\cdot, x)$. This proves the first part of the theorem.

The second part of the theorem is a direct application of Thm. 1 from Bertsimas et al. (2013) with appropriate adaptations to match our Bayesian framework. We omit the details. □

Proof of Thm. 2. (“If” direction) By continuity of probability, $\text{VaR}_{\theta|S}^\epsilon(v)$ is a closed function. It is positively homogenous by construction, and convex by assumption. Consequently, there exists a unique, closed convex $P^*$ such that $\delta^*(v|P^*) = \text{VaR}_{\theta|S}^\epsilon(v)$ (Nedic et al. 2003). Notice $P^*$ satisfies Eq. (11) with equality. Consequently, $P^*$ also satisfies the posterior feasibility guarantee and is a subset of any other convex set which also satisfies this guarantee.

(“Only if” direction) When $\text{VaR}_{\theta|S}^\epsilon(v)$ is non-convex, we can always identify two ambiguity sets $P_1, P_2$, both of which satisfy the posterior feasibility guarantee, but neither of which contains the other. We give an explicit construction. Let $v_1, v_2$ and $0 < \lambda < 1$ be such that

$$\text{VaR}_{\theta|S}^\epsilon(\lambda v_1 + (1 - \lambda)v_2) > \lambda \text{VaR}_{\theta|S}^\epsilon(v_1) + (1 - \lambda) \text{VaR}_{\theta|S}^\epsilon(v_2).$$
Lemma 1. Suppose for all \( 0 < \epsilon < 0.5 \), \( \text{VaR}_{\hat{\Theta}|S}^\epsilon(v) \leq \delta^*(v \mid \mathcal{P}_r(S)) \), \( ri(\mathcal{P}_r(S) \cup \Theta) \neq \emptyset \) and \( \delta^*(v \mid \mathcal{P}_r(S)) \) is continuous in \( \epsilon \). Then, \( \text{VaR}_{\hat{\Theta}|S}^\epsilon(v) \leq \delta^*(v \mid \mathcal{P}_r(S) \cap \Theta) \).

Proof of Lemma. From [Ben-Tal et al., 2015], \( \delta^*(v \mid \mathcal{P}_r(S) \cap \Theta) = \min_y \delta^*(v - y \mid \mathcal{P}_r(S)) + \delta^*(y \mid \Theta) \). Next, by the union bound and definition of Value at Risk, for any \( \epsilon_1, \epsilon_2 > 0 \) such that \( \epsilon_1 + \epsilon_2 = \epsilon \)

\[
\text{VaR}_{\hat{\Theta}|S}^\epsilon((v - y) + y) = \text{VaR}_{\hat{\Theta}|S}^{\epsilon_1}(v - y) + \text{VaR}_{\hat{\Theta}|S}^{\epsilon_2}(y) \\ \\
\leq \delta^*(v - y \mid \mathcal{P}_1(S)) + \delta^*(y \mid \Theta),
\]

where the last line follows by assumption and the fact that \( \text{VaR}_{\hat{\Theta}|S}^\epsilon(v) \leq \delta^*(v \mid \Theta) \) for all \( 0 < \epsilon \).

Taking the limit as \( \epsilon_1 \to \epsilon \) and minimizing over \( y \) proves the lemma.

Proof of Thm. 3 Let \( r = \text{rank}(\Sigma_N) \leq d \). Using Eq. (14) and the definition of value-at-risk, we have

\[
\text{VaR}_{\hat{\Theta}|S}^\epsilon(v) = v_{r+1,d}^T \beta + \text{VaR}_{\hat{\Theta}_{1,r}|S}^\epsilon(v_{1,r} + A^T v_{r+1,d}) \\
\leq v_{r+1,d}^T \beta + \left(v_{1,r} + A^T v_{r+1,d}\right) \Sigma_{1,r}^{-1} v_{1,r} + A^T v_{r+1,d} \\
= v_{r+1,d}^T \beta + \max_{\theta_{1,r}:\Sigma_{1,r}^{-1}(\theta_{1,r},\mu_{1,r}) \leq \frac{1}{\epsilon} - 1} \left(v_{1,r} + A^T v_{r+1,d}\right) \Sigma_{1,r}^{-1} \theta_{1,r},
\]

where the inequality follows from Eq. (12) and the last equality follows from a standard formula for the support function of an ellipse. Next, this last optimization is equivalent to

\[
\max_{\theta} \ v^T \theta \\
\text{s.t. } (\theta - \mu_N)^T \Sigma_N^{-1} (\theta - \mu_N) \leq \frac{1}{\epsilon} - 1 \\
\theta_{r+1,d} = \beta + A \theta_{1,r},
\]

by definition of the approximate inverse \( \Sigma_N^{-1} \) Eq. (15). Finally, since Eq. (14) holds almost surely, it must hold for all \( \theta \in \Theta \), i.e., \( \{ \theta \in \mathbb{R}^d : \theta_{r+1,d} = \beta + A \theta_{1,r} \} \subseteq \Theta \). Combining Lemma 1 with Thm. 1 proves the result.
Proof of Thm. 4. For the first part of the theorem, we show first that $\text{VaR}^\epsilon_{\hat{\theta}|S}(v)$ is convex. Note $\text{VaR}^\epsilon_{\hat{\theta}|S}(0) = 0$. Assuming $v \neq 0$, we have two cases.

First suppose $v_1 > v_2$. Then $\mathbb{P}(v_1 \hat{\theta}_1 + v_2 \hat{\theta}_2 \leq t \mid S) = \mathbb{P}(\theta_1 \leq \frac{t - v_2}{v_1 - v_2} \mid S)$ using the fact that $e^T \hat{\theta} = 1$ almost surely. Note that since $\hat{\theta}|S$ is Dirichlet with parameters $(\alpha_1, \alpha_2)$, we have $\theta_1|S$ follows a Beta distribution with parameter $(\alpha_1, \alpha_2)$. Thus, setting this probability equal to $1 - \epsilon$ and solving for $t$ yields, $\text{VaR}^\epsilon_{\hat{\theta}|S}(v) = v_2 + (v_1 - v_2)\beta_1 - \epsilon(\alpha_1, \alpha_2)$.

Next suppose $v_1 < v_2$. Again using $e^T \hat{\theta} = 1$ and rearranging terms yields
\[
\mathbb{P}(v_1 \hat{\theta}_1 + v_2 \hat{\theta}_2 \leq t \mid S) = 1 - \epsilon \iff \mathbb{P}\left(\hat{\theta}_1 \leq \frac{t - v_2}{v_2 - v_1} \mid S\right) = \epsilon.
\]
Solving for $t$ yields, $\text{VaR}^\epsilon_{\hat{\theta}|S}(v) = v_2 + (v_2 - v_1)\beta_1 - \epsilon(\alpha_1, \alpha_2)$.

Notice that for $0 < \epsilon < 0.5$, we have $0 < \beta_1 - \epsilon(\alpha_1, \alpha_2) \leq \beta_1 - \epsilon(\alpha_1, \alpha_2)$. Consequently, we can combine these two cases to write
\[
\text{VaR}^\epsilon_{\hat{\theta}|S}(v) = v_2 + \max((v_1 - v_2)\beta_1 - \epsilon(\alpha_1, \alpha_2), (v_2 - v_1)\beta_1 - \epsilon(\alpha_1, \alpha_2)).
\]  
(21)
As the maximum of two linear functions, $\text{VaR}^\epsilon_{\hat{\theta}}(v)$ is convex.

To obtain the representation for the optimal ambiguity set, we reconsider the case $v_1 < v_2$. Indeed, instead of proceeding as before, we treat this case symmetrically to the case $v_1 > v_2$ above to yield $\text{VaR}^\epsilon_{\hat{\theta}|S}(v) = v_1 + (v_2 - v_1)\beta_1 - \epsilon(\alpha_2, \alpha_1)$. Utilizing this alternate expression yields:
\[
\text{VaR}^\epsilon_{\hat{\theta}|S}(v) = \max(\beta_1 - \epsilon(\alpha_1, \alpha_2)v_1 + (1 - \beta_1 - \epsilon(\alpha_1, \alpha_2))v_2, (1 - \beta_1 - \epsilon(\alpha_2, \alpha_1))v_1 + \beta_1 - \epsilon(\alpha_2, \alpha_1)v_2).
\]  
(22)
One can check directly that the support of $P^\epsilon_\tau(S)$ is given by Eq. (22). This concludes the first part of the theorem.

The second part of the theorem can be validated numerically. For most examples, one can observe non-convexity along the line $\gamma \mapsto \text{VaR}^\epsilon_{\hat{\theta}|S}(1 - \gamma, 1 + \gamma, 0, \ldots, 0)$ for $\gamma$ slightly positive and slightly negative. We prove this formally in the special when $S$ is such that $\alpha = (1, 1, 1, 0, \ldots, 0)$. (Note this requires that we take an improper prior $\tau = 0$.)

By the merging property of the Dirichlet distribution, the random vector $(\hat{\theta}_1, \hat{\theta}_2, \sum_{i=3}^d \hat{\theta}_i)$ also has a Dirichlet distribution with parameter $(1, 1, 1)$, i.e., it is uniform over the simplex. For $0 < v_1 < t < v_2$ we can compute $\mathbb{P}(v_1 \hat{\theta}_1 + v_2 \hat{\theta}_2 \leq t)$ directly by integration,
\[
\mathbb{P}(v_1 \hat{\theta}_1 + v_2 \hat{\theta}_2 \leq t \mid S) = \frac{t^2 - 2tv_2 + v_1v_2}{v_2(v_1 - v_2)}.
\]
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 $0 < v_1 < v_2$, $\text{VaR}^\epsilon_{\hat{\theta}|S}(v_1, v_2, 0) = v_2 - \sqrt{\epsilon} \sqrt{v_2(v_1 - v_2)}$. This computation was entirely symmetric in $v_1, v_2$, so that when $0 < v_2 < v_1$, we have $\text{VaR}^\epsilon_{\hat{\theta}|S}(v_1, v_2, 0) = v_1 - \sqrt{\epsilon} \sqrt{v_1(v_2 - v_1)}$. One can then check directly that for $\gamma > 0$ sufficiently small, $\text{VaR}^\epsilon_{\hat{\theta}|S}(1 - \gamma, 1 + \gamma, 0) > \text{VaR}^\epsilon_{\hat{\theta}|S}(1 + 2\gamma, 1 - 2\gamma, 0) = \text{VaR}^\epsilon_{\hat{\theta}|S}(1 - 2\gamma, 1 + 2\gamma, 0)$. Thus, $\text{VaR}^\epsilon_{\hat{\theta}|S}(v)$ is non-convex. 

\(\square\)
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 \( Y_i \sim \text{Gamma}(\alpha, 1) \). Then \( (\tilde{Y}_1 / \sum_{i=1}^d \tilde{Y}_i, \ldots, \tilde{Y}_d / \sum_{i=1}^d \tilde{Y}_i) \) has Dirichlet distribution with parameter \( \alpha \).

We can now upperbound \( \text{VaR}_{\tilde{\theta}|S}(v) \) using a technique similar to [Nemirovski and Shapiro 2006]:

\[
\mathbb{P}(v^T \tilde{\theta} \leq t) = \mathbb{P} \left( \sum_{i=1}^d v_i \tilde{Y}_i \leq t \sum_{i=1}^d \tilde{Y}_i \right) \\
= \mathbb{P} \left( \sum_{i=1}^d (v_i - t) \tilde{Y}_i \leq 0 \right) \\
\leq \inf_{\lambda > \Delta} \prod_{i=1}^d \mathbb{E}[e^{v_i - t \tilde{Y}_i}] \\
= \inf_{\lambda > \Delta} \prod_{i=1}^d \left( 1 - \frac{v_i - t}{\lambda} \right)^{-\alpha_i},
\]

where the inequality follows from Markov’s inequality and the independence of the \( \tilde{Y}_i \), and the last equality follows from the formula for the moment generating function of a Gamma random variable. Throughout, \( \Delta \equiv \max_j (v_j - t)^+ \).

It follows that \( \text{VaR}_{\tilde{\theta}|S}(v) \leq t \) if there exists \( \lambda > \Delta \) such that \( \prod_{i=1}^d \left( 1 - \frac{v_i - t}{\lambda} \right)^{-\alpha_i} \leq \epsilon \), or, equivalently,

\[
\inf_{\lambda > \Delta} \frac{\log(1/\epsilon)}{\alpha_0} - \lambda \sum_{i=1}^d \mu_{N,i} \log \left( 1 - \frac{v_i - t}{\lambda} \right) \leq 0.
\]

Using Theorem 1 of [Ben-Tal et al. 2013], we recognize this inequality as \( \delta^*(v - te|Q) \leq 0 \) for \( Q = \{ \theta : \sum_{i=1}^d \mu_{N,i} \log \left( \frac{\mu_{N,i}}{\mu_{N,i}} \right) \leq \frac{\log(1/\epsilon)}{\alpha_0} \} \).

Finally, observe that since \( \theta \geq 0 \) for all \( \theta \in Q \), by rescaling we have

\[
(v - te)^T \theta \leq 0 \quad \forall \theta \in Q \iff (v - te)^T \theta \leq 0 \quad \forall \theta \in Q \cap \{ \theta : e^T \theta = 1 \},
\]

\[
\iff v^T \theta \leq t \quad \forall \theta \in Q \cap \{ \theta : e^T \theta = 1 \}.
\]

We recognize this last set as \( \mathcal{P}^{KL}(\mu_N, \sqrt{\frac{\log(1/\epsilon)}{\alpha_0}}) \). Since this inequality holds for arbitrary \( (v, t) \), we have shown \( \text{VaR}_{\tilde{\theta}|S}(v) \leq \delta^* \left( v | \mathcal{P}^{KL}(\mu_N, \sqrt{\frac{\log(1/\epsilon)}{\alpha_0}}) \right) \), completing the proof.

Proof of Thm. 7. Using the explicit formula for the value at risk of a Gaussian and Thm. 6, we have for any \( v \in \mathbb{R}^d, \|v\| = 1 \),

\[
\text{VaR}_{\tilde{\theta}|S}(v) \rightarrow v^T \theta^* + \frac{z_{1-\epsilon}}{\sqrt{N}} \sqrt{v^T \mathcal{I}^*(\theta^*)} v, \quad \text{a.s.}
\]

Since the convergence of Thm. 6 is in total variation, this limit holds uniformly in \( v \), i.e.,

\[
\sup_{v \in \mathbb{R}^d : \|v\| = 1} \left| \text{VaR}_{\tilde{\theta}|S}(v) - v^T \theta^* - \frac{z_{1-\epsilon}}{\sqrt{N}} \sqrt{v^T \mathcal{I}^*(\theta^*)} v \right| \rightarrow 0 \quad \text{a.s.}
\]
Thus, to prove Eq. (19), it suffices to show that
\[
\sup_{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\| = 1} \left| \mathbf{v}^T (\mathbf{\theta}^* - \mathbf{\mu}_N) - \frac{z_{1-\epsilon}^2}{N} \sqrt{\mathbf{v}^T \mathcal{I}(\mathbf{\theta}^*)} \mathbf{v} - z_{1-\epsilon} \sqrt{\mathbf{v}^T \Sigma_N \mathbf{v}} \right| \to 0, \quad \text{a.s.} \tag{23}
\]
From the Cauchy-Schwartz inequality, when \(\|\mathbf{v}\| = 1\),
\[
\|\mathbf{v}^T (\mathbf{\theta}^* - \mathbf{\mu}_N)\| \leq \|\mathbf{v}\| \|\mathbf{\theta}^* - \mathbf{\mu}_N\| = \|\mathbf{\theta}^* - \mathbf{\mu}_N\| \to 0 \quad \text{a.s.} \tag{24}
\]
where the last limit follows because \(\mathbf{\mu}_N \to \mathbf{\theta}^*\) almost surely, by Thm. 6. Similarly, when \(\|\mathbf{v}\| = 1\)
\[
\frac{1}{N} \mathbf{v}^T \mathcal{I}(\mathbf{\theta}^*) \mathbf{v} - \mathbf{v}^T \Sigma_N \mathbf{v} \leq \frac{1}{N} \mathcal{I}(\mathbf{\theta}^*) - \Sigma_N \|\mathbf{v}\|^2
\]
\[
= \frac{1}{N} \|\mathcal{I}(\mathbf{\theta}^*) - \Sigma_N\| \to 0 \quad \text{a.s.},
\]
where the last limit follows from Thm. 6. Combining Eq. (24) and Eq. (25) proves Eq. (23), which proves Eq. (19).

For the second part of the theorem, notice that since \(\mathbf{\theta}^* \in \text{int}(\Theta)\) and \(\mathbf{\mu}_N \to \mathbf{\theta}^*\) almost surely, \(\mathcal{P}(\mathbf{\mu}_N, \Sigma_N, z_{1-\epsilon}(1 + \kappa)) \subset \text{ri}(\Theta)\) for \(N\) sufficiently large. It follows that for all \(\mathbf{v} \in \mathbb{R}^d\),
\[
\delta^*(\mathbf{v}|\mathcal{P}(\mathbf{\mu}_N, \Sigma_N, z_{1-\epsilon}(1 + \kappa))) = \mathbf{v}^T \mathbf{\mu}_N + (1 + \kappa) z_{1-\epsilon} \sqrt{\mathbf{v}^T \Sigma_N \mathbf{v}}.
\]
Thus, for \(N\) sufficiently large, \(\delta^*(\mathbf{v}|\mathcal{P}(\mathbf{\mu}_N, \Sigma_N, z_{1-\epsilon}(1 + \kappa)))\) upper bounds \(\text{VaR}_{\tilde{\theta}_i}^\epsilon(\mathbf{v})\) uniformly for all \(\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\| = 1\), whereby from Eq. (11), \(\mathcal{P}(\mathbf{\mu}_N, \Sigma_N, z_{1-\epsilon}(1 + \kappa))\) satisfies the posterior guarantee. This proves the second statement of the theorem.

Finally, for the last statement, as above, notice that for \(N\) sufficiently large, \(\mathcal{P}(\mathbf{\mu}_N, \Sigma_N, z_{1-\epsilon}(1 - \kappa)) \subset \text{int}(\Theta)\) and for all \(\mathbf{v} \in \mathbb{R}^d\)
\[
\delta^*(\mathbf{v}|\mathcal{P}(\mathbf{\mu}_N, \Sigma_N, z_{1-\epsilon}(1 - \kappa))) = \mathbf{v}^T \mathbf{\mu}_N + (1 - \kappa) z_{1-\epsilon} \sqrt{\mathbf{v}^T \Sigma_N \mathbf{v}}.
\]
Consequently, for \(N\) sufficiently large, for all \(\mathbf{v} \in \mathbb{R}^d \) s.t. \(\|\mathbf{v}\| = 1\), \(\delta^*(\mathbf{v}|\mathcal{P}(\mathbf{\mu}_N, \Sigma_N, z_{1-\epsilon}(1 - \kappa))) \leq \text{VaR}_{\tilde{\theta}_i}^\epsilon(\mathbf{v})\), whereby \(\mathcal{P}(\mathbf{\mu}_N, \Sigma_N, z_{1-\epsilon}(1 - \kappa))\) is a subset of any ambiguity set which satisfies the posterior guarantee. This completes the theorem.

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

**Proof of Thm. 9** Writing a Taylor expansion of \(\mu_{N,i}\log(\mu_{N,i}/\theta_i) - \mu_{N,i} + \theta_i\) around \(\theta_i = \mu_{N,i}\),
we obtain for any \(\mathbf{\theta} \in \mathcal{P}^\text{KL}(\mathbf{\mu}_N, \sqrt{\log(1/\epsilon)/\sigma_0})\)
\[
\sum_{i=1}^d \frac{(\mu_{N,i} - \theta_i)^2}{2\mu_{N,i}} - \sum_{i=1}^d \mu_{N,i} \log(\mu_{N,i}/\theta_i) = \frac{1}{6} \sum_{i=1}^d \frac{(\mu_{N,i} - q_i)^3}{2\mu_{N,i}^2}, \tag{26}
\]
for some $q \in \mathcal{P}^{KL}(\mu_N, \sqrt{\log(1/\epsilon) \alpha_0})$ given by the mean-value theorem. Next, note that
\[
\|\mu_N - q\|_3 \leq \|\mu_N - q\|_1 \quad \text{(Monotonicity of norms)}
\leq \sqrt{\sum_{i=1}^{d} \mu_{N,i} \log(\mu_{N,i}/q_i) - \mu_{N,i} + q_i} \quad \text{(Plinsker’s Inequality)}
\leq \sqrt{2 \log(1/\epsilon) / \alpha_0},
\]
(since $q \in \mathcal{P}^{KL}(\mu_N, \sqrt{\log(1/\epsilon) \alpha_0})$)\tag{27}

Since $\theta^* \in \text{int}(\Theta) \implies \theta^* > 0$, and $\mu_N \to \theta^*$ almost surely, there exists a $\kappa' > 0$ such that for all $N$ sufficiently large, $\mu_N > \kappa'e$. Combining this observation with (27), for $N$ sufficiently large, we have
\[
\frac{1}{6} \sum_{i=1}^{d} \frac{(\mu_{N,i} - q_i)^3}{\mu_{N,i}^2} \leq \frac{1}{6\kappa'^2} \left( \frac{2 \log(1/\epsilon)}{N} \right)^{3/2} \equiv R_N
\leq \sqrt{2 \log(1/\epsilon) / \alpha_0} + R_N.
\]
Consequently, for any $\theta \in \mathcal{P}^{KL}(\mu_N, \sqrt{\log(1/\epsilon) \alpha_0})$
\[
\sum_{i=1}^{d} \frac{(\mu_{N,i} - \theta_i)^2}{2\mu_{N,i}} \leq \frac{2 \log(1/\epsilon)}{\alpha_0} + 2R_N. \tag{28}
\]
Fix $\kappa > 0$. Then for $N$ sufficiently large,
\[
\sum_{i=1}^{d} \frac{(\mu_{N,i} - \theta_i)^2}{\mu_{N,i}} \leq (\lambda + \kappa)^2 \frac{z^2_{1-\epsilon}}{N},
\]
for $\lambda = \sqrt{\frac{2 \log(1/\epsilon)}{z_{1-\epsilon}}}$. This proves the theorem.

Proof of Thm. 10 Suppose by contradiction that $\mathcal{P}_\epsilon(S) - \mu_N \leq \alpha \left( \mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon}) - \mu_N \right)$ for $\alpha < \frac{(1-\kappa)\sqrt{\chi_{r,1-\epsilon}}}{\sqrt{z_{1-\epsilon}}}$. By continuity, there exists $\kappa' > 0$ such that
\[
(1-\kappa)\sqrt{\chi_{r,1-\epsilon}} \leq \sqrt{\chi_{r,1-\epsilon-\kappa'}}.
\]
Since $\mathcal{P}_\epsilon(S)$ is a credible region,
\[
1 - \epsilon \leq \mathbb{P}(\tilde{\theta} \in \mathcal{P}_\epsilon(S) | S)
\leq \mathbb{P}(\tilde{\theta} - \mu_N \in \alpha \left( \mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon}) - \mu_N \right) | S)
\leq \mathbb{P} \left( (\tilde{\theta} - \mu_N)^T \Sigma_N^{-1} (\tilde{\theta} - \mu_N) \leq \alpha^2 z^2_{1-\epsilon} | S \right)
\]
For $N$ sufficiently large, from Thm. 6 this last probability is less than or equal $\kappa' + \mathbb{P}(\zeta^T \mathcal{I}(\theta)^{-1} \zeta \leq \alpha^2 z_{1-\epsilon}^2)$ where $\zeta$ is a $\mathcal{N}(0, \mathcal{I}(\theta^*))$ random variable. Notice $\zeta^T \mathcal{I}(\theta^*)^{-1} \zeta$ is therefore a $\chi^2$ random variable. Rearranging terms yields,

$$\chi^2_{r,1-\epsilon-\kappa'} \leq \alpha^2 z_{1-\epsilon}^2. \iff \frac{\sqrt{\chi^2_{r,1-\epsilon-\kappa'}}}{z_{1-\epsilon}} \leq \alpha,$$

a contradiction.  \qed

**Proof of Corollary 2.** Integrate $\mathbb{P}(\theta^* \in \mathcal{P}_\epsilon(S))$ against the prior density of $\tilde{\theta}$ yielding $\mathbb{P}(\tilde{\theta} \in \mathcal{P}_\epsilon(S)) \geq 1 - \epsilon$. Thus, it cannot be that $\mathbb{P}(\tilde{\theta} \in \mathcal{P}_\epsilon(S) | S) < 1 - \epsilon$ for all $S$, i.e., for some $S$, $\mathcal{P}_\epsilon(S)$ must be a credible region. The result follows from Thm. 10.  \qed

**Proof of Thm. 11.** Consider $y \in \mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon}) - \mu_N$. Then,

$$\frac{z_{1-\epsilon}^2}{\alpha_0 + 1} \geq \sum_{i=1}^d \frac{y_i^2}{\mu_{N,i}} \geq \|y\|_2^2 \geq \|y\|_\infty^2. \tag{29}$$

Thus,

$$|y_i| \leq \sqrt{\frac{z_{1-\epsilon}^2}{\alpha_0 + 1}} = O\left(\frac{1}{\sqrt{N}}\right), \text{ for } i = 1, \ldots, d. \tag{30}$$

We seek $\alpha$ such that $\mu_N + \alpha y \in \mathcal{P}^\phi(\mu_N, \sqrt{\phi'(1)\kappa^2_{d-1,1-\epsilon}/2N})$ for $N$ sufficiently large. To this end, write

$$\sum_{i=1}^d \mu_{N,i} \phi\left(\frac{\mu_{N,i} + \alpha y_i}{\mu_{N,i}}\right) = \sum_{i=1}^d \mu_{N,i} \phi\left(1 + \frac{\alpha y_i}{\mu_{N,i}}\right).$$

$$= \sum_{i=1}^d \mu_{N,i} \phi'(1) \frac{\alpha y_i}{\mu_{N,i}} + \sum_{i=1}^d \frac{1}{2} \mu_{N,i} \phi''(1) \frac{\alpha^2 y_i^2}{\mu_{N,i}^2} + \sum_{i=1}^d O(\alpha^3 y_i^3) \mu_{N,i},$$

where the last line follows from a Taylor expansion of $\phi(t)$ around $t = 1$. The first term disappears since $\mu_N + y \in \mathcal{P}(\mu_N, \Sigma_N, z_{1-\epsilon}) \implies \mathbf{e}^T y = 0$ and the last term is bounded by $\alpha^3 O(N^{-3/2})$ by Eq. (30), yielding

$$\frac{\phi'(1)\kappa^2_{d-1,1-\epsilon}/2N}{2N} + O(N^{-3/2}).$$

Taking $\alpha = (1 - \kappa) \sqrt{\frac{\chi^2_{d-1,1-\epsilon}}{z_{1-\epsilon}}}$ yields

$$\sum_{i=1}^d \mu_{N,i} \phi\left(\frac{\mu_{N,i} + \alpha y_i}{\mu_{N,i}}\right) \leq (1 - \kappa)^2 \frac{\phi'(1)\chi^2_{d-1,1-\epsilon}}{2N} + O(N^{-3/2}),$$

so that for $N$ sufficiently large, $\mu_N + \alpha y \in \mathcal{P}^\phi(\mu_N, \sqrt{\phi'(1)\kappa^2_{d-1,1-\epsilon}/2N})$, as was to be proven.  \qed

Gupta: Near Optimal Ambiguity Sets  
Operations Research 00(0), pp. 000–000, © 0000 INFORMS.
**Proof of Thm. 12** Suppose $y \in \mathcal{P}(\mu, \Sigma, z_{1-\epsilon}) - \mu_N$. We require $\lambda$ such that $\mu_N + \lambda y \in \Delta_d$. Since $e^T y = 0$, this only requires that $\lambda y_i > -\mu_{N,i}$ for all $i$, which is equivalent to

$$\lambda \leq \frac{\mu_{N,i}}{|y_i|} \quad \forall i \text{ s.t. } y_i < 0.$$ 

As in the proof of Thm. 11, $|y_i| = O(N^{-1/2})$. Moreover, since $\theta^* \in \text{int}(\Theta)$ and $\mu_N \rightarrow \theta^*$ almost surely,

$$\min_{i: y_i < 0} \mu_{N,i} \geq (1 - \kappa) \theta^*_{\text{min}}$$

for $N$ sufficiently large. This proves the theorem. \qed

**Proof of Thm. 13** Suppose the theorem does not hold, define $g(\theta, x) \equiv (\theta - \theta^*)^T \mathcal{I}(\theta^*)^{-1} (\theta - \theta^*) - \hat{x}^2$, and let $\hat{x}$ be robust feasible. Notice that $g(\theta, x)$ is a convex quadratic in $\theta$.

From the posterior feasibility guarantee,

$$1 - \epsilon \leq \mathbb{P}(g(\tilde{\theta}, \hat{x}) \leq 0 \mid S) = \mathbb{P}((\tilde{\theta} - \theta^*)^T \mathcal{I}(\theta^*)^{-1} (\tilde{\theta} - \theta^*) \leq \hat{x}^2).$$

Thus, using Thm. 6 for any $\kappa > 0$, for all $N$ sufficiently large,

$$\hat{x}^2 \geq \frac{\chi_{d,1-\epsilon-\kappa}^2}{N}. \quad (31)$$

On the other hand, since $\mathcal{P}_e(S) \subseteq (1 - \kappa) \mathcal{P} \left( \mu_N, \sqrt{\frac{\chi_{d,1-\epsilon}^2}{N}} \right)$, there must exist robust feasible $\hat{x}$ such that

$$\hat{x}^2 \leq \sup_{\theta \in (1 - \kappa) \mathcal{P}} (\theta - \theta^*)^T \mathcal{I}(\theta^*)^{-1} (\theta - \theta^*)$$

$$\leq (\mu_N - \theta^*)^T \mathcal{I}(\theta^*)^{-1} (\mu_N - \theta^*) + \sup_{\theta \in (1 - \kappa) \mathcal{P}} (\theta - \mu_N)^T \mathcal{I}(\theta^*)^{-1} (\theta - \mu_N),$$

where the last line follows from the triangle inequality. Notice next that by Thm. 6 the first term tends to zero as $N$ gets large. The second term is at most $(1 - \kappa) \chi_{d,1-\epsilon}^2 / N$. Thus, for $N$ sufficiently large, $\hat{x}^2 < (1 - \kappa/2) \chi_{d,1-\epsilon}^2 / N$. Taking the limit as $\kappa' \rightarrow 0$ in Eq. (31) yields a contradiction. \qed