

Bayesian distributionally robust variational inequalities: regularization and quantification

Wentao Ma

School of Mathematics and Statistics, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an, Shaanxi 710049, China; Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China, wentao.ma@connect.polyu.hk

Zhiping Chen

School of Mathematics and Statistics, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an, Shaanxi 710049, China, zchen@mail.xjtu.edu.cn

Xiaojun Chen

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China, maxjchen@polyu.edu.hk

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and are not intended to be a true representation of the article's final published form. Use of this template to distribute papers in print or online or to submit papers to another non-INFORM publication is prohibited.

Abstract. We propose a Bayesian distributionally robust variational inequality (DRVI) framework that models the data-generating distribution through a finite-mixture family, which allows us to study the DRVI on a tractable finite-dimensional parametric ambiguity set. To address distributional uncertainty, we construct a data-driven ambiguity set with posterior coverage guarantees via Bayesian inference. We also employ a regularization approach to ensure numerical stability. We prove the existence of solutions to the Bayesian DRVI and the asymptotic convergence to a solution as sample size grows to infinity and the regularization parameter goes to zero. Moreover, we derive quantitative stability bounds and finite-sample guarantees under data scarcity and contamination. Numerical experiments on a distributionally robust multi-portfolio Nash equilibrium problem validate our theoretical results and demonstrate the robustness and reliability of Bayesian DRVI solutions in practice.

Funding: This work was supported in part by grants from the Hong Kong Research Grants Council (PolyU15300123, PolyU15300124) and the National Key R&D Program of China (2022YFA1004000, 2022YFA1004001).

Key words: distributional robustness; Bayesian asymptotics; data-driven; regularization method; quantitative stability

MSC2000 subject classification: Primary: 90C15; secondary: 90C33

OR/MS subject classification: Primary: Programming: Stochastic; secondary: Programming: Complementarity

1. Introduction Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty closed convex set. For any $x \in \mathcal{X}$, let $\mathcal{N}_{\mathcal{X}}(x) := \{v \in \mathbb{R}^d \mid \langle v, y - x \rangle \leq 0, \forall y \in \mathcal{X}\}$ denote the normal cone to $\mathcal{X}$ at x . The stochastic variational

inequality (SVI) [23, 49] aims to find a vector $x^* \in \mathcal{X}$ such that

$$0 \in \mathbb{E}_{P^c} [\Phi(x^*, \xi)] + \mathcal{N}_{\mathcal{X}}(x^*), \quad (1)$$

where $\xi : \Omega \rightarrow \mathbb{R}^m$ is a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ whose true distribution $P^c := \mathbb{P} \circ \xi^{-1}$ is supported on $\Xi \subseteq \mathbb{R}^m$. The mapping $\Phi : \mathcal{X} \times \Xi \rightarrow \mathbb{R}^d$ is continuous in x for any fixed $\xi \in \Xi$ and measurable and integrable in ξ for any fixed $x \in \mathcal{X}$. We use $\mathbb{E}_{P^c}$ to emphasize that the expectation is with respect to P^c . In many applications of interest, particularly in multi-agent equilibrium problems, solutions are characterized by a coupled system of best-response optimality conditions and generally cannot be reformulated as the minimization of a single objective [17]. Such coupled equilibrium systems are naturally modeled within the SVI framework, with representative examples including traffic equilibria [9, 44], oligopolistic competition in the crude oil market [27], and forward-contracting equilibria in electricity markets [56].

In practice, the true distribution P^c is unknown. Decision-makers therefore seek solutions that remain reliable under distributional uncertainty [13], which may stem from policy shifts, macroeconomic cycles, or market dynamics [21]. Distributionally robust optimization (DRO) addresses this challenge by replacing P^c with an ambiguity set $\mathcal{P}$ designed to contain P^c . This approach is well-founded in risk-sensitive decision-making, as many coherent risk measures admit dual representations closely related to DRO formulations [41, 50]. The combination of DRO with SVI yields the distributionally robust variational inequality (DRVI) problem [26, 60], which aims to find a pair (x^*, P^*) such that

$$\begin{aligned} 0 &\in \mathbb{E}_{P^*} [\Phi(x^*, \xi)] + \mathcal{N}_{\mathcal{X}}(x^*), \\ P^* &\in \arg \max_{P' \in \mathcal{P}} \mathbb{E}_{P'} [\varphi(x^*, \xi)], \end{aligned} \quad (2)$$

where $\varphi : \mathcal{X} \times \Xi \rightarrow \mathbb{R}$ is continuous in x for any fixed $\xi \in \Xi$ and measurable and integrable in ξ for any fixed $x \in \mathcal{X}$. Representative choices of φ can be found in [26, 60]. Common choices for ambiguity sets include those based on first- and second-order moments [14], unimodality [24], ϕ -divergence [12, 32] and Wasserstein distance [20, 30]. Despite their theoretical superiority in robustness and interpretability, the construction and computational tractability of such infinite-dimensional ambiguity sets remain challenging and often lead to the curse of dimensionality and prohibitively large-scale optimization problems [4].

As demonstrated in prior work (e.g., [38, 58, 66]), it is natural to model ξ via a parametric family of distributions $\mathcal{P}_{\Theta} := \{P_{\theta} \mid \theta \in \Theta\}$, where each distribution P_{θ} is uniquely determined by a parameter

θ through the probability density function (pdf) $f(\cdot|\theta)$. Specifically, for any Borel set $A \subseteq \Xi$, $P_\theta(A) = \int_A f(\xi | \theta) d\xi$. Here, $\Theta \subseteq \mathbb{R}^n$ is a nonempty closed parameter set. Throughout the main development, we adopt a well-specified parametric setting and assume that there exists $\theta^c \in \Theta$ such that $P^c = P_{\theta^c}$. In misspecified settings, θ^c may instead be interpreted as the parameter associated with the Kullback–Leibler (KL) projection of P^c onto $\mathcal{P}_\Theta$ [58]. Such parametric assumptions are common in SVI applications, including Gamma and Gaussian families in game theory [5, 37], uniform distributions in system control [18], and Beta distributions in traffic flow management [10]. However, real-world data often exhibit multimodality, skewness, and heavy-tailed behavior [69], which may be inadequately captured by a single parametric family. To enhance modeling flexibility while retaining a parametric structure, we work with a finite-mixture family of distributions. Theoretically, under mild regularity conditions, any continuous pdf can be approximated arbitrarily well by a finite mixture of Gaussian kernels [16, 40]. Practically, mixture models have been applied to various areas, including asset returns in portfolio selection [11, 68], rewards in multi-armed bandit problems [62], and customer preference heterogeneity in marketing analytics [53]. Thus, the finite-mixture family serves both as a structured parametric representation and as a low-dimensional approximation class. Accordingly, we consider the finite-mixture family $P_\theta = \sum_{j=1}^n \theta_j Q_j$, where $\{Q_j\}_{j=1}^n$ are fixed, prespecified component distributions and Θ is the probability simplex, i.e., $\Theta := \{\theta \mid e^\top \theta = 1, \theta \in \mathbb{R}_+^n\}$, where $e \in \mathbb{R}^n$ denotes the all-ones vector. Let $\theta^c = (\theta_1^c, \dots, \theta_n^c)$ be the corresponding mixture weight vector and $\mathbb{B}(\theta^c, r^c) := \{\theta \mid \|\theta - \theta^c\|_\infty \leq r^c\} \subseteq \mathbb{R}^n$ be an ℓ_∞ norm ball centered at θ^c with radius $r^c > 0$. Motivated by parametric DRO models [22, 69], we propose the Bayesian DRVI formulation as

$$\begin{aligned} 0 &\in \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\Phi(x, \xi)] + \mathcal{N}_X(x), \\ \theta &\in \arg \max_{\theta' \in \Theta^c} \sum_{j=1}^n \theta'_j \mathbb{E}_{Q_j}[\varphi(x, \xi)], \end{aligned} \tag{B^c}$$

where $\Theta^c := \mathbb{B}(\theta^c, r^c) \cap \Theta$ is the true parametric ambiguity set. This parametric ambiguity set is defined over mixture weights in a fixed-dimensional parameter set, which is computationally attractive when compared with ambiguity sets whose effective dimension grows with the sample size. This computational benefit comes with the need to prescribe or estimate the component distributions $Q_1, \dots, Q_n$; see, e.g., [7, 36, 63]. For complex unstructured data such as language, one may instead consider DRO formulations built around application-specific topic or subpopulation

structure [45]. The prespecified radius r^c captures the decision-maker's tolerance for model perturbation or risk aversion. In particular, when $r^c = 0$, Θ^c degenerates to $\{\theta^c\}$ and $(\mathbf{B}^c)$ reduces to the risk-neutral SVI (1) under the distribution P_{θ^c} .

The true ambiguity set Θ^c in $(\mathbf{B}^c)$ is determined by θ^c , which is fixed but unknown and must be estimated from limited sample data $\mathcal{S}^N = \{\xi^1, \dots, \xi^N\}$, where $\xi^1, \dots, \xi^N$ are independent and identically distributed (i.i.d.) samples generated from the true distribution P_{θ^c} . Conventional approaches construct an approximation by replacing θ^c with the maximum likelihood estimate (MLE) $\hat{\theta}_N$ [54] and then taking the set $\mathbb{B}(\hat{\theta}_N, r^c) \cap \Theta$. However, this MLE approximation can be unstable due to finite-sample noise and model sensitivity, which may undermine robustness. To address this, we adopt a Bayesian perspective that explicitly quantifies parametric uncertainty through posterior distributions [58]. In the Bayesian interpretation adopted here, θ^c is treated as a realization of a belief random variable, whose prior pdf $p(\theta)$ represents our belief before observing data [66]. As new data arrive, this prior is updated recursively to refine our uncertainty quantification for θ^c . Formally, applying Bayes' rule to $\mathcal{S}^N$ yields the pdf of the posterior distribution $\mathbb{P}_{\mathcal{S}^N}$ as:

$$p(\theta|\mathcal{S}^N) = \frac{f(\mathcal{S}^N|\theta)p(\theta)}{\int_{\Theta} f(\mathcal{S}^N|\theta')p(\theta')d\theta'}, \quad (3)$$

where $f(\mathcal{S}^N|\theta) := \prod_{i=1}^N f(\xi^i|\theta)$. This posterior distribution coherently integrates prior knowledge—whether from historical data, expert judgment, or physical models—and new observations, providing an evolving quantification of parametric uncertainty. In the absence of informative priors, noninformative choices such as Jeffreys' or uniform priors are commonly used. Recent studies [25, 59, 66] replace ambiguity sets with Bayesian averaging to characterize uncertainty in the model parameter through the posterior distribution. Further, [58] constructs an ambiguity set around the corresponding parametric reference distribution using KL divergence, thereby adopting a worst-case perspective on Bayesian averaging to hedge against misspecification of the assumed parametric model. These methodologies lead to different tradeoffs. Bayesian averaging is typically less conservative but may lack robustness, whereas [58] provides additional protection at the cost of greater computational complexity. Our approach can be viewed as a compromise between these two perspectives: it retains a worst-case treatment of uncertainty, but does so within the parametric family. This allows the framework to preserve tractability while integrating naturally with the DRVI structure.

From the perspective of ambiguity-set construction, our approach follows the general Bayesian framework in [22], which studies ambiguity sets for general parametric distribution families characterized by finite-dimensional parameters. The construction developed below specializes this general

framework to the finite-mixture setting underlying our DRVI model, where the corresponding parameters are the mixture weights. Accordingly, our goal is to construct a data-driven ambiguity set $\hat{\Theta}_N \subseteq \Theta$ that captures parametric uncertainty as reflected by the prior knowledge and the dataset $\mathcal{S}^N$ (equivalently, by the posterior $\mathbb{P}_{\mathcal{S}^N}$). We consider an ambiguity set of the form $\hat{\Theta}_N := \mathbb{B}(\hat{\theta}_N, \hat{r}_N) \cap \Theta$ and require it to contain the true ambiguity set Θ^c with high posterior probability. Specifically, for a prespecified confidence level $\alpha \in (0, 1)$, the nominal parameter $\hat{\theta}_N$ and radius $\hat{r}_N$ are chosen such that $\mathbb{P}_{\mathcal{S}^N} \{ \Theta^c \subseteq \hat{\Theta}_N \} \geq 1 - \alpha$. To avoid overly conservative decisions in the robust formulation [8], $\hat{r}_N$ is chosen as the smallest value that satisfies the posterior coverage guarantee:

$$\hat{r}_N = \min \{ r \mid r \geq 0, \mathbb{P}_{\mathcal{S}^N} \{ \Theta^c \subseteq \mathbb{B}(\hat{\theta}_N, r) \cap \Theta \} \geq 1 - \alpha \}.$$

Based on this construction, replacing Θ^c by its approximation $\hat{\Theta}_N$ yields the data-driven Bayesian counterpart of (B^c):

$$\begin{aligned} 0 &\in \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j} [\Phi(x, \xi)] + \mathcal{N}_X(x), \\ \theta &\in \arg \max_{\theta' \in \hat{\Theta}_N} \sum_{j=1}^n \theta'_j \mathbb{E}_{Q_j} [\varphi(x, \xi)]. \end{aligned} \tag{B_N}$$

To address the potential numerical instability arising from the non-uniqueness of solutions to the lower-level problem in (B_N), let $\| \cdot \|$ denote the Euclidean norm and $\lambda > 0$ be a regularization parameter. We introduce the following regularized variant of (B_N):

$$\begin{aligned} 0 &\in \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j} [\Phi(x, \xi)] + \mathcal{N}_X(x), \\ \theta &\in \arg \max_{\theta' \in \hat{\Theta}_N} \sum_{j=1}^n \theta'_j \mathbb{E}_{Q_j} [\varphi(x, \xi)] - \lambda \|\theta'\|^2. \end{aligned} \tag{B_N^\lambda}$$

Building on these formulations and the Bayesian construction, the main contributions of this paper are summarized as follows.

- **Modeling.** We propose a Bayesian DRVI framework that integrates Bayesian inference with DRVI to unify robustness and flexibility in a data-driven setting. By modeling the unknown true distribution within a finite mixture family, the resulting ambiguity set is defined over mixture weights and remains fixed-dimensional as sample size grows, which facilitates posterior calibration and computation. The framework captures parametric uncertainty through Bayesian updating, which integrates prior knowledge with observed data. Moreover, we introduce a regularization term in the lower-level maximization problem to improve numerical stability.

- **Analysis.** We show that the lower-level maximization problem over the Bayesian parametric ambiguity set can be reformulated as a risk-averse formulation with Bayesian risk measures [59, 66]. We prove that the solutions of (B_N) and (B_N^λ) converge to a solution of (B^c) as $N \rightarrow \infty$ and $\lambda \downarrow 0$. Moreover, we show that the unique maximizer of the regularized lower-level problem in (B_N^λ) converges to the least-norm maximizer of the unregularized problem in (B_N) as $\lambda \downarrow 0$. We further derive finite-sample guarantees and quantitative stability bounds that account for both data perturbation and regularization effects, addressing practical challenges in data-driven Bayesian DRVI estimation.

- **Applications.** We apply the proposed Bayesian DRVI framework to a distributionally robust multi-portfolio Nash equilibrium problem, building on Bayesian parametric ambiguity sets as in [68, 69]. Our numerical results show that (i) the method yields more accurate equilibrium solutions and lower utility error in finite samples than classical DRVI and risk-neutral SVI benchmarks, (ii) the data-driven equilibrium solutions converge to a solution of the Bayesian DRVI problem, and (iii) the method achieves favorable out-of-sample performance, balancing payoff and tail-risk mitigation. These findings underscore the practical relevance and applicability of our method in real-world decision-making scenarios.

The paper is structured as follows. Section 2 introduces a motivating Bayesian DRVI example and explores the construction of data-driven ambiguity sets from the frequentist and Bayesian perspectives. Section 3 proves the existence of solutions to the data-driven Bayesian DRVI problem and analyzes the convergence of solutions of (B_N^λ) to a solution as $N \rightarrow \infty$ and $\lambda \downarrow 0$. Section 4 derives quantitative stability bounds, providing finite-sample guarantees and robustness assessments under varying data regimes. Section 5 uses a distributionally robust multi-portfolio Nash equilibrium problem to show numerically the efficiency of the data-driven Bayesian DRVI with an extragradient method.

2. Model setup and background We begin by illustrating how Bayesian parametric ambiguity sets can be incorporated into the DRVI framework. The motivations and theoretical foundations of DRVI have been discussed in [26, 60]. For completeness, we provide an illustrative example below to show how Bayesian perspectives arise naturally in optimization and equilibrium problems.

EXAMPLE 1. Finite mixture models provide a flexible framework for modeling complex distributions such as asset returns [46]. Two common interpretations of mixture models in finance

are as follows: (i) if $Q_1, \dots, Q_n$ represent asset return distributions under different market regimes from scenario analysis, then $\theta_1, \dots, \theta_n$ are the occurrence probabilities of these regimes; (ii) if $Q_1, \dots, Q_n$ reflect asset returns generated by different predictive models, then $\theta_1, \dots, \theta_n$ quantify the credibility weights of the models. Let $\xi \sim \sum_{j=1}^n \theta_j^c Q_j$ be the random asset return vector, and let $x \in \mathcal{X}$ denote the corresponding portfolio weights. Following [69], the distributionally robust portfolio selection problem can be formulated as:

$$\min_{x \in \mathcal{X}} \max_{\theta \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j} [\varphi(x, \xi)], \quad (4)$$

where the loss (e.g., disutility) function $\varphi : \mathcal{X} \times \Xi \rightarrow \mathbb{R}$ is convex and continuous in x . A pair $(x^*, \theta^*) \in \mathcal{X} \times \hat{\Theta}_N$ is a saddle point of (4) if and only if the following conditions hold:

$$x^* \in \arg \min_{x \in \mathcal{X}} \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j} [\varphi(x, \xi)] \text{ and } \theta^* \in \arg \max_{\theta \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j} [\varphi(x^*, \xi)]. \quad (5)$$

Assume that φ is continuously differentiable in x and the expectation $\mathbb{E}_{Q_j} [\nabla_x \varphi(x, \xi)]$ is well-defined for all $j = 1, \dots, n$, then (5) yields the following Bayesian DRVI formulation:

$$0 \in \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j} [\Phi(x^*, \xi)] + \mathcal{N}_{\mathcal{X}}(x^*) \text{ and } \theta^* \in \arg \max_{\theta' \in \hat{\Theta}_N} \sum_{j=1}^n \theta'_j \mathbb{E}_{Q_j} [\varphi(x^*, \xi)],$$

with $\Phi(x, \xi) := \nabla_x \varphi(x, \xi)$. Now consider the multi-portfolio case with I accounts as in [33]. Let $x^i \in \mathcal{X}_i \subseteq \mathbb{R}^{d_i}$ denote the portfolio for account i , and let x^{-i} represent the other $I - 1$ accounts' allocations. Let $\varphi_i : \mathcal{X}_1 \times \dots \times \mathcal{X}_I \times \Xi \rightarrow \mathbb{R}$ be the loss function for account i . Inspired by [37], a distributionally robust multi-portfolio Nash equilibrium aims to find $(x^{1*}, \dots, x^{I*})$ together with worst-case parameters $\theta^{i*} \in \hat{\Theta}_N$ for each account i such that

$$\begin{aligned} x^{i*} &\in \arg \min_{x^i \in \mathcal{X}_i} \sum_{j=1}^n \theta_j^{i*} \mathbb{E}_{Q_j} [\varphi_i(x^i, x^{-i*}, \xi)], \quad i = 1, \dots, I, \\ \theta^{i*} &\in \arg \max_{\theta^i \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j^i \mathbb{E}_{Q_j} [\varphi_i(x^{i*}, x^{-i*}, \xi)], \quad i = 1, \dots, I. \end{aligned} \quad (6)$$

By the same argument as for (5), any equilibrium $(x^{1*}, \dots, x^{I*}; \theta^{1*}, \dots, \theta^{I*})$ satisfying (6) also satisfies the following Bayesian DRVI problem

$$\begin{aligned} 0 &\in \sum_{j=1}^n \theta_j^{i*} \mathbb{E}_{Q_j} [\nabla_{x^i} \varphi_i(x^{i*}, x^{-i*}, \xi)] + \mathcal{N}_{\mathcal{X}_i}(x^{i*}), \quad i = 1, \dots, I, \\ \theta^{i*} &\in \arg \max_{\theta^i \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j^i \mathbb{E}_{Q_j} [\varphi_i(x^{i*}, x^{-i*}, \xi)], \quad i = 1, \dots, I, \end{aligned} \quad (7)$$

where $\Phi_i(x^1, \dots, x^I, \xi) := \nabla_{x^i} \varphi_i(x^1, \dots, x^I, \xi)$.

The example highlights two complementary roles of the DRVI formulation. While the single-agent model (4) can be treated as a standard min–max program, the multi-portfolio model (6) is inherently an equilibrium system characterized by coupled optimality conditions across agents and, in general, does not admit a reformulation as the minimization of a single objective. This makes the DRVI framework a natural representation of robust multi-agent equilibrium systems.

The single-agent model (4) also provides a convenient setting for comparing different posterior-based decision criteria. Define $g(x, \theta) := \mathbb{E}_{P_\theta}[\varphi(x, \xi)] = \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)]$. A Bayesian risk formulation applies a risk functional to the parameter-dependent expected loss under the posterior distribution [66]; a representative choice is posterior expectation [42, 59], for which

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\mathbb{P}_{\mathcal{S}N}}[g(x, \theta)] = \min_{x \in \mathcal{X}} \sum_{j=1}^n \bar{\theta}_{N,j} \mathbb{E}_{Q_j}[\varphi(x, \xi)], \quad \bar{\theta}_N := \mathbb{E}_{\mathbb{P}_{\mathcal{S}N}}[\theta]. \quad (8)$$

Thus, posterior expectation directly averages the parameter-dependent expected losses.

The Bayesian-DRO framework in [58] introduces an additional conditional robustification. For each θ , let $v_\theta(x) := \sup_{P \in \mathfrak{M}^\theta} \mathbb{E}_P[\varphi(x, \xi)]$, where $\mathfrak{M}^\theta$ is a conditional ambiguity set associated with the reference distribution P_θ . The resulting decision problem is

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\mathbb{P}_{\mathcal{S}N}}[v_\theta(x)]. \quad (9)$$

Hence, Bayesian-DRO first evaluates the conditional worst-case value for each θ and then averages these values with respect to the posterior distribution. If the supremum defining $v_\theta(x)$ is attained, a conditional worst-case distribution may be selected as $P_\theta^*(x) \in \arg \max_{P \in \mathfrak{M}^\theta} \mathbb{E}_P[\varphi(x, \xi)]$, in which case $v_\theta(x) = \mathbb{E}_{P_\theta^*(x)}[\varphi(x, \xi)]$. Thus, the posterior expectation in (9) averages the conditional worst-case values generated by the family $\{P_\theta^*(x)\}_{\theta \in \Theta}$.

In contrast, the single-agent optimization problem underlying our proposed Bayesian DRVI is

$$\min_{x \in \mathcal{X}} \sup_{\theta \in \hat{\Theta}_N} g(x, \theta), \quad (10)$$

where the posterior is used to construct the credible parameter set $\hat{\Theta}_N$, over which worst-case selection is performed.

Two singleton cases further clarify the relationships. If $\mathfrak{M}^\theta = \{P_\theta\}$ for every θ , then (9) reduces to (8). For the proposed formulation, in the present finite-mixture setting where $g(x, \theta)$ is linear in θ , if the parameter ambiguity set in (10) is specialized to $\{\bar{\theta}_N\}$, then (10) also coincides with (8). This equivalence relies on the linear dependence of $g(x, \theta)$ on the mixture weights and need

not hold when $g(x, \theta)$ depends nonlinearly on θ . Thus, posterior-expectation optimization serves as a non-robust reference point in the present finite-mixture setting. Outside these special cases, Bayesian-DRO averages the conditional worst-case values $v_\theta(x)$ over the posterior distribution, whereas the proposed formulation minimizes the worst-case expected loss over $\hat{\Theta}_N$. The two formulations therefore generally define different statistical decision criteria.

We next detail the posterior-based construction of $\hat{\Theta}_N$.

2.1. Bayesian parametric ambiguity sets for distributions We work with a Bayesian parametric ambiguity set defined over the mixture weights of a prescribed finite-mixture family. As a result, the ambiguity set is constructed in a parameter set whose dimension remains fixed as sample size grows, in contrast to data-driven ambiguity sets supported on the observed samples [30, 32], whose effective dimension typically scales with the number of observations. This representation is computationally attractive, but it comes with a stronger modeling commitment: the component distributions $Q_1, \dots, Q_n$ must be prescribed from structural knowledge or estimated from data. Hence, the proposed framework is most natural in settings where such mixture components have meaningful structural interpretations or can be reliably estimated. In practice, the components may reflect regimes, latent segments, or distinct demand-generation mechanisms [7, 36, 63], or they may be learned using standard mixture-estimation procedures such as the expectation–maximization (EM) algorithm [48, 61]. To construct practical ambiguity sets that contain the true parametric set Θ^c with high probability, we first introduce two fundamental concepts of confidence regions with coverage guarantees in the DRO context.

DEFINITION 1 (FREQUENTIST GUARANTEE). A set $\mathcal{P}_N \subseteq \mathcal{P}$ is called a level- α frequentist confidence region if $\mathbb{P}_{\mathcal{S}^N|P^c} \{P^c \in \mathcal{P}_N\} \geq 1 - \alpha$, where $\mathbb{P}_{\mathcal{S}^N|P^c} := (P^c)^N$ denotes the probability measure induced by N i.i.d. observations from P^c .

Definition 1 states that a frequentist confidence region $\mathcal{P}_N$ contains the true distribution P^c with probability at least $1 - \alpha$ under repeated sampling. Classical ambiguity sets satisfying such guarantees include constructions based on moments [14], ϕ -divergences [2, 12, 57], hypothesis tests [3], and Wasserstein/ ℓ_1 / ℓ_∞ metrics [28, 30]. In our setting, distributional uncertainty is modeled within a prescribed finite-mixture family through a Bayesian parametric ambiguity set over the mixture weights. This yields a fixed-dimensional parametric representation, while the uncertainty quantification is expressed in posterior terms. We therefore introduce below a Bayesian analogue of Definition 1 in the parametric setting.

DEFINITION 2 (POSTERIOR GUARANTEE). A set $\mathcal{P}_{\Theta_N} := \{P_\theta \mid \theta \in \Theta_N\} \subseteq \mathcal{P}_\Theta$ is called a level- α Bayesian confidence region if $\mathbb{P}_{\mathcal{S}^N} \{P_{\theta^c} \in \mathcal{P}_{\Theta_N}\} \geq 1 - \alpha$.

In contrast to Definition 1, the probability here is taken with respect to the posterior distribution $\mathbb{P}_{\mathcal{S}^N}$ induced by the observed data. The Bayesian confidence region $\mathcal{P}_{\Theta_N}$ in Definition 2 corresponds one-to-one to a parameter set $\Theta_N \subseteq \Theta$, which is commonly referred to as a level- α Bayesian confidence set in [35], satisfying $\mathbb{P}_{\mathcal{S}^N} \{\theta^c \in \Theta_N\} \geq 1 - \alpha$. For notational simplicity, we henceforth work directly with the parameter set Θ_N . Moreover, any Θ_N that satisfies the posterior guarantee at level α automatically satisfies the Bayesian feasible guarantee in [22, Definition 2] at the same confidence level. For a detailed discussion on the trade-offs between frequentist and Bayesian methods, we refer the reader to [15, 22].

As multiple sets may satisfy Definition 2, we aim to construct a computationally tractable minimal Bayesian confidence region Θ_N , thereby avoiding excessive conservatism [8]. To obtain an explicit construction, we invoke a normal approximation of the posterior under standard Bernstein–von Mises regularity conditions [22].

ASSUMPTION 1. Assume (i) the true parameter θ^c lies in the relative interior of Θ , (ii) the log-likelihood $\log f(\mathcal{S}^N \mid \theta)$ is locally twice continuously differentiable around θ^c , (iii) the Fisher information matrix $\mathcal{I}(\theta^c)$ is nonsingular, and (iv) the prior pdf $p(\theta)$ is continuous and strictly positive in a neighborhood of θ^c .

Under Assumption 1, the posterior $\mathbb{P}_{\mathcal{S}^N}$ can be approximated by a normal distribution $\mathcal{N}(\hat{\theta}_N, \Sigma_N)$ by the Bernstein–von Mises theorem [64], with $\Sigma_N = \frac{1}{N} \mathcal{I}(\hat{\theta}_N)^{-1}$ and $\mathcal{I}(\hat{\theta}_N) = -\frac{1}{N} \nabla_{\theta}^2 \log f(\mathcal{S}^N \mid \theta) \Big|_{\theta=\hat{\theta}_N}$. This approximation is widely employed in the literature; see, e.g., [11, 22, 68]. Motivated by this and applying a Bonferroni correction for simultaneous marginal coverage, we define the confidence radius $\hat{\delta}_N = \max_{1 \leq j \leq n} \sqrt{\Sigma_{N,jj}} z_{1-\frac{\alpha}{2n}}$, where $\Sigma_{N,jj}$ is the j -th diagonal element of Σ_N and $z_{1-\frac{\alpha}{2n}}$ is the $(1 - \frac{\alpha}{2n})$ -quantile of the standard normal distribution. This choice leads to the explicit construction

$$\Theta_N := \mathbb{B}(\hat{\theta}_N, \hat{\delta}_N) \cap \Theta = \{\theta \mid \|\theta - \hat{\theta}_N\|_\infty \leq \hat{\delta}_N, \theta \in \Theta\} \quad (11)$$

as a computationally convenient approximation of the minimal level- α Bayesian confidence region.

To quantify the convergence of the Bayesian confidence region Θ_N , the data-driven ambiguity set $\hat{\Theta}_N$, and later the corresponding solution sets of $(\mathbf{B}_N)$ and $(\mathbf{B}_N^\lambda)$, we employ the directed set

deviation $\mathbb{D}$ induced by the ℓ_∞ norm. Specifically, for a metric space $(\mathbb{X}, \|\cdot\|_\infty)$ and any two subsets $S_1, S_2 \subseteq \mathbb{X}$, the discrepancy measure is defined as:

$$\mathbb{D}(S_1, S_2) = \sup_{s_1 \in S_1} \mathbf{d}(s_1, S_2) = \sup_{s_1 \in S_1} \inf_{s_2 \in S_2} \|s_1 - s_2\|_\infty.$$

Further, the Hausdorff distance is defined as $\mathbb{H}(S_1, S_2) = \max\{\mathbb{D}(S_1, S_2), \mathbb{D}(S_2, S_1)\}$. We next impose mild conditions to establish almost-sure consistency of the Bayesian confidence region constructed above, which will be used later for convergence analysis. Following the setting in [66], the observed data $\mathcal{S}^N = \{\xi^1, \dots, \xi^N\}$ are generated i.i.d. from the true distribution P_{θ^c} , and the infinite data stream $(\xi^1, \xi^2, \dots)$ induces the product probability measure $(P_{\theta^c})^\mathbb{N}$ on the sample-path space $\Xi^\mathbb{N}$. Unless otherwise stated, “almost surely” (a.s.) convergence in this paper is with respect to $(P_{\theta^c})^\mathbb{N}$ by default.

ASSUMPTION 2. *For all $\theta \in \Theta$, (i) there exists an integrable function g such that $|\log f(\xi|\theta)| \leq g(\xi)$ with $\mathbb{E}_{P_{\theta^c}}[g(\xi)] < \infty$, and (ii) there exist constants $0 < c_1 < c_2$ such that the prior pdf $p(\theta)$ satisfies $c_1 \leq p(\theta) \leq c_2$.*

Under the above conditions, we can establish the convergence of the Bayesian confidence region Θ_N defined in (11) as $N \rightarrow \infty$.

LEMMA 1. *Under Assumptions 1-2, $\mathbb{H}(\Theta_N, \{\theta^c\})$ converges to 0 almost surely as $N \rightarrow \infty$.*

This follows from standard results (e.g., [58, Lemma 3.1]) and is referred to as Bayesian consistency in [59, Assumption 3.2]. In particular, $\|\hat{\theta}_N - \theta^c\|_\infty \rightarrow 0$ and $\hat{\delta}_N \rightarrow 0$ almost surely as $N \rightarrow \infty$. Based on the Bayesian confidence region Θ_N in (11) and Lemma 1, we now construct the Bayesian parametric ambiguity set $\hat{\Theta}_N$ with the minimal (i.e., smallest) radius that still contains Θ^c with posterior probability at least $1 - \alpha$.

THEOREM 1. *Suppose Assumption 1 holds. Let $\hat{r}_N := r^c + \hat{\delta}_N$ and $\hat{\Theta}_N := \mathbb{B}(\hat{\theta}_N, \hat{r}_N) \cap \Theta$. Then*

$$\mathbb{P}_{\mathcal{S}^N} \{\Theta^c \subseteq \hat{\Theta}_N\} \geq 1 - \alpha \text{ and } \hat{r}_N = \min \{r | r \geq 0, \mathbb{P}_{\mathcal{S}^N} \{\Theta^c \subseteq \mathbb{B}(\hat{\theta}_N, r) \cap \Theta\} \geq 1 - \alpha\}.$$

Furthermore, under Assumptions 1-2, $\mathbb{H}(\hat{\Theta}_N, \Theta^c)$ converges to 0 almost surely as $N \rightarrow \infty$.

Proof. Recall that the ambiguity set is defined using the ℓ_∞ norm. Suppose that $\theta^c \in \Theta_N$, then for any $\theta \in \Theta^c$, we have

$$\|\theta - \hat{\theta}_N\|_\infty \leq \|\theta - \theta^c\|_\infty + \|\theta^c - \hat{\theta}_N\|_\infty \leq r^c + \hat{\delta}_N.$$

Thus, whenever $\theta^c \in \Theta_N$, the true ambiguity set Θ^c is contained in $\hat{\Theta}_N$. In terms of probability, this implies $\mathbb{P}_{S^N} \{\Theta^c \subseteq \hat{\Theta}_N\} \geq \mathbb{P}_{S^N} \{\theta^c \in \Theta_N\} \geq 1 - \alpha$. To verify the minimality of $\hat{r}_N$, assume that there exists some $\delta'_N < \hat{\delta}_N$ such that

$$\mathbb{P}_{S^N} \{\Theta^c \subseteq \mathbb{B}(\hat{\theta}_N, \delta'_N + r^c) \cap \Theta\} \geq 1 - \alpha.$$

For $\Theta^c \subseteq \mathbb{B}(\hat{\theta}_N, \delta'_N + r^c) \cap \Theta$ to hold, we must have $\theta^c \in \mathbb{B}(\hat{\theta}_N, \delta'_N) \cap \Theta$. This implies

$$\mathbb{P}_{S^N} \{\theta^c \in \mathbb{B}(\hat{\theta}_N, \delta'_N) \cap \Theta\} \geq \mathbb{P}_{S^N} \{\Theta^c \subseteq \mathbb{B}(\hat{\theta}_N, \delta'_N + r^c) \cap \Theta\} \geq 1 - \alpha,$$

which contradicts the minimality of $\hat{\delta}_N$. Finally, under Assumption 2, the convergence of $\hat{\Theta}_N$ to Θ^c follows directly from Lemma 1. ■

Theorem 1 shows that, under the parametric mixture model and the Bernstein–von Mises theorem [64] with Bonferroni correction, the infinite-dimensional distributional ambiguity set over distributions can be represented by a finite-dimensional ambiguity set $\hat{\Theta}_N = \mathbb{B}(\hat{\theta}_N, \hat{r}_N) \cap \Theta$ over mixture weights with a posterior coverage guarantee.

2.2. Model component estimation and comparison with classical ambiguity sets

The construction in Section 2.1 assumes that the component distributions $Q_1, \dots, Q_n$ are fixed. In practice, however, these components may be estimated from data rather than specified a priori. To study this practically relevant setting, we consider a data-driven experiment in which Gaussian components are learned by EM while the number of mixture components is varied explicitly. The goal is to understand how the proposed Bayesian ambiguity set behaves when the component distributions are estimated from data and the number of mixture components varies.

We consider a continuous true distribution P^c given by a three-component mixture of Student- t distributions with unequal locations, scales, and degrees of freedom, namely

$$P^c = 0.55t_4(-2, 0.8^2) + 0.30t_6(1.2, 0.5^2) + 0.15t_3(4, 0.7^2),$$

where $t_\nu(\mu, \sigma^2)$ denotes the Student- t distribution with ν degrees of freedom, location μ , and scale parameter σ . This choice yields a multimodal, asymmetric, and heavy-tailed distribution, thereby providing a misspecified test case for Gaussian-mixture modeling. For each prescribed number of components n , we fit Gaussian components $Q_1, \dots, Q_n$ to the sample S^N by EM and then construct the Bayesian ambiguity set over the mixture weights as in Section 2.1.

We evaluate the resulting construction from two perspectives. First, we measure the nominal approximation quality of the fitted Gaussian-mixture model through the integrated cumulative distribution function (CDF) error

$$\int_{\mathbb{R}} |\hat{F}_n(\xi) - F^c(\xi)| d\xi,$$

where $\hat{F}_n$ denotes the CDF induced by the center of the fitted n -component Bayesian ambiguity set, and F^c is the CDF of the true data-generating distribution P^c . Second, we quantify conservatism through the average width of the induced ambiguity band, computed from the pointwise lower and upper CDF envelopes associated with the ambiguity set. Table 1 reports the Monte Carlo means over 100 replications.

TABLE 1. Integrated CDF error and average ambiguity-band width under different sample sizes N and numbers of Gaussian mixture components n .

N	Integrated CDF error				Average ambiguity-band width			
	$n = 2$	$n = 5$	$n = 8$	$n = 10$	$n = 2$	$n = 5$	$n = 8$	$n = 10$
50	0.4214	0.4162	0.4144	0.4087	0.0757	0.2298	0.3231	0.3582
100	0.3252	0.2875	0.2873	0.2831	0.0540	0.1666	0.2425	0.2763
200	0.2993	0.2397	0.2393	0.2390	0.0353	0.1247	0.1882	0.2248
500	0.2194	0.1292	0.1290	0.1285	0.0219	0.0793	0.1271	0.1558

Table 1 shows that, for each fixed sample size N , increasing the number n of Gaussian mixture components reduces the approximation error and enlarges the ambiguity-band width. At the same time, for each fixed n , increasing the sample size N reduces both the approximation error and the ambiguity-band width. These results suggest that moderate mixture orders provide a favorable balance between approximation quality and ambiguity-band width in this data-driven setting.

To place the above data-driven construction in context, we also report a supplementary comparison with classical ambiguity sets. For this benchmark comparison, we select the number of Gaussian mixture components n according to the Bayesian information criterion (BIC) [55] and estimate the corresponding Gaussian mixture model from the sample by the EM algorithm. As classical benchmarks, given samples $\mathcal{S}^N := \{\xi^1, \dots, \xi^N\}$ from P^c , the ϕ -divergence ambiguity set

(e.g., modified χ^2 divergence) in [2, 32] and the ℓ_1 norm ambiguity set in [19, 28, 65] that satisfy the frequentist guarantee at level α can be constructed as follows:

$$\ell_1 \text{ norm: } \mathcal{P}_N^{\ell_1} = \left\{ P \mid p = (p_1, \dots, p_N) \in \Delta^N, \sum_{i=1}^N |p_i - \hat{p}_{N,i}| \leq \sqrt{\frac{2}{N} \log \left(\frac{2}{\alpha} \right)} \right\},$$

$$\text{Modified } \chi^2 \text{ divergence: } \mathcal{P}_N^{\chi^2} = \left\{ P \mid p = (p_1, \dots, p_N) \in \Delta^N, \sum_{i=1}^N \frac{(p_i - \hat{p}_{N,i})^2}{\hat{p}_{N,i}} \leq \frac{\chi_{N-1, 1-\alpha}^2}{N} \right\}.$$

Here, $\Delta^N := \{p \mid \sum_{i=1}^N p_i = 1, p \in \mathbb{R}_+^N\}$ is the N -dimensional probability simplex and $\hat{p}_N := (1/N, \dots, 1/N)$ denotes the probability vector of the empirical distribution based on $\mathcal{S}^N$. Each distribution P supported on $\mathcal{S}^N$ is uniquely identified with its probability vector $p = (p_1, \dots, p_N) \in \Delta^N$, where $p_i := P(\xi = \xi^i)$ represents the probability mass at sample ξ^i . The term $\chi_{N-1, 1-\alpha}^2$ denotes the $(1 - \alpha)$ -quantile of the χ_{N-1}^2 distribution with $N - 1$ degrees of freedom.

To visualize the geometric structure induced by these constructions, we plot the pointwise lower and upper envelopes of the CDF associated with the fitted Bayesian ambiguity set $\hat{\Theta}_N$ and with the two benchmark sets at level $\alpha = 0.05$ for sample sizes $N \in \{50, 100, 200\}$; see Figure 1.

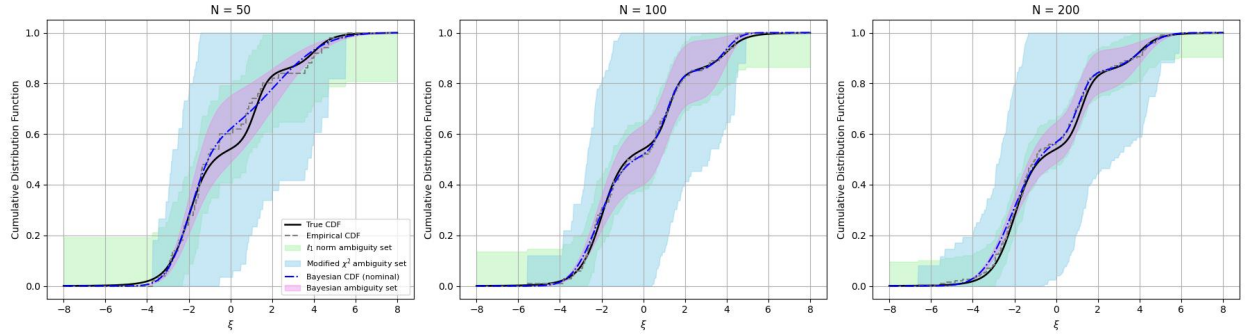


FIGURE 1. CDF envelopes induced by different ambiguity sets at $\alpha = 0.05$.

The CDF envelopes in Figure 1 suggest that under the fitted Gaussian-mixture model, the nominal Bayesian CDF provides a smooth approximation to the true data-generating distribution and the ambiguity band induced by the Bayesian mixture model is typically narrower than those induced by the ℓ_1 and modified χ^2 constructions.

3. Existence and convergence analysis of solutions In this section, we first prove the existence of solutions to (B_N) and (B_N^λ) under mild conditions. Next, we show that as $N \rightarrow \infty$ and $\lambda \downarrow 0$, the solutions of the data-driven problems (B_N) and (B_N^λ) converge almost surely to a solution of problem (B^c) .

3.1. Existence of solutions for $(\mathbf{B}_N)$ and $(\mathbf{B}_N^\lambda)$ We begin by verifying that the Bayesian DRVI problems are well posed. Specifically, for any fixed $\hat{\theta}_N$ and $\hat{r}_N$, we show that the x -solution sets of the Bayesian DRVI problem $(\mathbf{B}_N)$ and its regularized counterpart $(\mathbf{B}_N^\lambda)$ are nonempty and compact. This provides a basic regularity framework for the convergence and stability analysis developed below. To this end, we first introduce the necessary notation. For any fixed $x \in \mathcal{X}$, denote the optimal value and the set of maximizers (in θ) of the lower-level maximization problem in $(\mathbf{B}_N)$, as well as those associated with the regularized lower-level maximization problem in $(\mathbf{B}_N^\lambda)$, by

$$v(x, \hat{\theta}_N, \hat{r}_N) := \max_{\theta \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)], \quad v^\lambda(x, \hat{\theta}_N, \hat{r}_N) := \max_{\theta \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta\|^2,$$

$$\vartheta(x, \hat{\theta}_N, \hat{r}_N) := \arg \max_{\theta \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)], \quad \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) := \arg \max_{\theta \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta\|^2,$$

respectively. Moreover, let $\mathfrak{X}(\theta^c, r^c)$, $\mathfrak{X}(\hat{\theta}_N, \hat{r}_N)$ and $\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)$ denote the x -solution sets of Bayesian DRVI $(\mathbf{B}^c)$, the data-driven Bayesian DRVI $(\mathbf{B}_N)$ and the data-driven regularized Bayesian DRVI $(\mathbf{B}_N^\lambda)$, respectively.

PROPOSITION 1. *Suppose that $\mathbb{E}_{Q_j}[\Phi(\cdot, \xi)]$ and $\mathbb{E}_{Q_j}[\varphi(\cdot, \xi)]$ are continuous for $j \in \{1, \dots, n\}$ and the set $\mathcal{X}$ is bounded. Then $\mathfrak{X}(\hat{\theta}_N, \hat{r}_N)$ and $\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)$ are nonempty and compact.*

Proof. The result for $\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)$ follows immediately when $\lambda > 0$, together with the continuity of $\mathbb{E}_{Q_j}[\Phi(x, \xi)]$ and the compactness of $\mathcal{X}$. We only need to focus on the case $\lambda = 0$. For any fixed $x \in \mathcal{X}$, the maximization problem $\theta \in \arg \max_{\theta' \in \hat{\Theta}_N} \sum_{j=1}^n \theta'_j \mathbb{E}_{Q_j}[\varphi(x, \xi)]$ can be reformulated as $0 \in -\varphi^n(x) + \mathcal{N}_{\hat{\Theta}_N}(\theta)$, where $\varphi^n(x) := (\mathbb{E}_{Q_1}[\varphi(x, \xi)], \dots, \mathbb{E}_{Q_n}[\varphi(x, \xi)])^\top \in \mathbb{R}^n$ and $\mathcal{N}_{\hat{\Theta}_N}(\theta)$ is the normal cone to the compact set $\hat{\Theta}_N$ at θ . Therefore, $(\mathbf{B}_N)$ can be written as the following VI problem:

$$0 \in \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\Phi(x, \xi)] + \mathcal{N}_{\mathcal{X}}(x) \text{ and } 0 \in -\varphi^n(x) + \mathcal{N}_{\hat{\Theta}_N}(\theta).$$

Since $\mathcal{X} \times \hat{\Theta}_N$ is nonempty convex and compact and φ^n is continuous, the desired conclusion follows directly from Corollary 2.2.5 in [17]. ■

3.2. Convergence analysis In this subsection, we prove the convergence of the x -solution sets $\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)$ of problem $(\mathbf{B}_N^\lambda)$ to $\mathfrak{X}(\theta^c, r^c)$ of $(\mathbf{B}^c)$ as $N \rightarrow \infty$ and $\lambda \downarrow 0$. The convergence analysis proceeds in two steps: (i) the convergence of $\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)$ to $\mathfrak{X}(\hat{\theta}_N, \hat{r}_N)$ as $\lambda \downarrow 0$, and (ii) the convergence of $\mathfrak{X}(\hat{\theta}_N, \hat{r}_N)$ to $\mathfrak{X}(\theta^c, r^c)$ as $N \rightarrow \infty$. Due

to the continuity of the functions and the compactness of $\mathcal{X}$, there exists $M > 0$ such that $\sup_{x \in \mathcal{X}, j \in \{1, \dots, n\}} \{|\mathbb{E}_{Q_j}[\varphi(x, \xi)]|, |\mathbb{E}_{Q_j}[\Phi_1(x, \xi)]|, \dots, |\mathbb{E}_{Q_j}[\Phi_d(x, \xi)]|\} \leq M$. For each $j = 1, \dots, n$, we define the component-wise lower and upper bounds for the mixture weights as $\hat{\theta}_{N,j}^l := \max\{0, \hat{\theta}_{N,j} - \hat{r}_N\}$ and $\hat{\theta}_{N,j}^u := \min\{1, \hat{\theta}_{N,j} + \hat{r}_N\}$. The following result provides an alternative representation of the lower-level objective function, highlighting the risk-averse nature of our Bayesian DRVI framework.

THEOREM 2. *Let J be a categorical random variable taking values in $\{1, \dots, n\}$. Let P^l and P^{u-l} be the probability distributions of J with*

$$P^l(J = j) = \frac{\hat{\theta}_{N,j}^l}{\sum_{i=1}^n \hat{\theta}_{N,i}^l} \quad \text{and} \quad P^{u-l}(J = j) = \frac{\hat{\theta}_{N,j}^u - \hat{\theta}_{N,j}^l}{\sum_{i=1}^n (\hat{\theta}_{N,i}^u - \hat{\theta}_{N,i}^l)}.$$

Define $\beta_N := 1 - \frac{1 - \sum_{j=1}^n \hat{\theta}_{N,j}^l}{\sum_{i=1}^n (\hat{\theta}_{N,i}^u - \hat{\theta}_{N,i}^l)}$. Then, for any fixed $x \in \mathcal{X}$, we have

$$v(x, \hat{\theta}_N, \hat{r}_N) = \left(\sum_{j=1}^n \hat{\theta}_{N,j}^l \right) \mathbb{E}_{P^l} \mathbb{E}_{Q_j} [\varphi(x, \xi)] + \left(1 - \sum_{j=1}^n \hat{\theta}_{N,j}^l \right) \text{CVaR}_{P^{u-l}}^{\beta_N} \mathbb{E}_{Q_J} [\varphi(x, \xi)],$$

where $\text{CVaR}_{P^{u-l}}^{\beta_N} \mathbb{E}_{Q_J} [\varphi(x, \xi)]$ represents the conditional value-at-risk (CVaR) of $\mathbb{E}_{Q_J} [\varphi(x, \xi)]$ with respect to P^{u-l} at risk level β_N .

Proof. Without loss of generality, for a given x , we assume that the values $\varphi_j := \mathbb{E}_{Q_j} [\varphi(x, \xi)]$ are sorted in a non-decreasing order: $\varphi_1 \leq \dots \leq \varphi_n$. Recall that $v(x, \hat{\theta}_N, \hat{r}_N)$ equals the optimal value of the following optimization problem:

$$\begin{aligned} & \sup_{\theta \in \mathbb{R}^n} \sum_{j=1}^n \theta_j \varphi_j \\ & \text{s.t. } \hat{\theta}_{N,j}^l \leq \theta_j \leq \hat{\theta}_{N,j}^u, \quad j = 1, \dots, n, \quad e^\top \theta = 1. \end{aligned} \tag{12}$$

Introducing Lagrange multipliers $d_u^j \geq 0$ for the upper bounds, $d_l^j \geq 0$ for the lower bounds, and $d_0 \in \mathbb{R}$ for the equality constraint, the Lagrange dual problem of (12) can be reformulated as

$$\begin{aligned}
& \inf_{d_u, d_l \geq 0, d_0} \sup_{\theta \in \mathbb{R}^n} \sum_{j=1}^n \theta_j \varphi_j + \sum_{j=1}^n d_u^j (\hat{\theta}_{N,j}^u - \theta_j) + \sum_{j=1}^n d_l^j (\theta_j - \hat{\theta}_{N,j}^l) + d_0 \left(1 - \sum_{j=1}^n \theta_j \right) \\
&= \inf_{d_u, d_l \geq 0, d_0} d_0 + \sum_{j=1}^n (d_u^j \hat{\theta}_{N,j}^u - d_l^j \hat{\theta}_{N,j}^l) + \sum_{j=1}^n \sup_{\theta_j \in \mathbb{R}} (\varphi_j + d_l^j - d_u^j - d_0) \theta_j \\
&= \inf_{d_0} d_0 + \sum_{j=1}^n \inf_{\substack{d_u^j, d_l^j \geq 0 \\ \varphi_j + d_l^j - d_u^j - d_0 = 0}} \left\{ d_u^j \hat{\theta}_{N,j}^u - d_l^j \hat{\theta}_{N,j}^l \right\} \\
&= \inf_{d_0} d_0 + \sum_{j=1}^{j^*} (\varphi_j - d_0) \hat{\theta}_{N,j}^l + \sum_{j=j^*+1}^n (\varphi_j - d_0) \hat{\theta}_{N,j}^u \\
&= \inf_{d_0} d_0 + \sum_{j=1}^n \hat{\theta}_{N,j}^l (\varphi_j - d_0) + \sum_{j=j^*+1}^n (\hat{\theta}_{N,j}^u - \hat{\theta}_{N,j}^l) (\varphi_j - d_0) \\
&= \sum_{j=1}^n \hat{\theta}_{N,j}^l \varphi_j + \inf_{d_0} \left\{ \left(1 - \sum_{j=1}^n \hat{\theta}_{N,j}^l \right) d_0 + \sum_{j=1}^n (\hat{\theta}_{N,j}^u - \hat{\theta}_{N,j}^l) [\varphi_j - d_0]^+ \right\} \\
&= \left(\sum_{j=1}^n \hat{\theta}_{N,j}^l \right) \mathbb{E}_{P^l} \mathbb{E}_{Q_J} [\varphi(x, \xi)] \\
&\quad + \left(1 - \sum_{j=1}^n \hat{\theta}_{N,j}^l \right) \inf_{d_0} \left\{ d_0 + \frac{\sum_{i=1}^n (\hat{\theta}_{N,i}^u - \hat{\theta}_{N,i}^l)}{1 - \sum_{j=1}^n \hat{\theta}_{N,j}^l} \mathbb{E}_{P^{u-l}} [\mathbb{E}_{Q_J} [\varphi(x, \xi)] - d_0]^+ \right\} \\
&= \left(\sum_{j=1}^n \hat{\theta}_{N,j}^l \right) \mathbb{E}_{P^l} \mathbb{E}_{Q_J} [\varphi(x, \xi)] + \left(1 - \sum_{j=1}^n \hat{\theta}_{N,j}^l \right) \text{CVaR}_{P^{u-l}}^{\beta_N} \mathbb{E}_{Q_J} [\varphi(x, \xi)],
\end{aligned}$$

where $[a]^+ := \max\{a, 0\}$, $j^* := \max\{j \in \{1, \dots, n\} : \varphi_j < d_0\}$ if $\varphi_1 < d_0$, and $j^* := 0$ otherwise. The last two equations follow from the definitions of P^l and P^{u-l} and the standard Rockafellar-Uryasev CVaR formula in [51, Theorem 10], respectively. ■

REMARK 1. If $\sum_{j=1}^n \hat{\theta}_{N,j}^l = 0$ (resp. $\sum_{j=1}^n (\hat{\theta}_{N,j}^u - \hat{\theta}_{N,j}^l) = 0$), then the corresponding term in the representation vanishes and the distribution P^l (resp. P^{u-l}) can be chosen arbitrarily. In particular, if no truncation is active, i.e., $\hat{\theta}_{N,j} - \hat{r}_N \geq 0$ and $\hat{\theta}_{N,j} + \hat{r}_N \leq 1$ for all j , we have $\hat{\theta}_{N,j}^u - \hat{\theta}_{N,j}^l = 2\hat{r}_N$ for all j , and therefore $P^{u-l}(J = j) = 1/n$ (uniform over $\{1, \dots, n\}$) and $\beta_N = 1/2$.

Theorem 2 reveals that the original lower-level worst-case objective induced by the Bayesian ambiguity set can be expressed as a weighted combination of two kinds of Bayesian risk measures

$(\mathbb{E} - \mathbb{E}$ and $\text{CVaR} - \mathbb{E})$, which are widely used in stochastic optimization and control [59, 66]. This connects our model to the general relation between DRO and risk-averse optimization. However, in [59, 66], the posterior distribution serves as the outer distribution, and one then performs a prescribed posterior risk functional. In our model, by contrast, the posterior distribution is first used to construct the Bayesian ambiguity set, and the resulting risk representation is then induced endogenously by the geometry of that ambiguity set. Accordingly, the outer measures P^l and P^{u-l} , as well as the associated risk level β_N , are not specified a priori, but arise from the ambiguity-set construction. We next show that v is Lipschitz continuous.

LEMMA 2. *Suppose that $\mathbb{E}_{Q_j}[\varphi(\cdot, \xi)]$ is Lipschitz continuous with modulus $L_{\varphi,j} > 0$ for $j \in \{1, 2, \dots, n\}$. Then v is Lipschitz continuous.*

Proof. Let (x, θ, r) and (x', θ', r') be two points in $X \times \Theta \times [0, 1]$. Denote

$$\Delta := |v(x, \theta, r) - v(x', \theta', r')| \leq \Delta_1 + \Delta_2,$$

where $\Delta_1 := |v(x, \theta, r) - v(x', \theta, r)|$, and $\Delta_2 := |v(x', \theta, r) - v(x', \theta', r')|$.

(i) *Bound Δ_1 .* Since $\mathbb{B}(\theta, r) \cap \Theta \subseteq \Theta$ and Θ is the simplex, we have

$$\begin{aligned} \Delta_1 &= \left| \max_{\theta^\dagger \in \mathbb{B}(\theta, r) \cap \Theta} \sum_{j=1}^n \theta_j^\dagger \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \max_{\theta^\dagger \in \mathbb{B}(\theta, r) \cap \Theta} \sum_{j=1}^n \theta_j^\dagger \mathbb{E}_{Q_j}[\varphi(x', \xi)] \right| \\ &\leq \max_{\theta^\dagger \in \mathbb{B}(\theta, r) \cap \Theta} \sum_{j=1}^n \theta_j^\dagger \left| \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \mathbb{E}_{Q_j}[\varphi(x', \xi)] \right| \leq \max_j L_{\varphi,j} \|x - x'\|. \end{aligned}$$

(ii) *Bound Δ_2 .* Define $\Theta' := \mathbb{B}(\theta', r') \cap \Theta$, we first show

$$\mathbb{H}(\mathbb{B}(\theta, r) \cap \Theta, \Theta') \leq \|\theta - \theta'\|_\infty + |r - r'|. \quad (13)$$

By definition, it suffices to prove $\mathbb{D}(\mathbb{B}(\theta, r) \cap \Theta, \Theta') \leq \|\theta - \theta'\|_\infty + |r - r'|$. If $\mathbb{B}(\theta, r) \cap \Theta \subseteq \Theta'$, then $\mathbb{D}(\mathbb{B}(\theta, r) \cap \Theta, \Theta') = 0$ and the bound holds. Otherwise, take $\bar{\theta} \in \mathbb{B}(\theta, r) \cap \Theta \setminus \Theta'$ and define $\zeta := \frac{r'}{\|\bar{\theta} - \theta'\|_\infty}$. Since $\bar{\theta} \notin \Theta'$, we have $r' < \|\bar{\theta} - \theta'\|_\infty$, implying that $\zeta \in (0, 1)$. Define $\tilde{\theta} := \zeta \bar{\theta} + (1 - \zeta)\theta'$. Then,

$$\mathbf{d}(\tilde{\theta}, \theta') = \|\zeta \bar{\theta} + (1 - \zeta)\theta' - \theta'\|_\infty \leq \zeta \|\bar{\theta} - \theta'\|_\infty = r',$$

which means that $\tilde{\theta} \in \Theta'$. Hence

$$\begin{aligned} \mathbf{d}(\bar{\theta}, \Theta') &\leq \|\bar{\theta} - \tilde{\theta}\|_\infty = \|\bar{\theta} - \zeta \bar{\theta} - (1 - \zeta)\theta'\|_\infty \\ &\leq (1 - \zeta) \|\bar{\theta} - \theta'\|_\infty = \|\bar{\theta} - \theta'\|_\infty - r' \\ &\leq \|\bar{\theta} - \theta\|_\infty + \|\theta - \theta'\|_\infty - r' \leq r + \|\theta - \theta'\|_\infty - r' \\ &\leq \|\theta - \theta'\|_\infty + |r - r'|. \end{aligned}$$

By symmetry, the same bound holds for $\mathbb{D}(\Theta', \mathbb{B}(\theta, r) \cap \Theta)$, proving (13).

Now fix x' . Recall that $\sup_{x \in \mathcal{X}, j} |\mathbb{E}_{Q_j}[\varphi(x, \xi)]| \leq M$. Let $\theta^* \in \arg \max_{\theta^\dagger \in \mathbb{B}(\theta, r) \cap \Theta} \sum_{j=1}^n \theta_j^\dagger \mathbb{E}_{Q_j}[\varphi(x', \xi)]$. For any $\varepsilon > 0$, by the definition of $\mathbb{D}(\mathbb{B}(\theta, r) \cap \Theta, \Theta')$ there exists $\bar{\theta}' \in \Theta'$ such that $\|\theta^* - \bar{\theta}'\|_\infty \leq \mathbb{D}(\mathbb{B}(\theta, r) \cap \Theta, \Theta') + \varepsilon$. Then

$$\begin{aligned} & \max_{\theta^\dagger \in \mathbb{B}(\theta, r) \cap \Theta} \sum_{j=1}^n \theta_j^\dagger \mathbb{E}_{Q_j}[\varphi(x', \xi)] - \max_{\theta^\dagger \in \Theta'} \sum_{j=1}^n \theta_j^\dagger \mathbb{E}_{Q_j}[\varphi(x', \xi)] \\ & \leq nM \|\theta^* - \bar{\theta}'\|_\infty \leq nM (\mathbb{D}(\mathbb{B}(\theta, r) \cap \Theta, \Theta') + \varepsilon). \end{aligned}$$

Letting $\varepsilon \downarrow 0$ and using symmetry yields

$$\Delta_2 \leq nM \mathbb{H}(\mathbb{B}(\theta, r) \cap \Theta, \Theta') \leq nM (\|\theta - \theta'\|_\infty + |r - r'|).$$

Combining the bounds on Δ_1 and Δ_2 yields the desired Lipschitz continuity. $\blacksquare$

To quantify the discrepancy between the x -solution sets $\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)$ and $\mathfrak{X}(\theta^c, r^c)$, we first demonstrate that $\mathfrak{X}(\hat{\theta}_N, \hat{r}_N)$ closely approximates $\mathfrak{X}(\theta^c, r^c)$. To this end, we present the following lemma.

LEMMA 3. *Suppose that the conditions in Lemma 1 hold and $\mathbb{E}_{Q_j}[\Phi(\cdot, \xi)]$ is continuous for $j \in \{1, \dots, n\}$. Then $\mathbb{D}(\mathfrak{X}(\hat{\theta}_N, 0), \mathfrak{X}(\theta^c, 0)) \rightarrow 0$ almost surely as $N \rightarrow \infty$.*

Proof. Recall that $\mathfrak{X}(\hat{\theta}_N, 0)$ and $\mathfrak{X}(\theta^c, 0)$ can be regarded as the x -solution sets of the SVI problem. Since $\hat{\theta}_N \rightarrow \theta^c$ according to Lemma 1, the result follows directly from Lemma 2.1 in [67]. $\blacksquare$

THEOREM 3. *Suppose that Assumptions 1-2 hold, $\mathbb{E}_{Q_j}[\Phi(\cdot, \xi)]$ is continuous, and $\mathbb{E}_{Q_j}[\varphi(\cdot, \xi)]$ is Lipschitz continuous with modulus $L_{\varphi, j} > 0$ for $j \in \{1, 2, \dots, n\}$. Then $\mathbb{D}(\mathfrak{X}(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\theta^c, r^c)) \rightarrow 0$ almost surely as $N \rightarrow \infty$.*

Proof. Let $\{x_N\}$ be a sequence with $x_N \in \mathfrak{X}(\hat{\theta}_N, \hat{r}_N)$. From the compactness of $\mathcal{X}$, there exists a subsequence $\{x_{N_k}\}$ of $\{x_N\}$ such that $x_{N_k} \rightarrow x^*$ as $k \rightarrow \infty$. It suffices to demonstrate that $x^* \in \mathfrak{X}(\theta^c, r^c)$. For each $x_{N_k} \in \mathfrak{X}(\hat{\theta}_{N_k}, \hat{r}_{N_k})$, there exists $\theta_{N_k} \in \hat{\Theta}_{N_k}$ satisfying

$$\begin{aligned} 0 & \in \sum_{j=1}^n \theta_{N_k, j} \mathbb{E}_{Q_j}[\Phi(x_{N_k}, \xi)] + \mathcal{N}_{\mathcal{X}}(x_{N_k}), \\ \theta_{N_k} & \in \arg \max_{\theta \in \hat{\Theta}_{N_k}} \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x_{N_k}, \xi)]. \end{aligned}$$

Since Θ is compact, there exists a subsequence (not relabeled) $\{\theta_{N_k}\}$ such that $\theta_{N_k} \rightarrow \theta^*$. Thus

$$\begin{aligned} & \left| \sum_{j=1}^n \theta_{N_k,j} \mathbb{E}_{Q_j} [\varphi(x_{N_k}, \xi)] - \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j} [\varphi(x^*, \xi)] \right| \\ & \leq \left| \sum_{j=1}^n \theta_{N_k,j} \mathbb{E}_{Q_j} [\varphi(x_{N_k}, \xi)] - \sum_{j=1}^n \theta_{N_k,j} \mathbb{E}_{Q_j} [\varphi(x^*, \xi)] \right| + \left| \sum_{j=1}^n \theta_{N_k,j} \mathbb{E}_{Q_j} [\varphi(x^*, \xi)] - \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j} [\varphi(x^*, \xi)] \right| \\ & \leq \sum_{j=1}^n \theta_{N_k,j} L_{\varphi,j} \|x_{N_k} - x^*\| + \sum_{j=1}^n |\theta_j^* - \theta_{N_k,j}| |\mathbb{E}_{Q_j} [\varphi(x^*, \xi)]| \rightarrow 0, \end{aligned}$$

as $k \rightarrow \infty$. Based on the convergence in Theorem 1 and the Lipschitz property of v in Lemma 2, it follows that $v(x_{N_k}, \hat{\theta}_{N_k}, \hat{r}_{N_k}) \rightarrow v(x^*, \theta^c, r^c)$ almost surely as $k \rightarrow \infty$. Since $\sum_{j=1}^n \theta_{N_k,j} \mathbb{E}_{Q_j} [\varphi(x_{N_k}, \xi)] = v(x_{N_k}, \hat{\theta}_{N_k}, \hat{r}_{N_k})$, we have $\sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j} [\varphi(x^*, \xi)] = v(x^*, \theta^c, r^c)$, which leads to $\theta^* \in \vartheta(x^*, \theta^c, r^c)$.

As $k \rightarrow \infty$, we have $\sum_{j=1}^n \theta_{N_k,j} \mathbb{E}_{Q_j} [\Phi(x_{N_k}, \xi)] \rightarrow \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j} [\Phi(x^*, \xi)]$. Since $\mathcal{N}_{\mathcal{X}}(\cdot)$ is upper semicontinuous for the closed and convex $\mathcal{X}$ and $0 \in \sum_{j=1}^n \theta_{N_k,j} \mathbb{E}_{Q_j} [\Phi(x_{N_k}, \xi)] + \mathcal{N}_{\mathcal{X}}(x_{N_k})$, we obtain $0 \in \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j} [\Phi(x^*, \xi)] + \mathcal{N}_{\mathcal{X}}(x^*)$, which completes the proof. ■

REMARK 2. When we choose $r^c = 0$, Theorem 3 yields the convergence result $\mathbb{D}(\mathfrak{X}(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\theta^c, 0)) \rightarrow 0$, i.e., in the risk-neutral case, the Bayesian DRVI solutions converge to a solution of SVI (1) with the true distribution P_{θ^c} . This means that, in practice, our ambiguity-aware decisions become increasingly accurate and less conservative as parametric uncertainty diminishes with the accumulation of information or data.

In the following, for a fixed sample size N with the ambiguity set $\hat{\Theta}_N$, we analyze the convergence of the x -solution set $\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)$ of $(\mathbf{B}_N^\lambda)$ as λ tends to 0. We begin by establishing the uniform convergence of the regularized optimal value function to the corresponding optimal value function of problem $(\mathbf{B}_N)$.

LEMMA 4. *For any fixed $\hat{\theta}_N, \hat{r}_N$, the optimal value $v^\lambda(\cdot, \hat{\theta}_N, \hat{r}_N)$ converges to $v(\cdot, \hat{\theta}_N, \hat{r}_N)$ uniformly as $\lambda \downarrow 0$ over $x \in \mathcal{X}$. Moreover, for any fixed $x \in \mathcal{X}$, the unique maximizer of the regularized lower-level problem converges to the unique least-norm solution of the unregularized problem, i.e.,*

$$\vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) \rightarrow \theta_N^{\min} := \arg \min_{\theta \in \vartheta(x, \hat{\theta}_N, \hat{r}_N)} \|\theta\|.$$

Proof. We begin by proving the uniform convergence of $v^\lambda(\cdot, \hat{\theta}_N, \hat{r}_N)$ to $v(\cdot, \hat{\theta}_N, \hat{r}_N)$. For any fixed $x \in \mathcal{X}$, the functions $\theta \mapsto \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j} [\varphi(x, \xi)]$ and $\theta \mapsto \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j} [\varphi(x, \xi)] - \lambda \|\theta\|^2$ are

continuous on the compact set $\hat{\Theta}_N$. Hence the optimal values $\nu(x, \hat{\theta}_N, \hat{r}_N)$ and $\nu^\lambda(x, \hat{\theta}_N, \hat{r}_N)$ are well defined. For any given $\lambda > 0$, it is evident that $\nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) \leq \nu(x, \hat{\theta}_N, \hat{r}_N)$. Let $\theta^\lambda := \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)$, since $\|\theta\|^2 \leq 1$ for all $\theta \in \hat{\Theta}_N$, we have that

$$\begin{aligned} \nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) &= \sum_{j=1}^n \theta_j^\lambda \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta^\lambda\|^2 \\ &= \max_{\theta \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta\|^2 \geq \nu(x, \hat{\theta}_N, \hat{r}_N) - \lambda, \end{aligned} \quad (14)$$

which means $|\nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) - \nu(x, \hat{\theta}_N, \hat{r}_N)| \leq \lambda$ for all $x \in \mathcal{X}$. Thus, we can obtain that $\nu^\lambda(\cdot, \hat{\theta}_N, \hat{r}_N)$ converges to $\nu(\cdot, \hat{\theta}_N, \hat{r}_N)$ uniformly as $\lambda \downarrow 0$.

We now turn to the second claim. For any fixed $x \in \mathcal{X}$, the set $\vartheta(x, \hat{\theta}_N, \hat{r}_N)$ is nonempty compact and convex. Since $\|\cdot\|^2$ is strictly convex, there exists a unique $\theta_N^{\min}$ such that $\theta_N^{\min} = \arg \min\{\|\theta\| : \theta \in \vartheta(x, \hat{\theta}_N, \hat{r}_N)\}$. Consider any sequence $\{\lambda_k\}$ with $\lambda_k \downarrow 0$. Since $\theta^{\lambda_k} \in \hat{\Theta}_N$ and $\hat{\Theta}_N$ is compact, there exists a subsequence (not relabeled) $\{\theta^{\lambda_k}\}$ such that $\theta^{\lambda_k} \rightarrow \bar{\theta} \in \hat{\Theta}_N$. By the optimality of θ^{λ_k} , for any $\hat{\theta} \in \vartheta(x, \hat{\theta}_N, \hat{r}_N)$,

$$\sum_{j=1}^n \theta_j^{\lambda_k} \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda_k \|\theta^{\lambda_k}\|^2 \geq \sum_{j=1}^n \hat{\theta}_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda_k \|\hat{\theta}\|^2.$$

Taking the limit as $k \rightarrow \infty$, we obtain $\sum_{j=1}^n \bar{\theta}_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] \geq \sum_{j=1}^n \hat{\theta}_j \mathbb{E}_{Q_j}[\varphi(x, \xi)]$ for all $\hat{\theta} \in \vartheta(x, \hat{\theta}_N, \hat{r}_N)$, which implies $\bar{\theta} \in \vartheta(x, \hat{\theta}_N, \hat{r}_N)$. Taking $\hat{\theta} = \theta_N^{\min}$ leads to

$$\lambda_k (\|\theta^{\lambda_k}\|^2 - \|\theta_N^{\min}\|^2) \leq \sum_{j=1}^n \theta_j^{\lambda_k} \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \sum_{j=1}^n \theta_{N,j}^{\min} \mathbb{E}_{Q_j}[\varphi(x, \xi)] \leq 0,$$

so $\|\theta^{\lambda_k}\| \leq \|\theta_N^{\min}\|$ for all k , and therefore $\|\bar{\theta}\| \leq \liminf_{k \rightarrow \infty} \|\theta^{\lambda_k}\| \leq \|\theta_N^{\min}\|$. By the uniqueness of the least-norm element, it follows that $\bar{\theta} = \theta_N^{\min}$. Since every sequence $\{\theta^{\lambda_k}\}$ with $\lambda_k \downarrow 0$ admits a subsequence converging to $\theta_N^{\min}$, we conclude that $\theta^\lambda \rightarrow \theta_N^{\min}$ as $\lambda \downarrow 0$. ■

THEOREM 4. Suppose that $\mathbb{E}_{Q_j}[\Phi(\cdot, \xi)]$ and $\mathbb{E}_{Q_j}[\varphi(\cdot, \xi)]$ are Lipschitz continuous with modulus $L_{\Phi,j} > 0$ and $L_{\varphi,j} > 0$ for $j \in \{1, 2, \dots, n\}$, respectively. Then $\mathbb{D}(\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\hat{\theta}_N, \hat{r}_N)) \rightarrow 0$ as $\lambda \downarrow 0$.

Proof. Consider a non-increasing sequence $\{\lambda_k\}$ that converges to 0 as $k \rightarrow \infty$ with $x_N^k \in \mathfrak{X}^{\lambda_k}(\hat{\theta}_N, \hat{r}_N)$. By compactness of $\mathcal{X}$, there exists a subsequence (not relabeled) $\{x_N^k\}$ with $x_N^k \rightarrow x_N^*$

as $k \rightarrow \infty$. Our goal is to verify that $x_N^* \in \mathfrak{X}(\hat{\theta}_N, \hat{r}_N)$. By the definition of $\mathfrak{X}^{\lambda_k}(\hat{\theta}_N, \hat{r}_N)$, for each x_N^k , there exists $\theta_N^k \in \hat{\Theta}_N$ such that

$$0 \in \sum_{j=1}^n \theta_{N,j}^k \mathbb{E}_{Q_j} [\Phi(x_N^k, \xi)] + \mathcal{N}_X(x_N^k),$$

$$\theta_N^k \in \arg \max_{\theta \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j} [\varphi(x_N^k, \xi)] - \lambda_k \|\theta\|^2.$$

Since $\theta_N^k \in \hat{\Theta}_N$ and $\hat{\Theta}_N$ is compact, there exists a subsequence (not relabeled) $\{\theta_N^k\}$ with $\theta_N^k \rightarrow \theta_N^*$.

Then we have

$$\begin{aligned} & \left| \sum_{j=1}^n \theta_{N,j}^k \mathbb{E}_{Q_j} [\varphi(x_N^k, \xi)] - \lambda_k \|\theta_N^k\|^2 - \sum_{j=1}^n \theta_{N,j}^* \mathbb{E}_{Q_j} [\varphi(x_N^*, \xi)] \right| \\ & \leq \sum_{j=1}^n |\theta_{N,j}^k| |\mathbb{E}_{Q_j} [\varphi(x_N^k, \xi)] - \mathbb{E}_{Q_j} [\varphi(x_N^*, \xi)]| \\ & \quad + \sum_{j=1}^n |\theta_{N,j}^k - \theta_{N,j}^*| |\mathbb{E}_{Q_j} [\varphi(x_N^*, \xi)]| + \lambda_k \|\theta_N^k\|^2. \end{aligned}$$

The right-hand side converges to zero as $k \rightarrow \infty$. Meanwhile, by Lemmas 2 and 4, we have $\nu^{\lambda_k}(x_N^k, \hat{\theta}_N, \hat{r}_N) \rightarrow \nu(x_N^*, \hat{\theta}_N, \hat{r}_N)$. With this and the fact that $\nu^{\lambda_k}(x_N^k, \hat{\theta}_N, \hat{r}_N) = \sum_{j=1}^n \theta_{N,j}^k \mathbb{E}_{Q_j} [\varphi(x_N^k, \xi)] - \lambda_k \|\theta_N^k\|^2$, we can conclude that $\sum_{j=1}^n \theta_{N,j}^* \mathbb{E}_{Q_j} [\varphi(x_N^*, \xi)] = \nu(x_N^*, \hat{\theta}_N, \hat{r}_N)$, which leads to $\theta_N^* \in \vartheta(x_N^*, \hat{\theta}_N, \hat{r}_N)$.

Following a procedure similar to the proof of Theorem 3, along with $0 \in \sum_{j=1}^n \theta_{N,j}^k \mathbb{E}_{Q_j} [\Phi(x_N^k, \xi)] + \mathcal{N}_X(x_N^k)$, we can obtain that $0 \in \sum_{j=1}^n \theta_{N,j}^* \mathbb{E}_{Q_j} [\Phi(x_N^*, \xi)] + \mathcal{N}_X(x_N^*)$, which completes the proof. ■

REMARK 3. Under the conditions in Theorem 4, there exists a sequence $\{\lambda_k\}$ with $\lambda_k \downarrow 0$ such that $x_N^k \in \mathfrak{X}^{\lambda_k}(\hat{\theta}_N, \hat{r}_N)$ with $x_N^k \rightarrow x_N^*$. For $\theta_N^k \in \vartheta^{\lambda_k}(x_N^k, \hat{\theta}_N, \hat{r}_N)$, there exists a subsequence such that $\theta_N^k \rightarrow \theta_N^*$ and (x_N^*, θ_N^*) is a solution of $(\mathbf{B}_N)$. Moreover, from Lemma 4, θ_N^* is the unique least-norm optimal solution of $\vartheta(x_N^*, \hat{\theta}_N, \hat{r}_N)$, i.e., $\theta_N^* \equiv \theta_N^{\min} = \arg \min \{ \|\theta\| : \theta \in \vartheta(x_N^*, \hat{\theta}_N, \hat{r}_N) \}$.

We are now ready to present the main convergence theorem for this section.

THEOREM 5. Suppose that Assumptions 1-2 and the conditions in Theorem 4 hold. Then, almost surely,

$$\lim_{N \rightarrow \infty} \lim_{\lambda \downarrow 0} \mathbb{D}(\mathfrak{X}^{\lambda}(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\theta^c, r^c)) = 0.$$

Proof. Based on the triangle inequality, we have

$$\mathbb{D}(\mathfrak{X}^{\lambda}(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\theta^c, r^c)) \leq \mathbb{D}(\mathfrak{X}^{\lambda}(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\hat{\theta}_N, \hat{r}_N)) + \mathbb{D}(\mathfrak{X}(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\theta^c, r^c)).$$

The first term vanishes as $\lambda \downarrow 0$ by Theorem 4, and the second term vanishes almost surely as $N \rightarrow \infty$ by Theorem 3. Hence the left-hand side converges to zero almost surely as claimed. ■

Theorem 5 provides a sequential asymptotic consistency result, obtained by first letting $\lambda \downarrow 0$ for fixed N and then letting $N \rightarrow \infty$. While this ensures convergence to the Bayesian DRVI solution set, it does not yet determine how λ should be chosen as a function of N . In the next section, we address this issue through quantitative stability bounds, which make it possible to discuss the trade-off between sampling error and regularization bias.

4. Stability of solution sets While Section 3 proves a sequential asymptotic consistency result for solutions to $(\mathbf{B}_N^\lambda)$ as $N \rightarrow \infty$ and $\lambda \downarrow 0$, practical applications rely on finite, often noisy datasets and a fixed positive λ . This setting introduces two distinct sources of deviation from the idealized solution: (i) sampling noise, which perturbs the estimated ambiguity parameters $(\hat{\theta}_N, \hat{r}_N)$, and (ii) the regularization bias due to $\lambda > 0$. These facts highlight the importance of stability analysis in ensuring robustness. In particular, small perturbations in $\mathcal{S}^N$ may generate a contaminated dataset $\tilde{\mathcal{S}}^N = \{\tilde{\xi}^1, \dots, \tilde{\xi}^N\}$ with shifted ambiguity parameters $(\tilde{\theta}_N, \tilde{r}_N)$, leading to the perturbed Bayesian DRVI problem:

$$\begin{aligned} 0 &\in \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j} [\Phi(x, \xi)] + \mathcal{N}_X(x), \\ \theta &\in \arg \max_{\theta' \in \tilde{\Theta}_N} \sum_{j=1}^n \theta'_j \mathbb{E}_{Q_j} [\varphi(x, \xi)] - \lambda \|\theta'\|^2, \end{aligned} \tag{B}_N^\lambda$$

where the perturbed ambiguity set $\tilde{\Theta}_N := \mathbb{B}(\tilde{\theta}_N, \tilde{r}_N) \cap \Theta$ is constructed from the contaminated dataset $\tilde{\mathcal{S}}^N$.

This section investigates how the x -solution set responds to perturbations of ambiguity sets. In addition, we provide finite-sample guarantees that ensure statistical performance with high probability. Further, we analyze whether the perturbed solution set $\mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N)$ closely approximates the solution set $\mathfrak{X}(\theta^c, r^c)$ for sufficiently large N . If such stability holds, the perturbed solutions can be justified as reliable approximations to $(\mathbf{B}^c)$ even under data contamination. We begin by establishing the following lemma, which provides a bound on the Hausdorff distance between ambiguity sets corresponding to different parameters.

LEMMA 5 (Quantitative stability of Bayesian parametric ambiguity sets). *For any $\hat{\theta}_N, \tilde{\theta}_N \in \Theta$ and $\hat{r}_N, \tilde{r}_N \in [0, 1]$, we have*

$$\mathbb{H}(\hat{\Theta}_N, \tilde{\Theta}_N) \leq \|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|.$$

This is a direct consequence by the proof of Lemma 2 from (13), we omit the proof.

LEMMA 6. Suppose that the conditions in Theorem 5 hold. Then, for any $x \in X$ and $\lambda \geq 0$,

$$|\nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) - \nu^\lambda(x, \tilde{\theta}_N, \tilde{r}_N)| \leq (nM + 2\lambda)(\|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|).$$

Proof. Note that

$$\begin{aligned} & \nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) - \nu^\lambda(x, \tilde{\theta}_N, \tilde{r}_N) \\ &= \max_{\theta \in \hat{\Theta}_N} \left[\sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta\|^2 \right] - \max_{\theta' \in \tilde{\Theta}_N} \left[\sum_{j=1}^n \theta'_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta'\|^2 \right] \\ &\leq \max_{\theta \in \hat{\Theta}_N} \min_{\theta' \in \tilde{\Theta}_N} \left(\sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta\|^2 - \sum_{j=1}^n \theta'_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] + \lambda \|\theta'\|^2 \right) \\ &\leq nM \max_{\theta \in \hat{\Theta}_N} \min_{\theta' \in \tilde{\Theta}_N} \|\theta - \theta'\|_\infty + \lambda \max_{\theta \in \hat{\Theta}_N} \min_{\theta' \in \tilde{\Theta}_N} \|\theta\|^2 - \|\theta'\|^2 \\ &\leq nM \max_{\theta \in \hat{\Theta}_N} \min_{\theta' \in \tilde{\Theta}_N} \|\theta - \theta'\|_\infty + 2\lambda \max_{\theta \in \hat{\Theta}_N} \min_{\theta' \in \tilde{\Theta}_N} \|\theta - \theta'\|_\infty = (nM + 2\lambda) \mathbb{D}(\hat{\Theta}_N, \tilde{\Theta}_N). \end{aligned}$$

Similarly, we can show that $\nu^\lambda(x, \tilde{\theta}_N, \tilde{r}_N) - \nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) \leq (nM + 2\lambda) \mathbb{D}(\tilde{\Theta}_N, \hat{\Theta}_N)$. Thus, combining with Lemma 5, we obtain the desired conclusion. $\blacksquare$

The next result quantifies the stability of the lower-level maximizer set $\vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)$ under perturbations of the ambiguity set. Although the lower-level maximization problem admits a unique solution when $\lambda > 0$, we retain the set-valued notation ϑ^λ in order to cover the case $\lambda = 0$ as well. To derive quantitative stability estimates, it is not enough to know that the lower-level optimal value changes continuously under perturbations. One also needs a way to relate small value perturbations to small changes in the corresponding optimizer set. Inspired by the literature in stochastic optimization [52], we consider the growth function ψ_x^λ defined on $\mathbb{R}_+$ by

$$\psi_x^\lambda(\tau) := \min_{\theta \in \hat{\Theta}_N} \left\{ \nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) - \left(\sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta\|^2 \right) : \mathbf{d}(\theta, \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)) \geq \tau \right\},$$

together with the associated function $\Psi_x^\lambda(\eta) := \eta + (\psi_x^\lambda)^{-1}(2\eta)$, where $(\psi_x^\lambda)^{-1}(t) := \sup \{ \tau \in \mathbb{R}_+ : \psi_x^\lambda(\tau) \leq t \}$ and $\eta \in \mathbb{R}_+$. Intuitively, $\psi_x^\lambda(\tau)$ measures the smallest loss in the lower-level objective incurred by choosing a parameter θ that stays at least τ away from the maximizer set $\vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)$. Accordingly, the inverse $(\psi_x^\lambda)^{-1}$ converts lower-level value perturbations into distance bounds for the corresponding worst-case parameter, which is the key mechanism used in the following stability analysis.

THEOREM 6. *Suppose that the conditions in Theorem 5 hold. Then, for any fixed $x \in \mathcal{X}$, there exists a constant $\hat{L} \geq 1$ such that*

$$\begin{aligned} \vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N) &\subseteq \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) + \Psi_x^\lambda \left(\hat{L} \mathbb{D}(\tilde{\Theta}_N, \hat{\Theta}_N) \right) \mathbb{B}_\infty \\ &\subseteq \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) + \Psi_x^\lambda \left(\hat{L} (\|\tilde{\theta}_N - \hat{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|) \right) \mathbb{B}_\infty \end{aligned}$$

holds for any $\tilde{\theta}_N \in \Theta$ and $\tilde{r}_N \in [0, 1]$. Here $\mathbb{B}_\infty$ denotes the unit ball under the ℓ_∞ norm.

Proof. The argument closely follows the structure of [52, Theorem 9]. Let $\theta^* \in \vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N)$ and $\bar{\delta} := \mathbb{D}(\tilde{\Theta}_N, \hat{\Theta}_N)$, there exists $\bar{\theta} \in \hat{\Theta}_N$ such that $\|\theta^* - \bar{\theta}\|_\infty \leq \bar{\delta}$. From Lemma 6 and the definition of ψ_x^λ , we have

$$\begin{aligned} 2(nM + 2\lambda)\bar{\delta} &\geq (nM + 2\lambda)\bar{\delta} + \nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) - \nu^\lambda(x, \tilde{\theta}_N, \tilde{r}_N) \\ &= (nM + 2\lambda)\bar{\delta} + \nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) - \left(\sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j} [\varphi(x, \xi)] - \lambda \|\theta^*\|^2 \right) \\ &\geq \nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) - \left(\sum_{j=1}^n \bar{\theta}_j \mathbb{E}_{Q_j} [\varphi(x, \xi)] - \lambda \|\bar{\theta}\|^2 \right) \geq \psi_x^\lambda \left(\bar{\theta}, \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) \right) \\ &\geq \inf_{\theta \in \theta^* + \bar{\delta} \mathbb{B}_\infty} \psi_x^\lambda \left(\mathbf{d} \left(\theta, \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) \right) \right) = \psi_x^\lambda \left(\mathbf{d} \left(\theta^*, \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) + \bar{\delta} \mathbb{B}_\infty \right) \right), \end{aligned}$$

where the last inequality follows from the monotonicity of ψ_x^λ . This implies

$$\begin{aligned} \mathbf{d} \left(\theta^*, \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) \right) &\leq \bar{\delta} + \mathbf{d} \left(\theta^*, \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) + \bar{\delta} \mathbb{B}_\infty \right) \\ &\leq \bar{\delta} + (\psi_x^\lambda)^{-1} (2(nM + 2\lambda)\bar{\delta}) \leq \hat{L} \bar{\delta} + (\psi_x^\lambda)^{-1} (2\hat{L} \bar{\delta}) = \Psi_x^\lambda (\hat{L} \bar{\delta}) \end{aligned}$$

where $\hat{L} := \max \{1, nM + 2\lambda\}$. Thus we have $\vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N) \subseteq \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) + \Psi_x^\lambda \left(\hat{L} \mathbb{D}(\tilde{\Theta}_N, \hat{\Theta}_N) \right) \mathbb{B}_\infty$. Since Ψ_x^λ is monotonically increasing, we further conclude that

$$\vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) + \Psi_x^\lambda \left(\hat{L} \mathbb{D}(\tilde{\Theta}_N, \hat{\Theta}_N) \right) \mathbb{B}_\infty \subseteq \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) + \Psi_x^\lambda \left(\hat{L} (\|\tilde{\theta}_N - \hat{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|) \right) \mathbb{B}_\infty.$$

Thus, the theorem follows. ■

REMARK 4. Suppose that there exist $p_\lambda \geq 1$ and $\gamma_\lambda > 0$ such that for each sufficiently small $\tau \in \mathbb{R}_+$, we have $\psi_x^\lambda(\tau) \geq \gamma_\lambda \tau^{p_\lambda}$. Then for each θ close to $\vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)$, the following holds:

$$\nu^\lambda(x, \hat{\theta}_N, \hat{r}_N) \geq \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j} [\varphi(x, \xi)] - \lambda \|\theta\|^2 + \gamma_\lambda \mathbf{d} \left(\theta, \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) \right)^{p_\lambda}.$$

Under these conditions, the function $\Psi_x^\lambda(\eta)$ can be bounded as $\Psi_x^\lambda(\eta) \leq \eta + \left(\frac{2\eta}{\gamma_\lambda} \right)^{1/p_\lambda} \leq C\eta^{1/p_\lambda}$ for some constant $C > 0$ and sufficiently small $\eta \in \mathbb{R}_+$.

REMARK 5. When $\lambda > 0$, the regularized objective function $\sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta\|^2$ is strictly concave. Near the unique maximizer $\theta^* \in \hat{\Theta}_N$, a second-order expansion gives:

$$v^\lambda(x, \hat{\theta}_N, \hat{r}_N) - \left(\sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta\|^2 \right) \geq \lambda \|\theta - \theta^*\|^2 \geq c\lambda \|\theta - \theta^*\|_\infty^2$$

for some $c > 0$ by norm equivalence. Thus, the growth function satisfies $\psi_x^\lambda(\tau) \geq c\lambda\tau^2$. For $\lambda = 0$, the objective function $\sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\varphi(x, \xi)]$ becomes linear in θ . Define $\delta(x) := \max_j \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \max_{j \notin I(x)} \mathbb{E}_{Q_j}[\varphi(x, \xi)]$ and $I(x) := \arg \max_j \mathbb{E}_{Q_j}[\varphi(x, \xi)]$. Assuming that there exists $\underline{\delta} := \inf_{x \in \mathcal{X}} \delta(x) > 0$, then the growth function satisfies $\psi_x(\tau) \geq \frac{1}{2}\underline{\delta}\tau$.

When $\lambda = 0$, the lower-level problem is linear over $\hat{\Theta}_N$ and may therefore admit multiple maximizers, so first-order growth generally requires an additional nondegeneracy condition. By contrast, when $\lambda > 0$, the term $-\lambda \|\theta\|^2$ makes the lower-level objective strongly concave in θ , yields a unique maximizer, and automatically guarantees second-order growth. As a direct consequence of Remarks 4 and 5, Proposition 2 below establishes that the mapping $\vartheta^\lambda(x, \cdot, \cdot)$ is Hölder continuous at $(\hat{\theta}_N, \hat{r}_N)$ with exponent $\frac{1}{p_\lambda}$ in the regularized-to-regularized case, while the regularized-to-unregularized deviation is controlled with exponent $\frac{1}{p_0}$.

PROPOSITION 2. Suppose that the conditions in Theorem 5 hold. Assume that for any fixed $\lambda \geq 0$, there exist $p_\lambda \geq 1$ and $\gamma_\lambda > 0$ such that $\psi_x^\lambda(\tau) \geq \gamma_\lambda \tau^{p_\lambda}$ for all $x \in \mathcal{X}$. Then for any $\hat{\theta}_N, \tilde{\theta}_N \in \Theta$ and $\hat{r}_N, \tilde{r}_N \in [0, 1]$, we have

$$\sup_{x \in \mathcal{X}} \mathbb{H} \left(\vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N), \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) \right) \leq \left(\frac{1}{\gamma_\lambda} (nM + 2\lambda) (\|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|) \right)^{1/p_\lambda},$$

$$\sup_{x \in \mathcal{X}} \mathbb{D} \left(\vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N), \vartheta(x, \hat{\theta}_N, \hat{r}_N) \right) \leq \left(\frac{1}{\gamma_0} (nM + 2\lambda) (\|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|) + \frac{\lambda}{\gamma_0} \right)^{1/p_0}.$$

Proof. For the first part, let $\tilde{\theta}^* \in \vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N)$ and $\theta^* \in \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)$ be such that $\|\theta^* - \tilde{\theta}^*\|_\infty = \mathbf{d}(\tilde{\theta}^*, \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N))$. This leads to the following estimation:

$$\begin{aligned} \psi_x^\lambda(\|\theta^* - \tilde{\theta}^*\|_\infty) &= \psi_x^\lambda \left(\mathbf{d}(\tilde{\theta}^*, \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)) \right) \\ &\leq \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\theta^*\|^2 - \left(\sum_{j=1}^n \tilde{\theta}_j^* \mathbb{E}_{Q_j}[\varphi(x, \xi)] - \lambda \|\tilde{\theta}^*\|^2 \right) \\ &= v^\lambda(x, \hat{\theta}_N, \hat{r}_N) - v^\lambda(x, \tilde{\theta}_N, \tilde{r}_N) \leq (nM + 2\lambda) (\|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|). \end{aligned}$$

Thus, we have $\mathbb{D}(\vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N), \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)) \leq (\psi_x^\lambda)^{-1}((nM + 2\lambda)(\|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|))$. A similar result holds for $\mathbb{D}(\vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N), \vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N))$. By the assumed growth condition $\psi_x^\lambda(\tau) \geq \gamma_\lambda \tau^{p_\lambda}$, we have $(\psi_x^\lambda)^{-1}(t) \leq \left(\frac{t}{\gamma_\lambda}\right)^{1/p_\lambda}$. Thus, we obtain the desired bound for $\sup_{x \in \mathcal{X}} \mathbb{H}(\vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N), \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N))$.

For the second part, let $\tilde{\theta}^* \in \vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N)$ and $\theta^* \in \vartheta(x, \hat{\theta}_N, \hat{r}_N)$ be such that $\|\theta^* - \tilde{\theta}^*\|_\infty = \mathbf{d}(\tilde{\theta}^*, \vartheta(x, \hat{\theta}_N, \hat{r}_N))$. This leads to

$$\begin{aligned} \psi_x(\|\theta^* - \tilde{\theta}^*\|_\infty) &= \psi_x(\mathbf{d}(\tilde{\theta}^*, \vartheta(x, \hat{\theta}_N, \hat{r}_N))) \\ &\leq v(x, \hat{\theta}_N, \hat{r}_N) - v^\lambda(x, \hat{\theta}_N, \hat{r}_N) + v^\lambda(x, \hat{\theta}_N, \hat{r}_N) - v^\lambda(x, \tilde{\theta}_N, \tilde{r}_N) \\ &\leq \lambda + (nM + 2\lambda)(\|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|). \end{aligned}$$

Thus we can obtain the conclusion for $\sup_{x \in \mathcal{X}} \mathbb{D}(\vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N), \vartheta(x, \hat{\theta}_N, \hat{r}_N))$. $\blacksquare$

To further quantify the perturbation effects on the x -solution set, we define a growth function for $(\mathbf{B}_N^\lambda)$ inspired by [26]. For fixed $(\hat{\theta}_N, \hat{r}_N)$, define $\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ as

$$\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) := \inf_{x \in \mathcal{X}} \left\{ \mathbf{d} \left(0, \sum_{j=1}^n \theta_j \mathbb{E}_{Q_j}[\Phi(x, \xi)] + \mathcal{N}_X(x) \right) : \theta \in \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N), \mathbf{d}(x, \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)) \geq \tau \right\}.$$

Then its inverse $(\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}$ can be defined similarly as that of ψ_x^λ . To ease notation, we use $\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda := \psi_{\hat{\theta}_N, \hat{r}_N}^\lambda$ when $\lambda = 0$.

LEMMA 7. *For any $\epsilon > 0$, $t \geq 0$ and fixed $(\hat{\theta}_N, \hat{r}_N) \in \Theta \times [0, 1]$, there exists $\delta > 0$ such that $(\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t + \delta) \leq (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t) + \epsilon$.*

Proof. Suppose for contradiction that the conclusion does not hold. Then, there exist an $\epsilon_0 > 0$ and a sequence $\{\delta_k\} \downarrow 0$ such that $(\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t + \delta_k) \geq (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t) + \epsilon_0$ for $k \in \mathbb{N}$. Specifically, we have

$$(\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t + \delta_k) = \sup \left\{ \tau : \psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) \leq t + \delta_k \right\} \geq (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t) + \epsilon_0.$$

Then, we have that for every $k \in \mathbb{N}$, $\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda \left((\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t) + \frac{\epsilon_0}{2} \right) \leq t + \delta_k$. Therefore, as $k \rightarrow \infty$, we obtain $\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda \left((\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t) + \frac{\epsilon_0}{2} \right) \leq t$, which leads to a contradiction. $\blacksquare$

THEOREM 7 (Quantitative stability of x -solution set). *Suppose that the conditions in Proposition 2 hold. Then, for any $\lambda \geq 0$, $\hat{\theta}_N, \tilde{\theta}_N \in \Theta$ and $\hat{r}_N, \tilde{r}_N \in [0, 1]$, we have*

$$\begin{aligned} &\mathbb{D}(\mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N), \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)) \\ &\leq (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1} \left(\left(\frac{(nM)^{p_\lambda}(nM + 2\lambda)}{\gamma_\lambda} (\|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|) \right)^{1/p_\lambda} \right), \end{aligned} \quad (15)$$

$$\begin{aligned} & \mathbb{D} \left(\mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N), \mathfrak{X}(\hat{\theta}_N, \hat{r}_N) \right) \\ & \leq (\psi_{\hat{\theta}_N, \hat{r}_N})^{-1} \left(\left(\frac{(nM)^{p_0}(nM+2\lambda)}{\gamma_0} (\|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|) + \frac{\lambda}{\gamma_0} \right)^{1/p_0} \right). \end{aligned} \quad (16)$$

Proof. For the first part, given any $x \in \mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N)$, there exists $\tilde{\theta}^* \in \vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N)$ such that

$$0 \in \sum_{j=1}^n \tilde{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] + \mathcal{N}_X(x).$$

For any $\hat{\theta}^* \in \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)$, we have

$$\begin{aligned} & \left\| \sum_{j=1}^n \tilde{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] - \sum_{j=1}^n \hat{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] \right\|_\infty \\ & \geq \mathbf{d} \left(0, \sum_{j=1}^n \hat{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] + \mathcal{N}_X(x) \right) - \mathbf{d} \left(0, \sum_{j=1}^n \tilde{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] + \mathcal{N}_X(x) \right) \\ & = \mathbf{d} \left(0, \sum_{j=1}^n \hat{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] + \mathcal{N}_X(x) \right) \geq \psi_{\hat{\theta}_N, \hat{r}_N}^\lambda \left(\mathbf{d} \left(x, \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N) \right) \right), \end{aligned}$$

which implies that

$$\mathbf{d} \left(x, \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N) \right) \leq (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1} \left(\inf_{\hat{\theta}^* \in \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)} \left\| \sum_{j=1}^n \tilde{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] - \sum_{j=1}^n \hat{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] \right\|_\infty \right).$$

Note that

$$\begin{aligned} & \inf_{\hat{\theta}^* \in \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)} \left\| \sum_{j=1}^n \tilde{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] - \sum_{j=1}^n \hat{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] \right\|_\infty \\ & \leq nM \mathbf{d} \left(\tilde{\theta}^*, \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) \right) \\ & \leq nM \mathbb{D} \left(\vartheta^\lambda(x, \tilde{\theta}_N, \tilde{r}_N), \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N) \right) \\ & \leq \left(\frac{(nM)^{p_\lambda}(nM+2\lambda)}{\gamma_\lambda} (\|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|) \right)^{1/p_\lambda}, \end{aligned}$$

we obtain

$$\mathbb{D} \left(\mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N), \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N) \right) \leq (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1} \left(\left(\frac{(nM)^{p_\lambda}(nM+2\lambda)}{\gamma_\lambda} (\|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|) \right)^{1/p_\lambda} \right).$$

For the second part, based on Proposition 2, the proof of (16) follows similarly to that of (15), and we omit the details here. ■

For any fixed $\lambda \geq 0$, applying (15) twice—once with $(\hat{\theta}_N, \hat{r}_N)$ versus $(\tilde{\theta}_N, \tilde{r}_N)$, and once with the roles swapped—immediately yields the Hausdorff bound

$$\mathbb{H}\left(\mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N), \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)\right) \leq \max\left\{(\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}((C_\lambda \Delta_N)^{\frac{1}{p_\lambda}}), (\psi_{\tilde{\theta}_N, \tilde{r}_N}^\lambda)^{-1}((C_\lambda \Delta_N)^{\frac{1}{p_\lambda}})\right\}, \quad (17)$$

where $\Delta_N := \|\hat{\theta}_N - \tilde{\theta}_N\|_\infty + |\tilde{r}_N - \hat{r}_N|$ and $C_\lambda := \frac{(nM)^{p_\lambda}(nM+2\lambda)}{\gamma_\lambda}$. We next derive two immediate consequences of Theorem 7. The first provides a posterior finite-sample guarantee for the regularized data-driven solution set, while the second gives a direct deterministic bound for the perturbed regularized solution set relative to the true Bayesian DRVI solution set.

COROLLARY 1 (Posterior finite-sample guarantee). *Suppose that the conditions in Proposition 2 hold. For any given finite sample set $\mathcal{S}^N$, with $C_0 := \frac{(nM)^{p_0}(nM+2\lambda)}{\gamma_0}$, the following guarantee holds:*

$$\mathbb{P}_{\mathcal{S}^N} \left\{ \mathbb{D}\left(\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\theta^c, r^c)\right) \leq (\psi_{\theta^c, r^c})^{-1} \left(\left(C_0(\hat{r}_N + \hat{\delta}_N) + \frac{\lambda}{\gamma_0} \right)^{1/p_0} \right) \right\} \geq 1 - \alpha.$$

Proof. For any $x \in \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)$, there exists $\hat{\theta}^* \in \vartheta^\lambda(x, \hat{\theta}_N, \hat{r}_N)$ such that $0 \in \sum_{j=1}^n \hat{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] + \mathcal{N}_X(x)$. Then, for any $\theta^* \in \vartheta(x, \theta^c, r^c)$, by the same argument as in the proof of Theorem 7,

$$\mathbf{d}(x, \mathfrak{X}(\theta^c, r^c)) \leq (\psi_{\theta^c, r^c})^{-1} \left(\left\| \sum_{j=1}^n \hat{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] - \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] \right\|_\infty \right).$$

In the event that $\Theta^c \subseteq \hat{\Theta}_N$, we have $\|\theta^c - \hat{\theta}_N\|_\infty \leq \hat{r}_N$ and $|r^c - \hat{r}_N| = \hat{\delta}_N$. Note that

$$\left\| \sum_{j=1}^n \hat{\theta}_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] - \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j}[\Phi(x, \xi)] \right\|_\infty \leq \left(C_0(\hat{r}_N + \hat{\delta}_N) + \frac{\lambda}{\gamma_0} \right)^{1/p_0},$$

we obtain

$$\mathbb{D}\left(\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\theta^c, r^c)\right) \leq (\psi_{\theta^c, r^c})^{-1} \left(\left(C_0(\hat{r}_N + \hat{\delta}_N) + \frac{\lambda}{\gamma_0} \right)^{1/p_0} \right).$$

In terms of probability, based on Theorem 1, we have

$$\mathbb{P}_{\mathcal{S}^N} \left\{ \mathbb{D}\left(\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\theta^c, r^c)\right) \leq (\psi_{\theta^c, r^c})^{-1} \left(\left(C_0(\hat{r}_N + \hat{\delta}_N) + \frac{\lambda}{\gamma_0} \right)^{1/p_0} \right) \right\} \geq \mathbb{P}_{\mathcal{S}^N} \{ \Theta^c \subseteq \hat{\Theta}_N \} \geq 1 - \alpha,$$

which completes the proof. ■

Corollary 1 provides a posterior finite-sample guarantee based on the Bayesian ambiguity-set construction. In addition, by replacing $(\hat{\theta}_N, \hat{r}_N)$ with (θ^c, r^c) in (16), we obtain the following direct deterministic bound to the true solution set.

COROLLARY 2. *Suppose that the conditions in Proposition 2 hold. Then, for any contaminated sample set $\tilde{\mathcal{S}}^N$ with induced ambiguity parameters $(\tilde{\theta}_N, \tilde{r}_N)$,*

$$\mathbb{D}\left(\mathfrak{X}^{\lambda}(\tilde{\theta}_N, \tilde{r}_N), \mathfrak{X}(\theta^c, r^c)\right) \leq (\psi_{\theta^c, r^c})^{-1} \left(\left(C_0(\|\tilde{\theta}_N - \theta^c\|_{\infty} + |\tilde{r}_N - r^c|) + \frac{\lambda}{\gamma_0} \right)^{1/p_0} \right).$$

The following remark complements the sequential $\lim_{N \rightarrow \infty} \lim_{\lambda \downarrow 0}$ in Theorem 5 by providing a data-dependent rule for coupling λ_N with the sample size N .

REMARK 6. By applying (16) with (θ^c, r^c) and $(\hat{\theta}_N, \hat{r}_N)$, we obtain

$$\begin{aligned} & \mathbb{D}\left(\mathfrak{X}^{\lambda_N}(\hat{\theta}_N, \hat{r}_N), \mathfrak{X}(\theta^c, r^c)\right) \\ & \leq (\psi_{\theta^c, r^c})^{-1} \left(\left(\frac{(nM)^{p_0}(nM + 2\lambda_N)}{\gamma_0} (\|\hat{\theta}_N - \theta^c\|_{\infty} + \hat{\delta}_N) + \frac{\lambda_N}{\gamma_0} \right)^{1/p_0} \right). \end{aligned}$$

This bound separates the total approximation error into two components: the estimation error $\|\hat{\theta}_N - \theta^c\|_{\infty} + \hat{\delta}_N$, which is induced by finite-sample uncertainty in the ambiguity-set construction, and the regularization bias λ_N . Hence the bound yields a bias-variance type trade-off between statistical estimation error in the ambiguity-set construction and regularization bias in the lower-level problem, in the spirit of standard regularization analyses; see, e.g., [1, Section 7.6] for an analogous discussion. In particular, under the construction in Section 2 with $\|\hat{\theta}_N - \theta^c\|_{\infty} \rightarrow 0$ and $\hat{\delta}_N \rightarrow 0$, any choice satisfying $\lambda_N \rightarrow 0$ yields consistency of the regularized data-driven solution set. Moreover, under the Bernstein–von Mises approximation, one typically has $\|\hat{\theta}_N - \theta^c\|_{\infty} = O(N^{-1/2})$ and $\hat{\delta}_N = O(N^{-1/2})$, then choosing $\lambda_N \asymp N^{-1/2}$ balances the sampling error and the regularization bias at the same order.

In practice, data often deviate from the assumed true distribution P_{θ^c} due to contamination or outliers, especially in fields such as finance, biology and medicine [34]. Typically, one observes samples $\tilde{\mathcal{S}}^N = \{\tilde{\xi}^1, \dots, \tilde{\xi}^N\}$ generated from a perturbed distribution $P_{\tilde{\theta}^c}$ rather than P_{θ^c} . Following [47], we assume that $\tilde{\xi}^1, \dots, \tilde{\xi}^N$ are i.i.d. from $P_{\tilde{\theta}^c}$. In what follows, we further examine how the discrepancy between $P_{\tilde{\theta}^c}$ and P_{θ^c} affects the x -solution set of the Bayesian DRVI problem. In the following, whenever the perturbed sample sequence $\tilde{\mathcal{S}}^N$ is considered, the corresponding “almost surely” statements are understood with respect to $(P_{\tilde{\theta}^c})^{\mathbb{N}}$. Statements involving both the original and perturbed sample sequences are understood almost surely with respect to the joint probability distribution generated by these two sequences.

LEMMA 8. *Suppose that Assumptions 1-2 and the conditions in Lemma 6 hold. Moreover, for any fixed $\lambda \geq 0$, there exist a neighborhood U of (θ^c, r^c) in $\Theta \times [0, 1]$, constants $\bar{p}_{\lambda} \geq 1$, $\bar{\gamma}_{\lambda} > 0$, and*

$\tau_0 > 0$ such that for all $(\theta, r) \in U$ and $\tau \in [0, \tau_0]$, $\psi_{\theta, r}^\lambda(\tau) \geq \bar{\gamma}_\lambda \tau^{\bar{p}_\lambda}$. Then, for any $\tau > 0$ and $t > 0$, we have (i) $\liminf_{N \rightarrow \infty} \psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) \geq \psi_{\theta^c, r^c}^\lambda(\tau)$, and (ii) $\limsup_{N \rightarrow \infty} (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t) \leq (\psi_{\theta^c, r^c}^\lambda)^{-1}(t)$.

Proof. We prove part (i) by contradiction. Suppose that there exists $\delta > 0$ such that

$$\liminf_{N \rightarrow \infty} \psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) \leq \psi_{\theta^c, r^c}^\lambda(\tau) - \delta.$$

Thus, there exists a subsequence $\{N_k\}_{k \geq 1}$ such that $\psi_{\hat{\theta}_{N_k}, \hat{r}_{N_k}}^\lambda(\tau) \leq \psi_{\theta^c, r^c}^\lambda(\tau) - \delta$. By the definition of $\psi_{\hat{\theta}_{N_k}, \hat{r}_{N_k}}^\lambda(\tau)$, there exists a pair $(x_{N_k}, \theta_{N_k}^*)$ such that $\mathbf{d}(x_{N_k}, \mathfrak{X}^\lambda(\hat{\theta}_{N_k}, \hat{r}_{N_k})) \geq \tau$ and

$$\mathbf{d}\left(0, \sum_{j=1}^n \theta_{N_k, j}^* \mathbb{E}_{Q_j}[\Phi(x_{N_k}, \xi)] + \mathcal{N}_X(x_{N_k})\right) \leq \psi_{\theta^c, r^c}^\lambda(\tau) - \delta. \quad (18)$$

Letting $x_{N_k} \rightarrow x^*$, we claim $\mathbf{d}(x^*, \mathfrak{X}^\lambda(\theta^c, r^c)) \geq \tau$. Otherwise, there exists $x^\dagger \in \mathfrak{X}^\lambda(\theta^c, r^c)$ such that $\|x^* - x^\dagger\| < \tau$. From (17), we have

$$\mathbf{d}(x^\dagger, \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)) \leq (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}\left(C_\lambda^{1/p_\lambda}(\|\hat{\theta}_N - \theta^c\|_\infty + |\hat{r}_N - r^c|)^{1/p_\lambda}\right) := \epsilon_N.$$

From Theorem 1 and the assumed growth condition, for N large enough we have $(\hat{\theta}_N, \hat{r}_N) \in U$ and $\epsilon_N \rightarrow 0$ as $N \rightarrow \infty$. For each N_k large enough, there exists $y_{N_k} \in \mathfrak{X}^\lambda(\hat{\theta}_{N_k}, \hat{r}_{N_k})$ with $\|y_{N_k} - x^\dagger\|_\infty < \epsilon_{N_k}$. Thus, we have

$$\|x_{N_k} - y_{N_k}\|_\infty \leq \|x_{N_k} - x^*\|_\infty + \|x^* - x^\dagger\|_\infty + \|x^\dagger - y_{N_k}\|_\infty < (\epsilon' + \epsilon_{N_k}) + \|x^* - x^\dagger\|_\infty$$

where $\epsilon' \rightarrow 0$ as $N_k \rightarrow \infty$. Since $\|x^* - x^\dagger\|_\infty < \tau$ and ϵ_{N_k} can be made arbitrarily small, $\|x_{N_k} - y_{N_k}\|_\infty$ can be made less than τ for large enough N_k . This contradicts the assumption that $\mathbf{d}(x_{N_k}, \mathfrak{X}^\lambda(\hat{\theta}_{N_k}, \hat{r}_{N_k})) \geq \tau$. Therefore, our assumption must be false, implying $\mathbf{d}(x^*, \mathfrak{X}^\lambda(\theta^c, r^c)) \geq \tau$.

Using an argument similar to that in Lemma 3, it is known from (18) that

$$\mathbf{d}\left(0, \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j}[\Phi(x^*, \xi)] + \mathcal{N}_X(x^*)\right) \leq \psi_{\theta^c, r^c}^\lambda(\tau) - \frac{\delta}{2}.$$

However, by the definition of $\psi_{\theta^c, r^c}^\lambda(\tau)$, this is impossible because $\psi_{\theta^c, r^c}^\lambda(\tau)$ is the infimum of such distances for points at least τ away from $\mathfrak{X}^\lambda(\theta^c, r^c)$. Thus,

$$\psi_{\theta^c, r^c}^\lambda(\tau) \leq \mathbf{d}\left(0, \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j}[\Phi(x^*, \xi)] + \mathcal{N}_X(x^*)\right) \leq \psi_{\theta^c, r^c}^\lambda(\tau) - \frac{\delta}{2},$$

which is a contradiction. Therefore, the inequality $\liminf_{N \rightarrow \infty} \psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) \geq \psi_{\theta^c, r^c}^\lambda(\tau)$ must hold.

To establish part (ii), note that for any $\epsilon > 0$, part (i) implies the existence of N_0 such that for all $N \geq N_0$, $\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) \geq \psi_{\theta^c, r^c}^\lambda(\tau) - \epsilon$. By the definition of $(\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}$ and $(\psi_{\theta^c, r^c}^\lambda)^{-1}$, we have that $(\psi_{\theta^c, r^c}^\lambda)^{-1}(t + \epsilon) \geq (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t)$. Since ϵ is arbitrary, we conclude that $\limsup_{N \rightarrow \infty} (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t) \leq (\psi_{\theta^c, r^c}^\lambda)^{-1}(t)$, which completes the proof. ■

Recall that Theorem 7 establishes quantitative stability for the Bayesian DRVI problem. In conjunction with Lemma 7, it offers explicit upper bounds for the qualitative results derived in Theorem 5. To maintain self-containment, we will now demonstrate that, for any $\lambda \geq 0$, the right-hand side of (15) tends to 0 as $N \rightarrow \infty$ when the perturbation $\|\theta^c - \tilde{\theta}^c\|_\infty$ tends to 0.

THEOREM 8. *Suppose that the conditions in Lemma 8 hold. Then, for any $\lambda \geq 0$, almost surely*

$$\lim_{N \rightarrow \infty, \|\theta^c - \tilde{\theta}^c\|_\infty \rightarrow 0} \mathbb{D}(\mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N), \mathfrak{X}^\lambda(\theta^c, r^c)) = 0.$$

Proof. Note that

$$\begin{aligned} & \limsup_{N \rightarrow \infty, \|\theta^c - \tilde{\theta}^c\|_\infty \rightarrow 0} \mathbb{D}(\mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N), \mathfrak{X}^\lambda(\theta^c, r^c)) \\ & \leq \limsup_{N \rightarrow \infty, \|\theta^c - \tilde{\theta}^c\|_\infty \rightarrow 0} (\psi_{\theta^c, r^c}^\lambda)^{-1} \left((C_\lambda(\|\tilde{\theta}_N - \theta^c\|_\infty + |\tilde{r}_N - r^c|))^{1/p_\lambda} \right) \\ & \leq \limsup_{\|\theta^c - \tilde{\theta}^c\|_\infty \rightarrow 0} (\psi_{\theta^c, r^c}^\lambda)^{-1} \left((C_\lambda\|\theta^c - \tilde{\theta}^c\|_\infty)^{1/p_\lambda} \right) = 0. \end{aligned}$$

Since $(\psi_{\theta^c, r^c}^\lambda)^{-1} \left((C_\lambda\|\theta^c - \tilde{\theta}^c\|_\infty)^{1/p_\lambda} \right) \geq 0$, we complete the proof. ■

Theorem 8 demonstrates the asymptotic convergence of the x -solution set for the contaminated problem $(\tilde{\mathbf{B}}_N^\lambda)$ as the distance between the contaminated distribution $P_{\tilde{\theta}^c}$ and the true distribution P_{θ^c} approaches zero. We now derive a quantitative error bound for the deviation between x -solution sets when the distance between θ^c and $\tilde{\theta}^c$ is not zero. To this end, we need the following lemma.

LEMMA 9. *Suppose that the conditions in Theorem 8 hold and $\psi_{\theta^c, r^c}^\lambda$ is continuous for any fixed $\lambda \geq 0$. Then, for any $\tau > 0$, we have (i) $\limsup_{N \rightarrow \infty} \psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) \leq \psi_{\theta^c, r^c}^\lambda(\tau)$ and (ii) $\liminf_{N \rightarrow \infty} (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t) \geq (\psi_{\theta^c, r^c}^\lambda)^{-1}(t)$.*

Proof. To prove (i), let $\tau > \delta > 0$ and $\epsilon > 0$ be given. By the definition of $\psi_{\theta^c, r^c}^\lambda$, there exist $x^* \in \mathcal{X}$ and $\theta^* \in \vartheta^\lambda(x^*, \theta^c, r^c)$ such that $\mathbf{d}(x^*, \mathfrak{X}^\lambda(\theta^c, r^c)) \geq \tau + \delta$ and

$$\mathbf{d}\left(0, \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j}[\Phi(x^*, \xi)] + \mathcal{N}_X(x^*)\right) \leq \psi_{\theta^c, r^c}^\lambda(\tau + \delta) + \epsilon.$$

Since $(\hat{\theta}_N, \hat{r}_N) \rightarrow (\theta^c, r^c)$ and $\mathfrak{X}^\lambda$ is upper semicontinuous, there exists N_1 such that for all $N > N_1$, we have $\mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N) \subseteq \mathfrak{X}^\lambda(\theta^c, r^c) + \frac{\delta}{2}\mathbb{B}_\infty$. Hence, for any $x' \in \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)$, there exists $y \in \mathfrak{X}^\lambda(\theta^c, r^c)$ such that $\|x' - y\|_\infty \leq \frac{\delta}{2}$, which implies

$$\|x^* - x'\|_\infty \geq \|x^* - y\|_\infty - \|y - x'\|_\infty \geq \tau + \frac{\delta}{2}.$$

Thus, $\mathbf{d}(x^*, \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)) \geq \tau$.

On the other hand, due to the upper semicontinuity of $\vartheta^\lambda(x, \cdot, \cdot)$ from Theorem 6 and Proposition 2, there exist $N_2 \in \mathbb{N}$ and a sequence $\{\theta_N^*\}_{N \geq N_2}$ with $\theta_N^* \in \vartheta^\lambda(x^*, \hat{\theta}_N, \hat{r}_N)$ such that $\theta_N^* \rightarrow \theta^*$ as $N \rightarrow \infty$. Then, we have $\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) \leq \mathbf{d}\left(0, \sum_{j=1}^n \theta_{N,j}^* \mathbb{E}_{Q_j}[\Phi(x^*, \xi)] + \mathcal{N}_X(x^*)\right)$. Taking the limit yields

$$\begin{aligned} \limsup_{N \rightarrow \infty} \psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) &\leq \lim_{N \rightarrow \infty} \mathbf{d}\left(0, \sum_{j=1}^n \theta_{N,j}^* \mathbb{E}_{Q_j}[\Phi(x^*, \xi)] + \mathcal{N}_X(x^*)\right) \\ &= \mathbf{d}\left(0, \sum_{j=1}^n \theta_j^* \mathbb{E}_{Q_j}[\Phi(x^*, \xi)] + \mathcal{N}_X(x^*)\right) \leq \psi_{\theta^c, r^c}^\lambda(\tau + \delta) + \epsilon. \end{aligned}$$

Since $\epsilon > 0$ and $\delta > 0$ are arbitrary, we conclude $\limsup_{N \rightarrow \infty} \psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) \leq \psi_{\theta^c, r^c}^\lambda(\tau)$.

To prove (ii), let $t \geq 0$ be given and define $\tau^* = (\psi_{\theta^c, r^c}^\lambda)^{-1}(t) = \sup\{\tau \geq 0 \mid \psi_{\theta^c, r^c}^\lambda(\tau) \leq t\}$. Since $\psi_{\theta^c, r^c}^\lambda$ is continuous and non-decreasing, we have $\psi_{\theta^c, r^c}^\lambda(\tau^*) = t$ and $\psi_{\theta^c, r^c}^\lambda(\tau) < t$ for any $\tau < \tau^*$. By the result of part (i), $\limsup_{N \rightarrow \infty} \psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) \leq \psi_{\theta^c, r^c}^\lambda(\tau)$. This means that for the fixed $\tau < \tau^*$ and any $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$, we have $\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) \leq \psi_{\theta^c, r^c}^\lambda(\tau) + \epsilon$.

Since $\psi_{\theta^c, r^c}^\lambda(\tau) < t$, we can choose ϵ small enough such that $\psi_{\theta^c, r^c}^\lambda(\tau) + \epsilon < t$. Thus, for all $N \geq N_0$, $\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda(\tau) < t$, which means that for any $\tau < \tau^*$ and sufficiently large N , $(\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t) \geq \tau$. Since the above inequality holds for all $\tau < \tau^*$, we have $\liminf_{N \rightarrow \infty} (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}(t) \geq \tau^* = (\psi_{\theta^c, r^c}^\lambda)^{-1}(t)$. ■

Based on these conclusions, we can further derive the error bound corresponding to Theorem 8.

THEOREM 9. *Suppose that the conditions in Theorem 8 hold and $\psi_{\theta^c, r^c}^\lambda$ is continuous for any fixed $\lambda \geq 0$. Then, almost surely*

$$\lim_{N \rightarrow \infty} \mathbb{D}\left(\mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N), \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)\right) \leq (\psi_{\theta^c, r^c}^\lambda)^{-1}\left((C_\lambda \|\theta^c - \tilde{\theta}^c\|_\infty)^{1/p_\lambda}\right).$$

Proof. By Theorem 7, for any N and $\lambda \geq 0$,

$$\mathbb{D}\left(\mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N), \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N)\right) \leq (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}\left((C_\lambda \Delta_N)^{1/p_\lambda}\right).$$

Taking the limit on both sides of the inequality, for any $\delta \geq \Delta_N - \|\theta^c - \tilde{\theta}^c\|_\infty$, we have that

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathbb{D} \left(\mathfrak{X}^\lambda(\tilde{\theta}_N, \tilde{r}_N), \mathfrak{X}^\lambda(\hat{\theta}_N, \hat{r}_N) \right) &\leq \lim_{N \rightarrow \infty} (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1} \left((C_\lambda \Delta_N)^{1/p_\lambda} \right) \\ &\leq \lim_{N \rightarrow \infty} (\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1} \left((C_\lambda (\delta + \|\theta^c - \tilde{\theta}^c\|_\infty))^{1/p_\lambda} \right) \\ &\leq (\psi_{\theta^c, r^c}^\lambda)^{-1} \left((C_\lambda \|\theta^c - \tilde{\theta}^c\|_\infty)^{1/p_\lambda} \right), \end{aligned}$$

where the second inequality holds due to the non-decreasing property of $(\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1}$ and the last inequality follows from Lemma 9. Since δ can be taken arbitrarily small as $N \rightarrow \infty$, we obtain that as $N \rightarrow \infty$,

$$(\psi_{\hat{\theta}_N, \hat{r}_N}^\lambda)^{-1} \left((C_\lambda \Delta_N)^{1/p_\lambda} \right) \leq (\psi_{\theta^c, r^c}^\lambda)^{-1} \left((C_\lambda \|\theta^c - \tilde{\theta}^c\|_\infty)^{1/p_\lambda} \right).$$

Thus the proof is completed. ■

To clarify the relationships among different Bayesian DRVI problems discussed in Sections 3 and 4, we present a relationship diagram in Figure 2. Here, $(\tilde{\mathbf{B}}_N)$ denotes the perturbed formulation $(\tilde{\mathbf{B}}_N^\lambda)$ with $\lambda = 0$. Overall, Theorem 5 and Corollary 1 show that a properly designed Bayesian parametric ambiguity set—formulated in Theorem 1—ensures both finite-sample and asymptotic guarantees for the x -solution set of the regularized data-driven problem $(\mathbf{B}_N^\lambda)$. In addition, Theorems 7 and 9 establish quantitative stability of the x -solution set between $(\mathbf{B}_N^\lambda)$ and its perturbed counterpart $(\tilde{\mathbf{B}}_N^\lambda)$ under sample contamination. Theorem 8 further demonstrates that the perturbed problem $(\tilde{\mathbf{B}}_N^\lambda)$ asymptotically converges to the Bayesian DRVI formulation $(\mathbf{B}^c)$, thereby confirming the robustness of the x -solution set against data contamination.

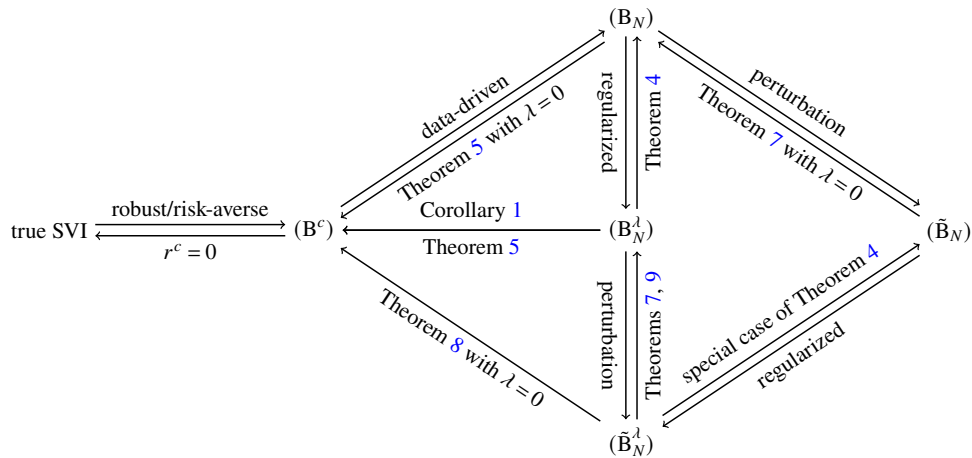


FIGURE 2. The relationship diagram of problems $(\mathbf{B}^c)$, $(\mathbf{B}_N)$, $(\mathbf{B}_N^\lambda)$, $(\tilde{\mathbf{B}}_N^\lambda)$.

5. Numerical experiment In this section, we illustrate the effectiveness of the proposed Bayesian DRVI model ($\mathbf{B}_N^\lambda$). We consider the distributionally robust multi-portfolio Nash equilibrium problem in Example 1 with $I = 2$ accounts, described as follows:

$$\min_{x^i \in \mathcal{X}_i} \max_{\theta^i \in \hat{\Theta}_N} \sum_{j=1}^n \theta_j^i \mathbb{E}_{Q_j} [\varphi(x^i, \xi)] + \frac{\kappa}{2} \|x^i - x^{-i}\|^2, \quad i = 1, 2. \quad (19)$$

Here, $\mathcal{X}_i = \{x^i \mid e^\top x^i \leq 1, x^i \in \mathbb{R}_+^d\}$ denotes the feasible portfolio set, where the inequality allows investment in a risk-free asset, and $\varphi(x^i, \xi) := -\log(1 + x^{i\top} \xi)$ represents the logarithmic disutility function which is convex in x^i . The term $\frac{\kappa}{2} \|x^i - x^{-i}\|^2$ penalizes deviations between the two portfolios, where $\kappa > 0$ is a common competition parameter (see [31]). In the following, let $x = (x^1, x^2)$ and denote

$$\begin{aligned} \varphi^n(x^i) &:= (\mathbb{E}_{Q_1}[\varphi(x^i, \xi)], \dots, \mathbb{E}_{Q_n}[\varphi(x^i, \xi)])^\top, \quad u_i(x^i, \theta^i) := \theta^{i\top} \varphi^n(x^i) - \lambda \|\theta^i\|^2, \\ U_i(x^i) &:= \max_{\theta^i \in \hat{\Theta}_N} u_i(x^i, \theta^i), \quad f_\lambda(x) := U_1(x^1) + U_2(x^2) + \frac{\kappa}{2} \|x^1 - x^2\|^2. \end{aligned}$$

As shown in Example 1, (19) leads to the following regularized Bayesian DRVI formulation:

$$\begin{aligned} 0 &\in \sum_{j=1}^n \theta_j^i \mathbb{E}_{Q_j} [\nabla_{x^i} \varphi(x^i, \xi)] + \kappa(x^i - x^{-i}) + \mathcal{N}_{\mathcal{X}_i}(x^i), \quad i = 1, 2, \\ \theta^i &\in \arg \max_{\theta' \in \hat{\Theta}_N} u_i(x^i, \theta'), \quad i = 1, 2. \end{aligned} \quad (20)$$

For each fixed $x^i \in \mathcal{X}_i$, the lower-level objective $u_i(x^i, \theta^i) = \theta^{i\top} \varphi^n(x^i) - \lambda \|\theta^i\|^2$ is 2λ -strongly concave in θ^i on the closed convex set $\hat{\Theta}_N$. Hence, the lower-level problem admits a unique solution $\theta^*(x^i) = \text{Proj}_{\hat{\Theta}_N} \left(\frac{1}{2\lambda} \varphi^n(x^i) \right)$ for $i = 1, 2$, where $\text{Proj}_S(\cdot)$ denotes the projection onto the set S . Moreover,

$$u_i(x^i, \theta^*(x^i)) - u_i(x^i, \theta) \geq \lambda \|\theta - \theta^*(x^i)\|^2 \geq \lambda \|\theta - \theta^*(x^i)\|_\infty^2.$$

Therefore, the growth condition introduced in Section 4 holds for the lower-level problem with $p_\lambda = 2$ and $\gamma_\lambda = \lambda$, which explains the stabilizing effect of regularization on the lower-level worst-case parameter selection. Substituting $\theta^*(x^i)$ into the upper-level problem simplifies the problem (20) to:

$$0 \in F(x) + \mathcal{N}_{\mathcal{X}_1 \times \mathcal{X}_2}(x), \quad (21)$$

where

$$F(x) = \begin{pmatrix} F_1(x) \\ F_2(x) \end{pmatrix}, \quad F_i(x) = \sum_{j=1}^n \theta_j^*(x^i) \mathbb{E}_{Q_j} [\nabla_{x^i} \varphi(x^i, \xi)] + \kappa(x^i - x^{-i}), \quad i = 1, 2.$$

For problem (21), we have the following conclusion:

PROPOSITION 3. Assume that for each $j = 1, \dots, n$, (i) there exists $\beta \in [0, 1)$ such that for each $i = 1, 2$ and all $x^i \in \mathcal{X}_i$, $x^{i\top} \xi \geq -\beta$ almost surely under Q_j , and (ii) there exists $\bar{\sigma}_j^2 < \infty$ such that $\mathbb{E}_{Q_j} \|\xi\|^2 \leq \bar{\sigma}_j^2$. Then, F is monotone and Lipschitz continuous on $\mathcal{X}_1 \times \mathcal{X}_2$ with modulus $L_F = (\sqrt{2} + \frac{\sqrt{2n}}{2\lambda}) \frac{\max_j \bar{\sigma}_j^2}{(1-\beta)^2} + 2\sqrt{2}\kappa$.

Proof. For each $x \in \mathcal{X}_1 \times \mathcal{X}_2$ and direction $d = (d^1, d^2)$, define $q(t) := [f_\lambda(x + td) - f_\lambda(x)]/t$, we have

$$\begin{aligned} \limsup_{t \downarrow 0} q(t) &= \limsup_{t \downarrow 0} \frac{f_\lambda(x + td) - f_\lambda(x)}{t} \\ &\geq \sum_{i=1}^2 \limsup_{t \downarrow 0} \frac{U_i(x^i + td^i) - U_i(x^i)}{t} + \kappa(x^1 - x^2)^\top (d^1 - d^2) \\ &\geq \sum_{i=1}^2 \limsup_{t \downarrow 0} \frac{u_i(x^i + td^i, \theta^*(x^i)) - u_i(x^i, \theta^*(x^i))}{t} + \kappa(x^1 - x^2)^\top (d^1 - d^2) \\ &= \sum_{i=1}^2 \nabla_{x^i} u_i(x^i, \theta^*(x^i))^\top d^i + \kappa(x^1 - x^2)^\top (d^1 - d^2). \end{aligned}$$

Similarly, we have $\liminf_{t \downarrow 0} q(t) \leq \sum_{i=1}^2 \nabla_{x^i} u_i(x^i, \theta^*(x^i))^\top d^i + \kappa(x^1 - x^2)^\top (d^1 - d^2)$. Thus, the directional derivative exists. Moreover, by Danskin's theorem, we have

$$\nabla_{x^i} f_\lambda(x) = \nabla_{x^i} u_i(x^i, \theta^*(x^i)) + \kappa(x^i - x^{-i}) = \sum_{j=1}^n \theta_j^*(x^i) \mathbb{E}_{Q_j} [\nabla_{x^i} \varphi(x^i, \xi)] + \kappa(x^i - x^{-i}) = F_i(x).$$

With the convexity of f_λ , we can deduce that $F = \nabla f_\lambda$ is monotone.

For the Lipschitz property, denote $g_j(x^i) = \mathbb{E}_{Q_j} [\nabla_{x^i} \varphi(x^i, \xi)]$, we have from the definition of φ that

$$\|g_j(x^i)\| \leq \frac{\bar{\sigma}_j}{(1-\beta)} =: L_{\varphi,j}, \quad \|\nabla g_j(x^i)\| \leq \mathbb{E}_{Q_j} \left[\frac{\|\xi\|^2}{(1-\beta)^2} \right] \leq \frac{\bar{\sigma}_j^2}{(1-\beta)^2} =: L_{g,j}.$$

Thus g_j is $L_{g,j}$ -Lipschitz, $\mathbb{E}_{Q_j} [\varphi(\cdot, \xi)]$ is $L_{\varphi,j}$ -Lipschitz and then $\theta^*(\cdot)$ is $\frac{\sqrt{n} \max_j L_{\varphi,j}}{2\lambda}$ -Lipschitz.

Based on the above results, we have

$$\begin{aligned} &\|F_i(x) - F_i(y)\| \\ &= \left\| \sum_j \theta_j^*(x^i) g_j(x^i) - \theta_j^*(y^i) g_j(y^i) + \kappa(x^i - y^i) - \kappa(x^{-i} - y^{-i}) \right\| \\ &\leq \sum_j \theta_j^*(x^i) \|g_j(x^i) - g_j(y^i)\| + \sum_j \|\theta_j^*(x^i) - \theta_j^*(y^i)\| \|g_j(y^i)\| + \kappa \|x^{-i} - y^{-i}\| + \kappa \|x^i - y^i\| \\ &\leq \max_j L_{g,j} \|x^i - y^i\| + \frac{n \max_j L_{\varphi,j}^2}{2\lambda} \|x^i - y^i\| + 2\kappa \|x - y\| \leq L_{\bar{F}} \|x - y\|, \end{aligned}$$

where $L_{\bar{F}} := (1 + \frac{n}{2\lambda}) \frac{\max_j \bar{\sigma}_j^2}{(1-\beta)^2} + 2\kappa$. By setting $L_F = \sqrt{2}L_{\bar{F}}$, the stated modulus is proved. ■

As for the assumptions in Proposition 3, assumption (i) is mild in financial contexts where $x^{i^\top} \xi \geq -1$ is often satisfied for $i = 1, 2$. For any fixed $\lambda > 0$, Proposition 3 shows that the operator F in (21) is monotone and L_F -Lipschitz continuous on the nonempty closed convex set $\mathcal{X}_1 \times \mathcal{X}_2$. Therefore, the standard convergence theory of the extragradient method for monotone Lipschitz variational inequalities applies to the fixed- λ problem (21); see, e.g., [39, 43]. In particular, Algorithm 1 is covered by this theory whenever the constant step size is chosen such that $\eta \in (0, \frac{1}{2L_F})$. Moreover, define the variational inequality gap function by $\text{Gap}(z) := \max_{u \in \mathcal{X}_1 \times \mathcal{X}_2} \langle F(u), z - u \rangle$. Then, from [29, 43], with a constant stepsize η of order $1/L_F$, the averaged auxiliary iterates $\bar{y}^K := \frac{1}{K} \sum_{k=0}^{K-1} y^k$ satisfy

$$\text{Gap}(\bar{y}^K) \leq \frac{C L_F}{K},$$

where $C > 0$ is a constant independent of K and L_F . Therefore, to achieve $\text{Gap}(\bar{y}^K) \leq \varepsilon$, it suffices to take $K = O\left(\frac{L_F}{\varepsilon}\right)$. Since Proposition 3 yields $L_F = O(1/\lambda)$ as $\lambda \downarrow 0$, the required number of extragradient iterations scales as $K = O\left(\frac{1}{\lambda \varepsilon}\right)$.

REMARK 7. As discussed in Remark 6, choosing $\lambda_N \asymp N^{-1/2}$ balances the estimation error and the regularization bias. Let K_N denote the number of extragradient iterations used to solve the corresponding problem with regularization parameter λ_N . By Proposition 3, $L_F = O(1/\lambda_N)$, and hence the preceding gap bound gives $\text{Gap}(\bar{y}^{K_N}) = O\left(\frac{1}{\lambda_N K_N}\right)$. To make the finite-iteration optimization error of the same order as the regularization bias $O(\lambda_N)$, it is sufficient to choose $K_N \asymp \lambda_N^{-2}$. Since $\lambda_N \asymp N^{-1/2}$, this gives $K_N \asymp N$ and $\text{Gap}(\bar{y}^{K_N}) = O(N^{-1/2})$. Thus, the finite-iteration optimization error has the same order as the estimation error and regularization bias.

To further specify the setting of problem (20), we consider $d = 10$ risky assets with $n = 3$, representing three potential market regimes: bear, oscillating and bull markets. To generate the necessary data for stock returns, we adopt the following factor model widely used in [11, 68]:

$$\xi_m^j = a_m^j + b_m^j \zeta_{\text{market}}^j + e_m^j, \quad j = 1, 2, 3, \quad m = 1, \dots, 10,$$

where the scalar market factors follow $\zeta_{\text{market}}^j \sim \mathcal{N}(\mu_j, \sigma_{z,j}^2)$, $j = 1, 2, 3$, with

$$(\mu_1, \sigma_{z,1}^2) = (0.005, 0.001), \quad (\mu_2, \sigma_{z,2}^2) = (0.010, 0.004), \quad (\mu_3, \sigma_{z,3}^2) = (0.015, 0.007).$$

The intercepts a_m^j capture each asset's baseline return and are set to -0.0005 for bear market ($j = 1$), 0 for oscillating market ($j = 2$), and 0.0005 for bull market ($j = 3$); the factor loadings b_m^j

Algorithm 1 Extragradient method for the fixed- λ regularized Bayesian DRVI problem (21)**Require:** Regularization parameter $\lambda > 0$, step size $\eta \in (0, \frac{1}{2L_F})$, tolerance $\varepsilon > 0$, initial point

$$x^0 \in \mathcal{X}_1 \times \mathcal{X}_2$$

```

1: for  $k = 0, 1, 2, \dots$  do
2:   Compute  $F$  at  $x^k$ :
3:   for  $i = 1, 2$  do
4:      $z^{i,k} \leftarrow \frac{1}{2\lambda} \varphi^n(x^{i,k})$ 
5:      $\theta^{\lambda,i} \leftarrow \text{Proj}_{\hat{\Theta}_N}(z^{i,k})$ 
6:      $F_i^k \leftarrow \sum_{j=1}^n \theta_j^{\lambda,i} \mathbb{E}_{Q_j}[\nabla_{x^i} \varphi(x^{i,k}, \xi)] + \kappa(x^{i,k} - x^{-i,k})$ 
7:   end for
8:    $F^k \leftarrow (F_1^k, F_2^k)^\top$ 
9:   Extragradient step:  $y^k \leftarrow \text{Proj}_{\mathcal{X}_1 \times \mathcal{X}_2}(x^k - \eta F^k)$ 
10:  Recompute  $F$  at  $y^k$ :
11:  for  $i = 1, 2$  do
12:     $\tilde{z}^{i,k} \leftarrow \frac{1}{2\lambda} \varphi^n(y^{i,k})$ 
13:     $\tilde{\theta}^{\lambda,i} \leftarrow \text{Proj}_{\hat{\Theta}_N}(\tilde{z}^{i,k})$ 
14:     $\tilde{F}_i^k \leftarrow \sum_{j=1}^n \tilde{\theta}_j^{\lambda,i} \mathbb{E}_{Q_j}[\nabla_{x^i} \varphi(y^{i,k}, \xi)] + \kappa(y^{i,k} - y^{-i,k})$ 
15:  end for
16:   $\tilde{F}^k \leftarrow (\tilde{F}_1^k, \tilde{F}_2^k)^\top$ 
17:  Update:  $x^{k+1} \leftarrow \text{Proj}_{\mathcal{X}_1 \times \mathcal{X}_2}(x^k - \eta \tilde{F}^k)$  and  $\bar{y}^{k+1} \leftarrow \frac{1}{k+1} \sum_{t=0}^k y^t$ 
18:  if  $\text{Gap}(\bar{y}^{k+1}) \leq \varepsilon$  then
19:    return  $x^* \leftarrow \bar{y}^{k+1}$ 
20:  end if
21: end for

```

determine sensitivity to the market factor and increase linearly with m : $b_m^1 = 0.60 + 0.02(m - 1)$ in bear market, $b_m^2 = 0.80 + 0.015(m - 1)$ in oscillating market, and $b_m^3 = 1.00 + 0.01(m - 1)$ in bull market. The residuals $e_m^j \sim \mathcal{N}(0, \sigma_{e,j}^2)$ follow independent normal distributions with variances $\sigma_{e,1}^2 = 2 \times 10^{-5}$, $\sigma_{e,2}^2 = 1 \times 10^{-5}$, and $\sigma_{e,3}^2 = 0.5 \times 10^{-5}$, reflecting higher asset-specific risk in downturns and lower risk in upturns. The mixture weight is set to $\theta^c = (0.5, 0.25, 0.25)$. Here, we set $\kappa = 0.1$, $\lambda = 0.01$ and $r^c = 0$ to model a risk-neutral decision-maker.

We generate a large dataset $\{\xi^s\}_{s=1}^{N_{\text{sim}}}$ with $N_{\text{sim}} = 10^6$ to simulate the true equilibrium solution x^c as a benchmark. For comparison, we compute the approximate solution x^* by different data-driven methods on training samples with size $N \in \{20, 50, 500, 1000, 3000\}$: (i) the empirical SVI solved by sample average approximation (SAA) [67]; (ii) the posterior predictive SVI, which replaces the true distribution by the Bayesian posterior predictive distribution [59, 66] serving as a non-robust Bayesian benchmark; (iii) two frequentist robust benchmarks with ℓ_1 and modified χ^2 ambiguity sets; and (iv) our Bayesian DRVI method. For the ℓ_1 -, modified χ^2 -, and our Bayesian DRVI methods, the ambiguity sets are calibrated under the same level $\alpha = 0.05$. Each method is run over 100 independent trials. For each trial, to quantify the accuracy of the solution x^* , we compute the residual distance $d(x^*) = \|x^* - x^c\|$ and the squared relative utility error [66]:

$$e(x^*) = \left(\frac{U(x^*) - U(x^c)}{U(x^c)} \right)^2, \text{ where } U(x) = \frac{1}{N_{\text{sim}}} \sum_{s=1}^{N_{\text{sim}}} -\log(1 + \xi^s \top x^1) + \frac{\kappa}{2} \|x^1 - x^2\|^2.$$

Here, since the game is symmetric across the two accounts, we focus on account 1's utility without loss of generality. Furthermore, to quantify out-of-sample performance, we generate a fixed test set $\{\xi^t\}_{t=1}^{N_{\text{test}}}$ with $N_{\text{test}} = 5000$. Based on the single-period returns $\{R_t = \xi^t \top x^{1*}\}_{t=1}^{N_{\text{test}}}$, we further assess the out-of-sample performance of the optimal portfolio with additional metrics in [6]—Sharpe ratio (measuring return per unit of total risk) and risk-adjusted return on capital (RAROC; measuring return per unit of tail risk) at 10% tail level. Specifically, let the out-of-sample mean and standard deviation of the returns be

$$\bar{R} = \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} R_t, \quad \hat{\sigma}_R = \sqrt{\frac{1}{N_{\text{test}} - 1} \sum_{t=1}^{N_{\text{test}}} (R_t - \bar{R})^2}.$$

The Sharpe ratio is then $\widehat{\text{SR}} = \frac{\bar{R}}{\hat{\sigma}_R}$.

For the RAROC at 10% tail level, define the losses $l_t := -R_t$. Let the empirical value-at-risk be $\widehat{\text{VaR}}_{10\%}$ as the 10%-quantile of $\{l_t\}_{t=1}^{N_{\text{test}}}$, and let the empirical expected shortfall of losses be

$$\widehat{\text{ES}}_{10\%} = \frac{1}{\#\{t : l_t \geq \widehat{\text{VaR}}_{10\%}\}} \sum_{t: l_t \geq \widehat{\text{VaR}}_{10\%}} l_t,$$

where $\#\{\cdot\}$ denotes the cardinality (number of elements) of a set. Then the RAROC at 10% tail level is calculated as $\widehat{\text{R}}_{10\%} = \frac{\bar{R}}{\widehat{\text{ES}}_{10\%}}$.

We report in Tables 2-5 the sample means and variances of all performance metrics across 100 trials, allowing a comprehensive assessment of each method's estimation accuracy, sensitivity to sampling variability, and risk-adjusted performance as the training sample size increases.

The observations derived from Tables 2-5 lead to the following conclusions.

TABLE 2. Residual distance $d(x^*)$: means and variances across training sample sizes

	Method	$N = 20$	$N = 50$	$N = 500$	$N = 1000$	$N = 3000$
Mean ($\times 10^{-1}$)	ℓ_1 norm	7.311	6.914	5.629	3.203	0.669
	Modified χ^2	7.111	7.027	6.456	6.336	6.221
	SAA	6.205	5.325	1.758	0.805	0.217
	Bayesian	6.123	5.019	1.555	0.775	0.158
	Posterior predictive	6.182	5.266	1.694	0.806	0.214
Variance ($\times 10^{-2}$)	ℓ_1 norm	4.2579	4.1647	3.4892	3.2749	2.5608
	Modified χ^2	3.6275	3.2301	2.2205	1.2423	0.7599
	SAA	4.9777	4.2403	3.5268	2.4571	0.8582
	Bayesian	4.1647	4.0356	3.3361	2.2159	0.7046
	Posterior predictive	4.7948	4.1757	3.5347	2.2902	0.8152

TABLE 3. Squared relative utility error $e(x^*)$: means and variances across training sample sizes

	Method	$N = 20$	$N = 50$	$N = 500$	$N = 1000$	$N = 3000$
Mean ($\times 10^{-3}$)	ℓ_1 norm	3.843	3.440	2.625	1.636	0.368
	Modified χ^2	3.300	3.973	4.024	4.095	4.273
	SAA	1.464	1.210	1.034	0.330	0.151
	Bayesian	1.360	1.113	0.828	0.315	0.125
	Posterior predictive	1.423	1.181	1.044	0.330	0.151
Variance ($\times 10^{-6}$)	ℓ_1 norm	2.3010	1.9958	1.5284	1.4141	0.8297
	Modified χ^2	2.3468	1.0422	0.5144	0.2289	0.1330
	SAA	1.8827	1.2819	0.8018	0.4076	0.0649
	Bayesian	1.5179	1.2662	0.7540	0.3783	0.0367
	Posterior predictive	1.8601	1.2795	0.7946	0.4071	0.0653

TABLE 4. Sharpe ratio $\widehat{SR}$: means and variances across training sample sizes

	Method	$N = 20$	$N = 50$	$N = 500$	$N = 1000$	$N = 3000$
Mean ($\times 10^{-1}$)	ℓ_1 norm	1.798	1.794	1.800	1.808	1.821
	Modified χ^2	1.795	1.791	1.789	1.790	1.789
	SAA	1.826	1.828	1.829	1.827	1.826
	Bayesian	1.820	1.824	1.829	1.827	1.826
	Posterior predictive	1.827	1.828	1.829	1.827	1.826
Variance ($\times 10^{-6}$)	ℓ_1 norm	1.7977	1.6436	0.8393	1.1894	0.7318
	Modified χ^2	1.2556	0.4248	0.1006	0.0889	0.0738
	SAA	2.0414	1.0569	0.2636	0.1862	0.1605
	Bayesian	1.7583	0.9653	0.2527	0.1542	0.1268
	Posterior predictive	1.9704	1.0689	0.2611	0.1941	0.1607

TABLE 5. RAROC at 10% tail level $\widehat{R}_{10\%}$: means and variances across training sample sizes

	Method	$N = 20$	$N = 50$	$N = 500$	$N = 1000$	$N = 3000$
Mean ($\times 10^{-1}$)	ℓ_1 norm	1.163	1.160	1.164	1.171	1.180
	Modified χ^2	1.160	1.157	1.155	1.155	1.154
	SAA	1.184	1.185	1.185	1.184	1.183
	Bayesian	1.180	1.183	1.185	1.184	1.183
	Posterior predictive	1.184	1.185	1.185	1.184	1.183
Variance ($\times 10^{-6}$)	ℓ_1 norm	1.2628	1.1974	0.6883	0.5591	0.1941
	Modified χ^2	0.9043	0.3301	0.0989	0.0881	0.0623
	SAA	1.6741	0.9903	0.1419	0.0941	0.0675
	Bayesian	1.2089	0.6837	0.1366	0.0908	0.0629
	Posterior predictive	1.3723	0.8958	0.1401	0.0936	0.0668

- **Consistency and estimation accuracy.** Across all sample sizes, the Bayesian DRVI attains the smallest mean residual distances to the benchmark x^c and the lowest squared relative utility errors (Tables 2-3), demonstrating its superior estimation accuracy. Moreover, Table 2 shows that the Bayesian solutions improve steadily as N increases, consistent with the theoretical convergence results. In contrast, the ℓ_1 norm method is outperformed by Bayesian DRVI, posterior predictive SVI, and empirical SVI, while the modified χ^2 scheme exhibits a persistent bias, indicative of over-conservatism that does not vanish with additional data.

- **Stability and variance reduction.** Tables 2-3 show that the Bayesian DRVI maintains low variance across sample sizes, particularly for large N , and achieves the lowest variance among all methods for $N = 3000$. The posterior predictive SVI is also more stable than SAA in most cases, which is consistent with the fact that it remains within a fixed parametric model rather than relying on the empirical distribution. However, because it incorporates posterior information only through Bayesian averaging under a single predictive distribution, it is still less stable than Bayesian DRVI, which explicitly accounts for such uncertainty through posterior-credible robustification. While the modified χ^2 approach also displays small variances, this is accompanied by pronounced bias, reflecting variance reduction driven by excessive pessimism rather than genuine robustness. In contrast, among the essentially unbiased approaches (ℓ_1 norm, Bayesian and SAA), the Bayesian method consistently attains both the lowest mean errors and the smallest variances, highlighting its superior stability and reliability in finite samples.

- **Risk-return performance.** The SAA and Bayesian methods form a top tier of performance for $N \geq 500$, achieving nearly identical and superior results in both the Sharpe ratio and RAROC at 10% tail level (Tables 4-5). In smaller samples ($N = 20, 50$), SAA may achieve a slightly higher mean performance, but the Bayesian method typically exhibits smaller variance, confirming that it provides constantly stable risk-return trade-offs across different data realizations. The inherent conservatism of the ℓ_1 and modified χ^2 methods leads to suboptimal decisions, resulting in consistently lower performance. This demonstrates the substantial opportunity cost of using overly pessimistic ambiguity sets, while the Bayesian approach delivers excellent risk-adjusted and tail-controlled returns with improved reliability.

Finally, we illustrate the regularization-dependent iteration rule derived in Remark 7. We set $\lambda_N = 0.3N^{-1/2}$ and consider $N \in \{50, 100, 200, 500, 1000, 2000, 3000\}$. For each λ_N , the extragradient method is run with the three iteration budgets $K_N = 0.1\lambda_N^{-2}$, $K_N = 0.2\lambda_N^{-2}$, and $K_N = 0.4\lambda_N^{-2}$. Each setting is repeated independently 100 times, and the average variational inequality gap $\text{Gap}(\bar{y}^{K_N})$,

where $\bar{y}^{K_N}$ is the averaged auxiliary iterate defined above, is reported. Figure 3 shows that the average gap decreases as λ_N decreases under all three iteration budgets. For each fixed λ_N , increasing the constant multiplying λ_N^{-2} also produces a smaller gap. These observations are consistent with the relation $K_N \propto \lambda_N^{-2}$ derived in Remark 7.

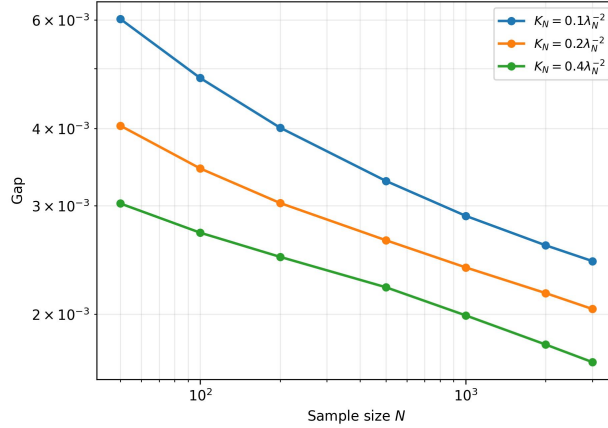


FIGURE 3. Average variational inequality gap $\text{Gap}(\bar{y}^{K_N})$ under $K_N = 0.1\lambda_N^{-2}$, $K_N = 0.2\lambda_N^{-2}$, and $K_N = 0.4\lambda_N^{-2}$.

In summary, the proposed Bayesian DRVI method effectively balances the trade-off between robustness and optimality. Compared with posterior predictive and empirical SVI models, the proposed method demonstrates the additional value of posterior robustification beyond Bayesian updating alone and provides more reliable decisions. Compared with the frequentist robust benchmarks, it also mitigates excessive conservatism.

6. Concluding remarks Building on the DRVI framework in [60], we consider a Bayesian formulation $(\mathbf{B}^c)$ which captures distributional uncertainty through the parametric ambiguity set Θ^c . To ensure tractability, we construct a data-driven ambiguity set with posterior coverage guarantees and introduce the regularized Bayesian DRVI model $(\mathbf{B}_N^\lambda)$. Under mild assumptions, we prove the existence of solutions for both Bayesian DRVI $(\mathbf{B}_N)$ and its regularized counterpart $(\mathbf{B}_N^\lambda)$. We further establish the convergence of the solutions of $(\mathbf{B}_N^\lambda)$ to a solution of the problem $(\mathbf{B}^c)$ as sample size grows and the regularization vanishes. Moreover, as $\lambda \downarrow 0$, the regularized maximizer of the lower-level problem in $(\mathbf{B}_N^\lambda)$ converges to the least-norm maximizer of the lower-level problem in $(\mathbf{B}_N)$. To address practical finite-sample and contamination effects, we derive finite-sample guarantees for the optimal solutions of $(\mathbf{B}_N^\lambda)$ and quantify the solution stability between this formulation and its perturbed variant $(\tilde{\mathbf{B}}_N^\lambda)$. The effectiveness of the proposed framework is demonstrated

through its application to a distributionally robust multi-portfolio Nash equilibrium problem. By integrating Bayesian learning with DRVI, our framework provides a robust and practical foundation for decision-making under uncertainty.

It is natural to extend the Bayesian DRVI framework to stochastic generalized equations. While statistical robustness results for stochastic generalized equations have been established in [21], analogous theoretical guarantees for the Bayesian DRVI framework remain an open question. We leave these extensions to future work.

References

- [1] Bach F (2024) *Learning Theory from First Principles* (MIT Press).
- [2] Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357.
- [3] Bertsimas D, Gupta V, Kallus N (2018) Robust sample average approximation. *Mathematical Programming* 171:217–282.
- [4] Bertsimas D, Shtern S, Sturt B (2022) Two-stage sample robust optimization. *Operations Research* 70(1):624–640.
- [5] Bichler M, Fichtl M, Oberlechner M (2025) Computing Bayes–Nash equilibrium strategies in auction games via simultaneous online dual averaging. *Operations Research* 73(2):1102–1127.
- [6] Bielecki TR, Cialenco I, Pitera M (2018) A unified approach to time consistency of dynamic risk measures and dynamic performance measures in discrete time. *Mathematics of Operations Research* 43(1):204–221.
- [7] Cao Y, Rusmevichientong P, Topaloglu H (2023) Revenue management under a mixture of independent demand and multinomial logit models. *Operations Research* 71(2):603–625.
- [8] Chan G, Van Parys B, Bennouna A (2024) From distributional robustness to robust statistics: a confidence sets perspective. *arXiv preprint arXiv:2410.14008*.
- [9] Chen X, Pong TK, Wets RJB (2017) Two-stage stochastic variational inequalities: an ERM-solution procedure. *Mathematical Programming* 165:71–111.
- [10] Chen X, Wets RJB, Zhang Y (2012) Stochastic variational inequalities: residual minimization smoothing sample average approximations. *SIAM Journal on Optimization* 22(2):649–673.
- [11] Chen Z, Peng S, Liu J (2018) Data-driven robust chance constrained problems: a mixture model approach. *Journal of Optimization Theory and Applications* 179:1065–1085.
- [12] Coppens P, Patrinos P (2023) Ordered risk minimization: Learning more from less data. *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2003–2009 (IEEE).
- [13] Cutler J, Drusvyatskiy D, Harchaoui Z (2023) Stochastic optimization under distributional drift. *Journal of Machine Learning Research* 24(147):1–56.

- [14] Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* 58(3):595–612.
- [15] Efron B, Hastie T (2021) *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, volume 6 (Cambridge University Press).
- [16] Everitt BS, Hand DJ (2013) *Finite Mixture Distributions* (Springer Science & Business Media).
- [17] Facchinei F, Pang JS (2003) *Finite-Dimensional Variational Inequalities and Complementarity Problems* (Springer).
- [18] Fang H, Chen X, Fukushima M (2007) Stochastic R_0 matrix linear complementarity problems. *SIAM Journal on Optimization* 18(2):482–506.
- [19] Farokhi F (2023) Distributionally robust optimization with noisy data for discrete uncertainties using total variation distance. *IEEE Control Systems Letters* 7:1494–1499.
- [20] Gao R, Chen X, Kleywegt AJ (2024) Wasserstein distributionally robust optimization and variation regularization. *Operations Research* 72(3):1177–1191.
- [21] Guo S, Xu H (2023) Data perturbations in stochastic generalized equations: statistical robustness in static and sample average approximated models. *Mathematical Programming* 202(1):135–168.
- [22] Gupta V (2019) Near-optimal Bayesian ambiguity sets for distributionally robust optimization. *Management Science* 65(9):4242–4260.
- [23] Gürkan G, Yonca Özge A, Robinson SM (1999) Sample-path solution of stochastic variational inequalities. *Mathematical Programming* 84(2).
- [24] Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2015) A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming* 151:35–62.
- [25] Jaiswal P, Honnappa H, Rao VA (2023) Bayesian joint chance constrained optimization: approximations and statistical consistency. *SIAM Journal on Optimization* 33(3):1968–1995.
- [26] Jiang J (2024) Distributionally robust variational inequalities: relaxation, quantification and discretization. *Journal of Optimization Theory and Applications* 203(1):227–255.
- [27] Jiang J, Shi Y, Wang X, Chen X (2019) Regularized two-stage stochastic variational inequalities for cournot-nash equilibrium under uncertainty. *Journal of Computational Mathematics* 813–842.
- [28] Jiang R, Guan Y (2018) Risk-averse two-stage stochastic program with distributional ambiguity. *Operations Research* 66(5):1390–1405.
- [29] Juditsky A, Nemirovski A, Tauvel C (2011) Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems* 1(1):17–58.
- [30] Kuhn D, Mohajerin Esfahani P, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: theory and applications in machine learning. *Operations Research & Management Science in the Age of Analytics*, 130–166 (INFORMS).

- [31] Lacker D, Zariphopoulou T (2019) Mean field and n-agent games for optimal investment under relative performance criteria. *Mathematical Finance* 29(4):1003–1038.
- [32] Lam H (2019) Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research* 67(4):1090–1105.
- [33] Lampariello L, Neumann C, Ricci JM, Sagratella S, Stein O (2021) Equilibrium selection for multi-portfolio optimization. *European Journal of Operational Research* 295(1):363–373.
- [34] Lecué G, Lerasle M (2020) Robust machine learning by median-of-means. *The Annals of Statistics* 48(2):906–931.
- [35] Lehmann EL, Romano JP (2005) *Testing Statistical Hypotheses* (Springer).
- [36] Li H, Webster S, Mason N, Kempf K (2019) Product-line pricing under discrete mixed multinomial logit demand: winner—2017 M&SOM practice-based research competition. *Manufacturing & Service Operations Management* 21(1):14–28.
- [37] Liu J, Su Z, Xu H (2024) Bayesian distributionally robust Nash equilibrium and its application. *arXiv preprint arXiv:2410.20364*.
- [38] Lu S, Budhiraja A (2013) Confidence regions for stochastic variational inequalities. *Mathematics of Operations Research* 38(3):545–568.
- [39] Malitsky Y (2015) Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization* 25(1):502–520.
- [40] McLachlan GJ, Peel D (2000) *Finite Mixture Models* (John Wiley & Sons).
- [41] Mehta R, Roulet V, Pillutla K, Liu L, Harchaoui Z (2023) Stochastic optimization for spectral risk measures. *International Conference on Artificial Intelligence and Statistics*, 10112–10159 (PMLR).
- [42] Milz J, Shapiro A, Zhou E (2026) Stochastic optimal control with side information and Bayesian learning. *arXiv preprint arXiv:2602.22047*.
- [43] Nemirovski A (2004) Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization* 15(1):229–251.
- [44] Ordóñez F, Stier-Moses NE (2010) Wardrop equilibria with risk-averse users. *Transportation Science* 44(1):63–86.
- [45] Oren Y, Sagawa S, Hashimoto TB, Liang P (2019) Distributionally robust language modeling. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4227–4237.
- [46] Peel D, McLachlan GJ (2000) Robust mixture modelling using the t distribution. *Statistics and Computing* 10:339–348.
- [47] Pichler A, Xu H (2022) Quantitative stability analysis for minimax distributionally robust risk optimization. *Mathematical Programming* 191(1):47–77.

- [48] Redner RA, Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 26(2):195–239.
- [49] Robinson SM (1996) Analysis of sample-path optimization. *Mathematics of Operations Research* 21(3):513–528.
- [50] Rockafellar RT (2024) Distributional robustness, stochastic divergences, and the quadrangle of risk. *Computational Management Science* 21(1):34.
- [51] Rockafellar RT, Uryasev S (2002) Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance* 26(7):1443–1471.
- [52] Römisch W (2003) Stability of stochastic programming problems. *Handbooks in Operations Research and Management Science* 10:483–554.
- [53] Rossi PE, Allenby GM (2003) Bayesian statistics and marketing. *Marketing Science* 22(3):304–328.
- [54] Schöttle K, Werner R (2009) Robustness properties of mean-variance portfolios. *Optimization* 58(6):641–663.
- [55] Schwarz G (1978) Estimating the dimension of a model. *The Annals of Statistics* 461–464.
- [56] Shanbhag UV, Infanger G, Glynn PW (2011) A complementarity framework for forward contracting under uncertainty. *Operations Research* 59(4):810–834.
- [57] Shapiro A (2017) Distributionally robust stochastic programming. *SIAM Journal on Optimization* 27(4):2258–2275.
- [58] Shapiro A, Zhou E, Lin Y (2023) Bayesian distributionally robust optimization. *SIAM Journal on Optimization* 33(2):1279–1304.
- [59] Shapiro A, Zhou E, Lin Y, Wang Y (2026) Episodic Bayesian optimal control with unknown randomness distributions. *Operations Research* 74(3):1437–1455.
- [60] Sun H, Shapiro A, Chen X (2023) Distributionally robust stochastic variational inequalities. *Mathematical Programming* 200(1):279–317.
- [61] Tseng P (2004) An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research* 29(1):27–44.
- [62] Urteaga I, Wiggins CH (2018) Nonparametric Gaussian mixture models for the multi-armed bandit. *arXiv preprint arXiv:1808.02932*.
- [63] van de Geer R, den Boer AV (2022) Price optimization under the finite-mixture logit model. *Management Science* 68(10):7480–7496.
- [64] Van der Vaart AW (2000) *Asymptotic Statistics*, volume 3 (Cambridge University Press).
- [65] Wasserman L (2006) *All of Nonparametric Statistics* (Springer Science & Business Media).
- [66] Wu D, Zhu H, Zhou E (2018) A Bayesian risk approach to data-driven stochastic optimization: formulations and asymptotics. *SIAM Journal on Optimization* 28(2):1588–1612.

- [67] Xu H (2010) Sample average approximation methods for a class of stochastic variational inequality problems. *Asia-Pacific Journal of Operational Research* 27(01):103–119.
- [68] Zhu S, Fan M, Li D (2014) Portfolio management with robustness in both prediction and decision: a mixture model based learning approach. *Journal of Economic Dynamics and Control* 48:1–25.
- [69] Zhu S, Fukushima M (2009) Worst-case conditional value-at-risk with application to robust portfolio management. *Operations Research* 57(5):1155–1168.