Percentile Criterion Optimization in Offline Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

In reinforcement learning, robust policies for high-stakes decision-making problems with limited data are usually computed by optimizing the *percentile criterion*. The percentile criterion is optimized by constructing an *uncertainty set* that contains the true model with high probability and optimizing the policy for the worst model in the set. Since the percentile criterion is non-convex, constructing uncertainty sets is often challenging. Existing works use *Bayesian credible regions* as uncertainty sets, but they are often unnecessarily large and result in learning overly conservative policies. To overcome these shortcomings, we propose a novel Value-at-Risk based dynamic programming algorithm to optimize the percentile criterion without explicitly constructing any uncertainty sets. Our theoretical and empirical results show that our algorithm implicitly constructs much smaller uncertainty sets and learns less conservative robust policies.

1 Introduction

Batch Reinforcement Learning (Batch RL) (18) is popularly used for solving sequential decision-making problems using limited data. These algorithms are crucial in high-stakes domains where exploration is either infeasible or expensive, and policies must be learned from limited data. In model-based Batch RL algorithms, transition probabilities are learned from the data as well. Due to insufficient data, these transition probabilities are often imprecise. Errors in transition probabilities can accumulate, resulting in low-performing policies that fail when deployed.

To account for the uncertainty in transition probabilities, prior works use Bayesian models (29; 10) to model uncertainty and optimize the policy to maximize the returns corresponding to the worst α-percentile transition probability model. These policies guarantee that the true expected returns will be at least as large as the optimal returns with high confidence. This technique is commonly referred to as the *percentile-criterion* optimization. Unfortunately, the percentile criterion is NP-hard to optimize. Thus, current work uses *Robust Markov Decision Processes* (RMDPs) to optimize a lower bound on the percentile criterion. An RMDP takes as input an uncertainty set that contains the true transition probability model with high confidence and finds a policy that maximizes the returns of the worst model in the uncertainty set.

Unfortunately, the percentile criterion is non-convex, and therefore, constructing uncertainty sets is a challenging problem. Existing work uses Bayesian credible regions (BCR) (29) as uncertainty sets. However, these uncertainty sets are often unnecessarily large (29; 12) and result in learning conservative robust policies. Some of the recent work approximates uncertainty sets using various heuristics (3; 29), but we show that they remain too conservative. Thus, the question of the *best possible near-optimal set that can be constructed for the percentile criterion* remains unanswered.

Our Contributions In this paper, we answer two important questions: a) *Are Bayesian credible regions the most optimal ambiguity sets for optimizing the percentile criterion?* b) *Can we obtain*

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

a less conservative solution to the percentile criterion without explicitly constructing ambiguity sets? Our theoretical findings show that Bayesian credible regions grow significantly as the number of states increases and therefore, tend to be unnecessarily large, resulting in highly conservative policies. As our main contribution, we provide a dynamic programming framework (Section 3), which we name the VaR framework, for optimizing a lower bound on the percentile criterion without explicitly constructing ambiguity sets. Specifically, we propose a new robust Bellman operator, the Value at Risk (VaR) Bellman operator, for optimizing the percentile criterion. We show that it is a valid contraction mapping that optimizes a tighter lower bound on the percentile criterion, compared to RMDPs with BCR ambiguity sets (Section 3). We theoretically analyze and bound the performance loss of our framework (Section 3.1). We also show that even span-optimized Bayesian credible regions can grow unnecessarily large with the number of states in the MDP and result in a conservative solution, whereas the ambiguity sets implicitly optimized by the VaR Bellman operator tend to be smaller and independent of the number of states (Section 4). Finally, we empirically demonstrate the efficacy of our framework in several domains (Section 5).

1.1 Related Work

Several works propose different methods for solving the percentile criterion, as well as other robust measures for handling uncertainty in the transition model estimates. Russel and Petrik (29) and Behzadian et al. (3) propose various heuristics for minimizing the size of the ambiguity sets constructed for the percentile-criterion. Russel and Petrik (29) propose a method that interleaves robust value iteration with ambiguity set size optimization. Behzadian et al. (3) propose an iterative algorithm that optimizes the weights of ℓ_1 and ℓ_∞ ambiguity sets while optimizing the robust policy. However, these methods still construct Bayesian credible sets which can be unnecessarily large and result in conservative policies, as we show in Section 5.

Other works consider partial correlations between uncertain transition model parameters to mitigate the conservativeness of learned policies (12; 22; 13; 21; 4). These approaches mitigate the conservativeness of S- and SA-rectangular ambiguity sets by capturing correlations between the uncertainty and by limiting the number of times the uncertain parameters deviate from the mean parameters. Despite these heuristics, most of these works (29; 35; 2; 14) either rely on weak statistical concentration bounds to construct frequentist ambiguity sets, or use Bayesian credible regions as ambiguity sets. These sets still tend to be unnecessarily large (29; 12), resulting in conservative policies. Robust RL work (37; 11; 20; 14; 2; 24) proposes other robust measures for handling uncertainty in transition models; however, these approaches do not provide probabilistic guarantees on the expected returns, and compute overly conservative policies.

2 Preliminaries

In the standard reinforcement learning setting, a sequential decision task is modeled as a Markov Decision Process (MDP) (26; 33). An MDP is a tuple $\langle \mathcal{S}, \mathcal{A}, P, r, d_0, \gamma \rangle$ that consists of (a) a set of states $\mathcal{S} = \{1, 2, \ldots, S\}$, (b) a set of actions $\mathcal{A} = \{1, 2, \ldots, A\}$, (c) a deterministic reward function $R\colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, (d) a transition probability function $P\colon \mathcal{S} \times \mathcal{A} \to \Delta^S$, (e) an initial state distribution $\boldsymbol{p}_0 \in \Delta^S$, where Δ^S represents the S-dimensional probability simplex, and (f) a discount factor $\gamma \in [0, 1]$. We use $\boldsymbol{p}_{s,a}$ to denote the vector of transition probabilities $P(s, a, \cdot)$ corresponding to the state s and action a. Similarly, for any reward function R, we use $\boldsymbol{r}_{s,a}$ to denote the vector of rewards $R(s, a, \cdot)$ corresponding to state s and action a. A Markovian policy $\pi\colon \mathcal{S} \to \Delta^A$ maps each state s to a distribution over actions $\mathcal{A}$. In a general RL setting, the goal is to compute a policy π that maximizes the expected discounted return $\rho(\pi, P)$ over an infinite horizon,

$$\max_{\pi \in \Pi} \rho(\pi, P) = \max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \mid s_0 \sim \boldsymbol{p}_0, a_t \sim \pi(s_t), s_{t+1} \sim P(s_t, a_t, \cdot)\right] .$$

The value of a policy π at any state s is the discounted sum of rewards received by an RL agent, if it starts from state s, i.e., $v^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) | s_0 = s, a_t \sim \pi, s_{t+1} \sim P(s_t, a_t, \cdot)\right]$. We assume a *batch reinforcement learning* setting (18) where the reward function is known, but the true transition probabilities $P^\star$ are unknown. Following prior work on robust Bayesian RL (10; 29; 36; 7), we use parametric Bayesian models to represent uncertainty over the true transition probabilities $P^\star$. Given a batch of sample data $\mathcal{D}$, one can derive a posterior distribution over the random variable representing transition probabilities $\tilde{P}$. We use the tilde to indicate that $\tilde{P}$ is a random variable. To

avoid unnecessary computational technicalities, we assume that $\tilde{P}$ is a discrete random variable taking on values $\tilde{P}(\omega), \omega \in \Omega$ for some $\Omega = \{1, \ldots, N\}$ with a distribution f. That is, the random variable $\tilde{P}$ represents a discrete approximation of the true, possibly continuous posterior, as is common in methods like Sample Average Approximation (SAA) (31).

Percentile Criterion The α-percentile criterion is popularly used to derive robust policies under model uncertainty (10). It aims to compute a policy π that maximizes the returns corresponding to the worst α-percentile model:

$$\underset{\pi \in \Pi,\, y \in \mathbb{R}}{\arg\max} \left\{ y \,\middle|\, \Pr_{\tilde{P} \sim f} \left[\rho(\pi, \tilde{P}) \geq y \right] > 1 - \alpha \right\}. \tag{1}$$

The value y lower-bounds the true expected discounted returns with confidence $1 - \alpha$ where $\alpha \in (0, 0.5)$. Optimizing the percentile criterion is equivalent to optimizing the Value at Risk (VaR_α) of expected discounted returns when there exists uncertainty in transition probabilities $\tilde{P}$ and the expected returns function ρ is lower-semicontinuous. The optimization in (1) is equivalent to

$$\max_{\pi \in \Pi} \mathrm{VaR}_\alpha \left[\rho(\pi, \tilde{P}) \right] \quad , \tag{2}$$

where VaR_α of a bounded random variable $\tilde{X}$ with a CDF function $F : \mathbb{R} \to [0, 1]$ is defined as (27)

$$\mathrm{VaR}_\alpha[\tilde{X}] = \inf \left\{ z \in \mathbb{R} | F(z) > \alpha \right\} \quad . \tag{3}$$

A lower value of α in (1) indicates a higher confidence in the returns achieved in expectation. For example, $\mathrm{VaR}_{0.05}[\rho(\pi, \tilde{P})] = x$ indicates that the true returns will be at least equal to the robust returns x for 95% of the transition probability models. When clear from context, we use VaR to denote the Value at Risk at confidence level α. Unfortunately, the optimization problem in (1) is NP-hard to optimize and is usually approximately solved using Robust MDPs.

Robust MDPs Robust MDPs (RMDPs) generalize MDPs to account for uncertainty, or ambiguity, in the transition model. An *ambiguity set* for an RMDP is constructed such that it contains the true model with high confidence. The optimal policy of a Robust MDP $\pi^\star$ maximizes the returns of the worst model in the ambiguity set: $\pi^\star = \arg\max_{\pi \in \Pi} \min_{P \in \mathcal{P}} \rho(\pi, P)$. General RMDPs are NP-hard to solve (35), but they are tractable for broad classes of ambiguity sets. The simplest such type is the SA-rectangular ambiguity set (35; 25), defined as

$$\mathcal{P} = \left\{ P \in (\Delta^S)^{S \times A} \middle| \boldsymbol{p}_{s,a} \in \mathcal{P}_{s,a},\ \forall s \in \mathcal{S},\ \forall a \in \mathcal{A} \right\} \quad ,$$

for a given $\mathcal{P}_{s,a} \subseteq \Delta^{\mathcal{S}}, s \in \mathcal{S}, a \in \mathcal{A}$. SA-rectangular ambiguity sets (29; 3) assume that the transition probabilities corresponding to each state-action pair are independent. Similarly to MDPs, the optimal robust value function $\boldsymbol{v}^\star \in \mathbb{R}^S$ for an SA-rectangular RMDP is the unique fixed point of the robust Bellman optimality operator $\mathcal{T} \colon \mathbb{R}^S \to \mathbb{R}^S$ defined as $(\mathcal{T}\boldsymbol{v})(s) = \max_{a \in \mathcal{A}} \min_{\boldsymbol{p}_{s,a} \in \mathcal{P}_{s,a}} \boldsymbol{p}_{s,a}^\mathsf{T} \left(\boldsymbol{r}_{s,a} + \gamma \cdot \boldsymbol{v} \right)$.

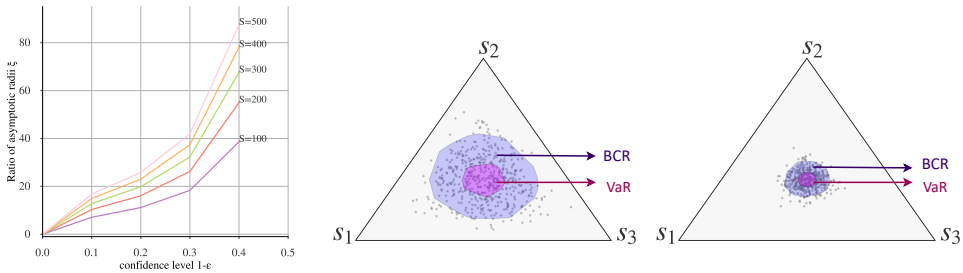
To optimize the percentile criterion, an SA-rectangular ambiguity set $\mathcal{P}$ is constructed such that it contains the true model with high probability, and thus, the following equation holds.

$$\Pr \left[\rho(\pi, \tilde{P}) \geq \min_{P \in \mathcal{P}} \rho(\pi, P) \right] \geq 1 - \alpha \ .$$

Although RMDPs have been used to approximately solve the percentile criterion (3), the quality of the robust policies it computes depends mainly on the size of the ambiguity sets. The larger the ambiguity sets, the more conservative the robust policy (22). SA-rectangular ambiguity sets are most commonly studied; thus we focus our attention on SA-rectangular Robust MDPs. We investigate whether Bayesian credible regions are optimal ambiguity sets for optimizing the percentile criterion. We simply refer to SA-rectangular RMDPs and SA-rectangular ambiguity sets as Robust MDPs and ambiguity sets respectively.

Our work focuses on Bayesian (rather than frequentist) ambiguity sets. Bayesian ambiguity sets are usually constructed from Bayesian credible regions (BCR) (3; 29). Given a state s and an action a, let $\psi_{s,a}$ represent the size of the BCR ambiguity sets; $\mathcal{P}_{s,a}^{\text{BCR}}$ and $\bar{p}_{s,a}$ represent the mean transition model. The set $\mathcal{P}_{s,a}^{\text{BCR}}$ is constructed as

$$\mathcal{P}_{s,a}^{\text{BCR}} = \mathcal{P}_{s,a}(\boldsymbol{w}, \psi, q) = \left\{ \boldsymbol{p} \in \Delta^S \middle| \|\boldsymbol{p}_{s,a} - \tilde{\boldsymbol{p}}_{s,a}\|_{q,\boldsymbol{w}} \leq \psi_{s,a} \right\} \quad , \tag{4}$$



(a) Asymptotic Radii (b) Dirichlet(5,5,5) (c) Dirichlet(30,30,30)

Figure 1: Figure 1a (left) compares the asymptotic radius of BCR ambiguity sets to VaR ambiguity sets. The size of the BCR ambiguity sets significantly grows with the number of states. Figure 1b and Figure 1c (right) compare the asymptotic forms of BCR and VaR ambiguity sets under high and low uncertainty in $\tilde{P}$ at confidence level $\delta = 0.2$ respectively. The grey dots represent the transition probabilities samples from a Dirichlet distribution. The sizes of BCR and VaR ambiguity sets increase with an increase in the variance of the Dirichlet distribution, however, the VaR ambiguity sets are consistently smaller than the BCR ambiguity sets.

where $q \in \{1, \infty\}$ represents the norm of the weighted ball in (4) and $\boldsymbol{w} \in \mathbb{R}^S_+$ is a weight vector. Here, $\boldsymbol{w}$ is jointly optimized with $\psi \in \mathbb{R}$ to minimize the span of the ambiguity sets such that the true model is contained in the ambiguity set with high confidence, i.e., $\Pr\left(\tilde{\boldsymbol{p}}_{s,a} \in \mathcal{P}_{s,a}(\boldsymbol{w}, \psi, q)\right) \geq 1-\alpha$. We refer to BCR ambiguity sets with non-uniform weights as weighted BCR ambiguity sets. We refer to the Robust Bellman optimality operator with BCR ambiguity sets $\mathcal{T}_{\text{BCR}}$ as the BCR Bellman optimality operator, and to RMDPs with BCR ambiguity sets as BCR RMDPs. For any $\delta \in (0, 1)$, setting the confidence level α in $\mathcal{T}_{\text{BCR}}$ to δ/SA for all state-action pairs yields $1 - \delta$ confidence on the returns of the optimal robust policy (3). However, we show that even span-optimized BCR RMDPs are suboptimal for optimizing the percentile criterion.

We use the shorthand $\boldsymbol{w}_{s,a}$ for any $s \in \mathcal{S}$, $a \in \mathcal{A}$ to denote the vector of values associated with value $\boldsymbol{v} \in \mathbb{R}^S$ and the one-step transition from state s and action a, i.e., $\boldsymbol{w}_{s,a} = \boldsymbol{r}_{s,a} + \gamma \boldsymbol{v}$. We use $\bar{\boldsymbol{p}}_{s,a} \in \mathbb{R}^s$ and $\boldsymbol{\Sigma}_{s,a} \in \mathbb{R}^{S \times S}$ for any $s \in \mathcal{S}$, $a \in \mathcal{A}$ to represent the empirical mean and covariance of transition model $\tilde{\boldsymbol{p}}_{s,a}$ estimated from $\mathcal{D}$. We use $\phi(\cdot)$ and $\Phi(\cdot)$ to represent the probability distribution function (PDF) and cumulative distribution function (CDF) respectively of the normal distribution with mean 0 and variance 1. The Z-Minkowski norm $\|\boldsymbol{x}\|_Z$ for a vector x given some positive-definite matrix Z is defined as $\|\boldsymbol{x}\|_Z = \sqrt{\boldsymbol{x}^\intercal Z^{-1} \boldsymbol{x}}$.

Example 2.1. Consider an MDP with four states $\{s_0, s_1, s_2, s_3\}$ and a single action $\{a_0\}$. The state s_0 is the initial state and the states s_1, s_2, s_3 are terminal states with zero rewards. The posterior of the transition probability $\tilde{\boldsymbol{p}}_{s_0,a_0}$ follows a Dirichlet distribution $\mathit{Dir}(10, 10, 1)$ with mean $[0.48, 0.48, 0.04]$. The rewards for transitions from state 0 are given by $\boldsymbol{r}_{s_0,a_0} = [0.25, 0.25, -1]$.

We wish to optimize the percentile criterion with confidence level $\delta = 0.2$. Following the sampling procedure proposed by Russel and Petrik (29) to construct a uniformly weighted BCR ambiguity set for $\tilde{\boldsymbol{p}}_{s_0,a}$ with 100 posterior samples, yields an ambiguity set $\mathcal{P}^{\text{BCR}}_{s_0,a} = \left\{\boldsymbol{p} \in \Delta^S \middle| \|\boldsymbol{p} - \tilde{\boldsymbol{p}}_{s_0,a}\|_1 \leq 0.277\right\}$. In this case, the reward estimate against the worst model in the ambiguity set $p = [0.50, 0.32, 0.18]$ is $\rho^{\text{BCR}} = 0.025$. Since we have a single non-terminating state in the MDP, the percentile returns are given by $\rho^{\mathrm{VaR}_\alpha} = \mathrm{VaR}_{0.2}[\tilde{\boldsymbol{p}}^\intercal_{s_0,a} \boldsymbol{r}_{s_0,a}]$. Computing $\rho^{\mathrm{VaR}_\alpha}$ for Dirichlet distribution $\mathit{Dir}(10, 10, 1)$, we get $\rho^{\mathrm{VaR}_\alpha} = 0.17 > \rho^{\text{BCR}}$. Thus, this example shows that BCR ambiguity sets can be unnecessarily large and, thus, result in conservative policies.

3 VaR Framework

We introduce the VaR Bellman optimality operator $\mathcal{T}_{\mathrm{VaR}}$ for approximately solving the percentile criterion. We show that $\mathcal{T}_{\mathrm{VaR}}$ is a valid Bellman operator: it is a contraction mapping and lower bounds the percentile criterion. For any value function $\boldsymbol{v} \in \mathbb{R}^{\mathcal{S}}$, state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, we

define the VaR Bellman optimality operator $\mathcal{T}_{\mathrm{VaR}}$ as

$$(\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v})(s) = \max_{a\in\mathcal{A}} \mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}\right] \quad . \tag{5}$$

For each state s, $\mathcal{T}_{\mathrm{VaR}}$ maximizes the value corresponding to the worst α-percentile model. In contrast to the BCR Bellman optimality operator $\mathcal{T}_{\mathrm{BCR}}$, computing $\mathcal{T}_{\mathrm{VaR}}$ does not require constructing ambiguity sets from confidence regions; it can simply be estimated from samples of the model posterior distribution, as we later show.

In Proposition A.4, we formally prove that $\mathcal{T}_{\mathrm{VaR}}$ is a contraction mapping, and thus has a unique fixed point. This fixed point is the value of the optimal policy that maximizes a tight lower bound on the percentile criterion.

We now show that the VaR Bellman optimality operator $\mathcal{T}_{\mathrm{VaR}}$ optimizes a lower bound on the percentile criterion. Given a policy π, a state $s \in \mathcal{S}$, and a transition model $\tilde{P}$, let

$$(\mathcal{T}^\pi\boldsymbol{v})(s) = \sum_{a\in\mathcal{A}} \pi(s,a)\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a} \text{ and } (\mathcal{T}_{\mathrm{VaR}}^\pi\boldsymbol{v})(s) = \sum_{a\in\mathcal{A}} \pi(s,a)\,\mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}\right] \quad ,$$

represent the Bellman evaluation operator and the VaR Bellman evaluation operator, respectively. Finally, let $\hat{\boldsymbol{v}}^\pi$ and $\boldsymbol{v}^\pi$ denote the fixed point of $\mathcal{T}_{\mathrm{VaR}}^\pi$ and $\mathcal{T}^\pi$ respectively. We use $\tilde{\boldsymbol{v}}^\pi$ to represent the random value function of policy π computed using a sample from the posterior distribution of $\tilde{P}$.

Proposition 3.1 (Lower Bound Percentile Criterion)**.** *For any $\delta \in (0,1)$, if we set the confidence level α in the operator $\mathcal{T}_{\mathrm{VaR}}^\pi$ to δ/S, then for every policy $\pi \in \Pi : \Pr_{\tilde{P}}\left[\hat{\boldsymbol{v}}^\pi \preceq \tilde{\boldsymbol{v}}^\pi | \mathcal{D}\right] \geq 1-\delta$, where $\tilde{P}$ is a realization of the posterior distribution of the transition model P conditioned on observed transitions $\mathcal{D}$.*

See Appendix B.1 for the proof. Proposition 3.1 shows that for any policy π and state s, the VaR value at state s, $\hat{\boldsymbol{v}}^\pi(s)$ lower bounds the true value $\tilde{\boldsymbol{v}}^\pi(s)$ with high confidence. Comparing Proposition 3.1 with the definition of the percentile-criterion in (1), it is easy to see that the percentile-criterion requires confidence guarantees only on the returns computed from the initial states, whereas the equation in Proposition 3.1 provides confidence guarantees on the value of every state. Therefore, for any policy π, the value $\boldsymbol{p}_0^\mathsf{T}\hat{\boldsymbol{v}}^\pi$ is a lower bound on the percentile-criterion objective $\mathrm{VaR}_\delta[\rho(\pi,\tilde{P})]$. Since $\mathcal{T}_{\mathrm{VaR}}$ finds a policy π that maximizes the value $\boldsymbol{p}_0^\mathsf{T}\hat{\boldsymbol{v}}^\pi$, it follows (26) that $\mathcal{T}_{\mathrm{VaR}}$ optimizes a lower bound on the percentile criterion in (1).

Proposition 3.2. *Suppose that $\tilde{\boldsymbol{p}}_{s,a}$ for any state s and action a, is a multivariate sub-Gaussian with mean $\bar{\boldsymbol{p}}_{s,a}$ and covariance factor $\boldsymbol{\Sigma}_{s,a}$, i.e., $\mathbb{E}\left[\exp\left(\lambda(\tilde{\boldsymbol{p}}_{\boldsymbol{s},\boldsymbol{a}} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\boldsymbol{w}\right)\right] \leq \exp\left(\lambda^2\boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}/2\right), \forall\lambda\in\mathbb{R}, \forall\boldsymbol{w}\in\mathbb{R}^S$. Then, for any state $s\in\mathcal{S}$, $\mathcal{T}_{\mathrm{VaR}}$ satisfies*

$$(\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v})(s) \geq \max_{a\in\mathcal{A}}\left(\bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a} - \sqrt{2\ln(1/\alpha)}\sqrt{\boldsymbol{w}_{s,a}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}\right) \quad .$$

See Appendix B.2 for the proof. Proposition 3.2 shows that by assuming that the transition probabilities are sub-Gaussian, we can easily compute a lower bound of the VaR Bellman update $(\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v})$ for a given value function using only the mean and the covariance matrix of $\tilde{P}$. In Proposition A.5, we provide the exact form of the VaR_α Bellman optimality operator $\mathcal{T}_{\mathrm{VaR}}$ under normal conditions, which is a special case of Proposition 3.2.

3.1 Performance Guarantees

We now derive finite-sample and asymptotic bounds on the loss of the VaR framework.

Theorem 3.3 (Performance)**.** *Let $\hat{\boldsymbol{v}}$ be the fixed point of the* VaR *Bellman optimality operator $\mathcal{T}_{\mathrm{VaR}}$, and $\pi^\star$ be the optimal policy in* (1)*. Let $\rho^\star = \mathrm{VaR}_\alpha\left[\rho(\pi^\star,\tilde{P})\right]$ denote the optimal percentile returns and $\hat{\rho} = \boldsymbol{p}_0^\mathsf{T}\hat{\boldsymbol{v}}$ denote the lower bound on the percentile returns computed using the Bellman operator $\mathcal{T}_{\mathrm{VaR}}$. For any $\delta\in(0,1)$, we set the confidence level $\alpha = \delta/(2SA)$ in $\mathcal{T}_{\mathrm{VaR}}$. Then, with probability at least $1-\delta$, the performance loss with respect to $\rho^\star$ is*

$$\rho^\star - \hat{\rho} \leq \frac{1}{1-\gamma}\max_{s\in\mathcal{S}}\max_{a\in\mathcal{A}}(\mathrm{VaR}_{1-\alpha}[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\hat{\boldsymbol{w}}_{s,a}] - \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\hat{\boldsymbol{w}}_{s,a}]) \quad . \tag{6}$$

See Appendix B.3 for the proof. Theorem 3.3 bounds the finite sample performance loss of the VaR framework. The loss varies proportionally to the maximum difference between the α and $1-\alpha$ percentile of the one-step Bellman update for the optimal robust value function $\hat{\boldsymbol{v}}$. As expected, the VaR framework performs better when the uncertainty in the transition models is small.

Theorem 3.4 (Asymptotic Performance)**.** *For any $\delta \in (0,1)$, set $\alpha = \delta/(2SA)$ in $\mathcal{T}_{\mathrm{VaR}}$. Let $I(\boldsymbol{p}^\star_{s,a})^{-1}$ for any state s and action a, be the Fisher Information matrix corresponding to the true transition probabilities $\boldsymbol{p}^\star_{s,a}$. Furthermore, let $\sigma_{\max} = \max_{s\in\mathcal{S},a\in\mathcal{A}} \sqrt{\hat{\boldsymbol{w}}^\mathsf{T}_{s,a} I(\boldsymbol{p}^\star_{s,a})^{-1}\hat{\boldsymbol{w}}_{s,a}}$ represent the maximum asymptotic standard deviation of the returns estimate $\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\hat{\boldsymbol{w}}_{s,a}$ for any state-action pair (s,a). Then, with probability at least $1-\delta$, the asymptotic performance of the* VaR *framework $\hat{\rho}$ w.r.t. the optimal percentile returns $\rho^\star$ satisfies*

$$\lim_{N\to\infty} \sqrt{N}(\rho^\star - \hat{\rho}) \leq \frac{1}{1-\gamma}\left(2\Phi^{-1}(1-\alpha)\sigma_{\max}\right) \leq \frac{1}{1-\gamma}\sqrt{8\ln(1/\alpha)}\sigma_{\max} \ .$$

See Appendix B.4 for the proof. Theorem 3.4 shows that almost surely, the asymptotic loss in performance of the VaR framework convergence to 0, i.e., $\lim_{N\to\infty}(\rho^\star - \hat{\rho}) = 0$.

3.2 Dynamic Programming Algorithm

We provide a detailed description of the VaR value iteration algorithm (Algorithm 3.1) below. We also bound the number of samples required to estimate the VaR Bellman update $(\mathcal{T}^\pi_{\mathrm{VaR}}\boldsymbol{v})(s)$ for any given policy π and state s with high confidence $1-\zeta$.

Algorithm 3.1: Generalized VaR Value Iteration Algorithm

Input: Confidence α, Posterior distribution f, target Bellman residual ε
Output: Robust policy π, lower bound $\boldsymbol{v}$

1 Initialize π with arbitrary π_0, robust value-function $\boldsymbol{v}$ with arbitrary $\boldsymbol{v}_0, k=0$;
2 Sample N models $\tilde{P}(\omega_1), \tilde{P}(\omega_2), \ldots, \tilde{P}(\omega_N)$ from posterior f ;
3 **repeat**
4 **for** $s \leftarrow 1$ **to** S **do**
5 Initialize $\boldsymbol{q} \leftarrow []$;
6 **for** $a \leftarrow 1$ **to** A **do**
7 $\boldsymbol{q}[a] \leftarrow \widehat{\mathrm{VaR}}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}(\boldsymbol{r}_{s,a} + \gamma\boldsymbol{v}_k)]$ OR ;
8 $\boldsymbol{q}[a] \leftarrow \bar{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a} - \Phi^{-1}(1-\alpha)\sqrt{\boldsymbol{w}^\mathsf{T}_{s,a}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}$ (under normal Assumptions);
9 **end**
10 $\boldsymbol{v}_k(s) \leftarrow \max(\boldsymbol{q})$; $\quad \pi_k(s) \leftarrow \arg\max(\boldsymbol{q})$;
11 **end**
12 $k \leftarrow k+1$;
13 **until** $\|\boldsymbol{v}_k - \boldsymbol{v}_{k-1}\|_\infty \leq \varepsilon$;
14 **return** $\pi_k, \boldsymbol{v}_k$;

In each iteration of Algorithm 3.1, we compute the one-step VaR Bellman update $\mathcal{T}_{\mathrm{VaR}}(\boldsymbol{v})$ using the current value function $\boldsymbol{v}$. When $\tilde{P}$ is not normally distributed, we use the Quick Select algorithm (16) to efficiently compute the empirical estimate of the α-percentile of returns for any state s and action a, i.e., $\widehat{\mathrm{VaR}}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}(\boldsymbol{r}_{s,a} + \gamma\boldsymbol{v})]$ in $\mathcal{O}(SAN)$ time (Proposition 3.5). On the other hand, when $\tilde{P}$ is normally distributed, we compute the VaR Bellman update $\mathcal{T}_{\mathrm{VaR}}(v)$ using the empirical estimate of mean $(\bar{\boldsymbol{p}}_{s,a})_{s\in\mathcal{S},a\in\mathcal{A}}$ and covariance $(\boldsymbol{\Sigma}_{s,a})_{s\in\mathcal{S},a\in\mathcal{A}}$ of the transition probabilities derived from the data $\mathcal{D}$ (Proposition A.5). We repeat these steps until convergence.

Proposition 3.5 (Time Complexity)**.** *The time complexity of a single iteration of the loop in line 3 of the* VaR *Value Iteration (Algorithm 3.1) is in $\mathcal{O}(SAN)$, where N is the number of samples of the posterior samples of $\tilde{P}$.*

Proposition 3.6 (Empirical Error Bound)**.** *For any state s, action a and value function $\boldsymbol{v}$, let $\widehat{\mathrm{VaR}}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a}]$ represent the empirical estimate of α-percentile of returns $\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a}]$ and Φ_f represent the cumulative density function (CDF) of the random estimate of returns $\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a}$. Suppose that Φ_f is differentiable at the point $\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a}]$ and let $m = \Phi'(\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a}])$*

represents the density of estimate of returns at point $\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}]$. *Let* $N^\star$ *be the number of posterior samples required to obtain empirical error* $\varepsilon \in \mathbb{R}$, *with confidence* $1-\zeta$, *i.e.,* $\Pr\left[|\widehat{\mathrm{VaR}}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}] - \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}]| > \varepsilon\right] \le \zeta$. *Then,* $\lim_{\varepsilon\to 0} N^\star\varepsilon^2 = \ln(2/\zeta)/2m^2$.

4 Comparison with Bayesian Credible Regions

We are now ready to answer the question: *Are Bayesian credible regions the optimal ambiguity sets for optimizing the percentile criterion?* For this, we compare the VaR framework with BCR Robust MDPs. First, we derive the robust form of the VaR framework and show that in contrast to the BCR Bellman operator, the VaR Bellman optimality operator implicitly constructs value function dependent ambiguity sets and thus, these sets tend to be smaller (Proposition 4.1). Then, we show that the solution given by the VaR operator $\mathcal{T}_{\mathrm{VaR}}$ is never worse than the solution given by the BCR operator $\mathcal{T}_{\mathrm{BCR}}$ (Proposition 4.2). Finally, we compare the asymptotic sizes of the BCR ambiguity sets and the VaR ambiguity sets implicitly constructed by $\mathcal{T}_{\mathrm{VaR}}$. For any given confidence level α, the VaR ambiguity sets are asymptotically smaller than BCR ambiguity sets (Theorem 4.4), which grow with the number of states. Precisely, the ratio of the sizes of VaR ambiguity sets to BCR ambiguity sets is at least $\sqrt{\chi^2_{S,1-\alpha}}/\Phi^{-1}(1-\alpha)$, where $\chi^2_{S,1-\alpha}$ is the CDF inverse of $1-\alpha$ percentile of Chi-squared distribution with degree of freedom S and $\Phi^{-1}(1-\alpha)$ is the $1-\alpha$ percentile of $\mathcal{N}(0,1)$. Thus, we conclude that VaR framework is better suited for optimizing the percentile criterion than BCR RMDPs.

For any policy π, define the VaR ambiguity set $\mathcal{P}^{\mathrm{VaR},\boldsymbol{v}^\pi}$ as

$$\mathcal{P}^{\mathrm{VaR},\boldsymbol{v}^\pi} = \bigtimes_{s\in\mathcal{S},a\in\mathcal{A}} \mathcal{P}_{s,a}^{\mathrm{VaR},\boldsymbol{v}^\pi} \text{ where } \mathcal{P}_{s,a}^{\mathrm{VaR},\boldsymbol{v}^\pi} = \left\{\boldsymbol{p}_{s,a}\in\Delta^{\mathcal{S}} \middle| \boldsymbol{p}_{s,a}^\mathsf{T}\boldsymbol{v}^\pi \ge \mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{v}^\pi\right]\right\} \ . \tag{7}$$

Proposition 4.1 (Equivalence)**.** *Let* $\hat{\boldsymbol{v}}$ *be the fixed point of the* VaR *Bellman optimality operator* $\mathcal{T}_{\mathrm{VaR}}$, *i.e,* $\hat{\boldsymbol{v}} = (\mathcal{T}_{\mathrm{VaR}}^\pi\hat{\boldsymbol{v}})$. *Let* Π^D *be the set of all deterministic policies. Then,* $\hat{\boldsymbol{v}}$ *is the value function of policy* $\hat{\pi}$ *where* $\hat{\pi}$ *is an optimal solution of the following problem,*

$$\arg\max_{\pi\in\Pi^D} \min_{P\in\mathcal{P}^{\mathrm{VaR},\boldsymbol{v}^\pi}} \rho(\pi,P) \ . \tag{8}$$

See Appendix B.7 for the proof. Proposition 4.1 shows that the VaR Bellman optimality operator optimizes a unique robust MDP whose ambiguity sets are SA-rectangular and policy dependent. Notice that for any state s and action a, the ambiguity set is a half-plane $\left\{\boldsymbol{p}_{s,a}\in\mathbb{R}^S : \boldsymbol{p}_{s,a}^\mathsf{T}\boldsymbol{v}^\pi \le \mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{v}^\pi\right]\right\}$ dependent on the value function $\boldsymbol{v}^\pi$ of the current policy π. In contrast, BCR ambiguity sets are independent of any policy or value function and are constructed such that it provides high-confidence guarantees on returns of all policies simultaneously. As a result, BCR ambiguity sets tend to be unnecessarily large.

Proposition 4.2. *For any policy* π, *the fixed point of the* VaR *policy evaluation operator* $\mathcal{T}_{\mathrm{VaR}}^\pi$ *dominates the fixed point of the Bellman evaluation operator* $\mathcal{T}_{BCR}^\pi$, *i.e.,* $\mathcal{T}_{BCR}^\pi\cdots\mathcal{T}_{BCR}^\pi\boldsymbol{v} \preceq \mathcal{T}_{\mathrm{VaR}}^\pi\cdots\mathcal{T}_{\mathrm{VaR}}^\pi\boldsymbol{v}$ *for any* $\boldsymbol{v}$. *Similar results hold for policy optimization operators* $\mathcal{T}_{\mathrm{VaR}}$ *and* $\mathcal{T}_{BCR}$, *i.e.,* $\mathcal{T}_{BCR}\cdots\mathcal{T}_{BCR}\boldsymbol{v} \preceq \mathcal{T}_{\mathrm{VaR}}\cdots\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v}$ *for any* $\boldsymbol{v}$.

Proposition 4.2 proves that the lower bound on the percentile-criterion optimized by the VaR Bellman operator $\mathcal{T}_{\mathrm{VaR}}$ is never worse than the lower-bound optimized by the BCR Bellman operator $\mathcal{T}_{\mathrm{BCR}}$. Although this is a weaker claim, in the results that follow, we show that asymptotically, the BCR ambiguity sets tend to grow unnecessarily large with the number of states, and therefore, BCR RMDPs are more susceptible to generating overly conservative solutions.

We now compute the asymptotic sizes of BCR ambiguity sets and VaR ambiguity sets.

The Bernstein Von-Mises Theorem (34) establishes the asymptotic properties of the posterior distribution of $\tilde{P}$ constructed from N independent samples. We assume that this theorem holds for transition probabilities $\tilde{P}$ for computing the asymptotic sizes of the VaR and BCR ambiguity sets.

Theorem 4.3 (Asymptotic Size of VaR Ambiguity Sets)**.** *Let* $\bar{P} = (\bar{\boldsymbol{p}}_{s,a})_{s\in\mathcal{S},a\in\mathcal{A}}$ *be the Maximum Likelihood estimate of transition probabilities computed from data* $\mathcal{D}$ *and* $\Sigma = (I(\boldsymbol{p}_{s,a}^\star)^{-1})_{s\in\mathcal{S},a\in\mathcal{A}}$ *be the corresponding covariance matrix. Then, as* $N\to\infty$, $\forall s\in\mathcal{S}, a\in\mathcal{A}$,

$$\mathcal{P}_{s,a}^{\mathrm{VaR}} = \left\{\boldsymbol{p}_{s,a}\in\Delta^{\mathcal{S}} \middle| \frac{1}{\sqrt{N}}\|\boldsymbol{p}_{s,a}-\bar{\boldsymbol{p}}_{s,a}\|_{\boldsymbol{\Sigma}_{s,a}^{-1}} \le \frac{\Phi^{-1}(1-\alpha)}{\sqrt{N}}\right\} \ . \tag{9}$$

Theorem 4.3 shows that the asymptotic form of the VaR ambiguity set is an ℓ_2 ellipsoid ball with radius $\frac{\Phi^{-1}(1-\alpha)}{\sqrt{N}}$. It is important to notice that, in contrast to the finite-sample VaR ambiguity set $\mathcal{P}_{s,a}^{\mathrm{VaR},\boldsymbol{v}^\pi}$ in problem (7), the asymptotic VaR ambiguity set $\mathcal{P}_{s,a}^{\mathrm{VaR}}$ is independent of the value function $\boldsymbol{v}$. This is because $\mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{v}^\pi\right]$ is convex in the value function $\boldsymbol{v}$, (28) and as a consequence, the asymptotic ambiguity set $\mathcal{P}_{s,a}^{\mathrm{VaR}}$ is simply the intersection of closed half-planes in $\mathcal{P}_{s,a}^{\mathrm{VaR},\boldsymbol{v}^\pi}$, computed over all value functions $\boldsymbol{v} \in \mathbb{R}^{\mathcal{S}}$ (8). It is also worth noting that the radius of the asymptotic VaR ambiguity set $\mathcal{P}_{s,a}^{\mathrm{VaR}}$ is a constant. In contrast, the asymptotic radius of the BCR ambiguity sets grows with the number of states, as we show in the following proposition.

Theorem 4.4 (Asymptotic size of Bayesian Credible Regions)**.** *For any state s and action a, let $\mathcal{P}_{s,a}^{BCR}$ be any Bayesian credible region and $\bar{\boldsymbol{p}}_{s,a}$ be the maximum likelihood estimator based on data $\mathcal{D}$. Furthermore, $\xi < \sqrt{\chi^2_{S,1-\alpha}}/\Phi^{-1}(1-\alpha)$. Then, as $N \to \infty$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$,*

$$\mathcal{P}_{s,a}^{BCR} - \bar{\boldsymbol{p}}_{s,a} \not\subseteq \xi(\mathcal{P}_{s,a}^{\mathrm{VaR}} - \bar{\boldsymbol{p}}_{s,a}) \ . \tag{10}$$

See Appendix B.10 for the proof. Theorem 4.4 is surprising and crucial as it proves that Bayesian credible regions can grow very large very fast with the number of states, thereby rendering them suboptimal for optimizing the percentile criterion. Figure 1a shows the growth in the ratio of BCR to VaR ambiguity sets with an increasing number of states, and Figure 1b and Figure 1c show the asymptotic forms of BCR and VaR ambiguity sets under high and low uncertainty in transition probabilities $\tilde{P}$.

5 Experiments

We now empirically analyze the robustness of the VaR framework in three different domains. Specifically, we investigate if the VaR_α framework learns robust policies that are less conservative than BCR Robust MDPs.

Riverswim: The Riverswim MDP (32) consists of five states and two actions. The state represents the coordinates of the swimmer in the river and action represents the direction of the swim. The task of the agent is to learn a policy that would take the swimmer to the other end of the river.

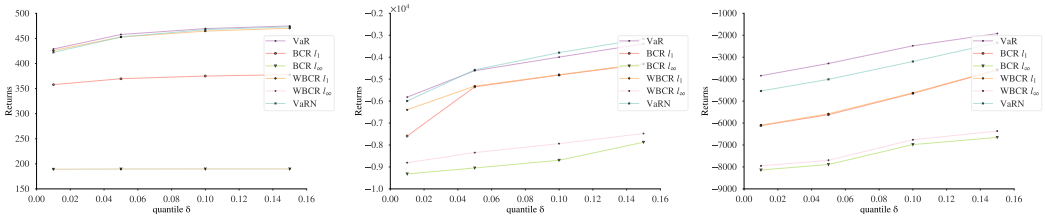
Population Growth Model: The Population Growth MDP (17) models the population growth of pests and consists of 50 states and 5 actions. The states represent the pest population and actions represent the pest control measures. In our experiments, we use two different instantiations of the Population Growth Model: Population-Small and Population, which vary in the number of posterior samples.

Inventory Management: The Inventory Management MDP (38) models the classical inventory management problem and consists of 30 states and 30 actions. States represent the inventory level and actions represent the inventory to be purchased. The sale price s, holding cost c and purchase costs p are 3.99, 0.03, and 2.219. The demand is normally distributed with mean=$s/4$ and standard deviation $s/6$.

Implementation details: For each domain in our experiments, we sample a dataset $\mathcal{D}$ consisting of n tuples of the form $\{s, a, r, s'\}$, corresponding to the state s, the action taken a, the reward r and the next state s'. We construct a posterior distribution over the models using $\mathcal{D}$, assuming Dirichlet priors over the model parameters. We obtain a set of M train and K test models using MCMC sampling. We use the train models to learn robust policies in the VaR framework and other baselines for a range of values of confidence level δ. We consider two instances of the VaR framework: one (denoted by *VaRN*) that assumes that $\tilde{P}$ is a multivariate normal, and another (denoted by *VaR*) that does not assume any structure over $\tilde{P}$. We compare the robust performance of the VaR framework with that of four baselines: BCR Robust MDPs with weighted ℓ_1 ambiguity sets (*WBCR* ℓ_1), weighted ℓ_∞ ambiguity sets (*WBCR* ℓ_∞), unweighted ℓ_1 ambiguity sets (*BCR* ℓ_1) and unweighted ℓ_∞ ambiguity sets(*BCR* ℓ_1) ambiguity sets. See Appendix C in the appendix for more details.

To evaluate the VaR framework, we use the robust returns of the optimal policy, i.e., the δ-percentile returns of the optimal policy on the test model dataset. We compare the robust returns of the VaR framework with that of the baselines, for different values of δ.

Experimental Results Figure 2 shows how the robust performance of the VaR framework and the baselines varies for different values of confidence level $(1-\delta)$. As expected, the robust performance



(a) Inventory Domain (b) Population Domain (c) Population-Small Domain

Figure 2: Comparison of robust returns achieved by *VaR*, *VaRN*, *BCR* ℓ_1, *BCR* ℓ_∞, *WBCR* ℓ_1 and *WBCR* ℓ_∞ for different confidence levels δ in Inventory and Population and Population-Small domain.

Domain	*VaR*	*VaRN*	*BCR* (ℓ_1)	*BCR* (ℓ_∞)	*WBCR* (ℓ_1)	*WBCR* (ℓ_∞)
Riverswim	**67.27**	**67.27**	67.27	67.27	67.27	67.27
Population	**-4614.89**	**-4576.65**	-5349.45	-9046.94	-5315.84	-8346.4
Population-Small	**-3284.43**	**-4008.32**	-5620.51	-7889.71	-5575.77	-7695.88
Inventory	**457.83**	452.78	369.67	189.65	452.78	189.65

Table 1: The robust returns at $\delta = 0.05$ achieved by *VaR*, *VaRN*, *BCR* ℓ_1, *BCR* ℓ_∞, *WBCR* ℓ_1 and *WBCR* ℓ_∞ in Riverswim, Inventory, Population-Small, and Population domain.

of the *VaR* is higher for smaller values of confidence level $(1 - \delta)$, but more importantly, the robust performance of *VaR* and *VaRN* dominates the robust performance of the baselines in most domains. We also observe that all the baselines and the VaR framework have the same robust performance in the Riverswim domain. We conjecture that this is due to the simplicity of the domain. Furthermore, as expected, we find that the robust performance of BCR Robust MDPs with span-optimized (weighted) ambiguity sets (*WBCR* ℓ_1, WBCR ℓ_∞) is relatively higher than the robust performance of Robust MDPs with unweighted BCR ambiguity sets (*BCR* ℓ_1, BCR ℓ_∞). However, we find that even Robust MDPs with span-optimized BCR ambiguity sets are unable to outperform the robust performance of our VaR_α framework.

Table 1 summarizes the performance of the VaR framework and the baselines for confidence level $\delta = 0.05$ (Table 2 in the appendix summarizes the results for $\delta = 0.1$). The VaR framework consistently outperforms the baselines in terms of robust performance in all domains for $\delta = 0.05$. Figure 3 compares the robust performance of the VaR framework and the baselines on both train and test models. The trends in the robust performance of the VaR framework and the baselines are similar on both train and test models.

6 Limitations and Conclusion

The main limitation of the VaR framework is that it does not consider the correlations in the uncertainty of transition probabilities across states and actions (22; 13; 21). However, due to the non-convex nature of the percentile-criterion (29; 3), constructing a tractable VaR_α Bellman operator that considers these correlations is not feasible. One plausible solution is to use a Conditional Value at Risk Bellman operator (19; 9).which is convex and lower bounds the Value at Risk measure. We leave the analysis of this approach for future work. Empirical analysis of the VaR_α framework in domains with continuous state-action spaces is also an interesting avenue for future work. It would be valuable to rigorously test the VaR_α framework in large RL domains and compare its performance against other Robust RL methods (11; 13; 22).

In conclusion, we propose a novel dynamic programming algorithm that optimizes a tight lower-bound approximation on the percentile criterion without explicitly constructing ambiguity sets. We theoretically show that our algorithm implicitly constructs tight uncertainty sets that are smaller in size than any optimized Bayesian credible region, and thus computes less conservative policies with the same confidence guarantees on returns. We also derive finite-sample and asymptotic bounds on the performance loss due to our approximation. Finally, our experimental results demonstrate the efficacy of our method in several domains.

References

[1] Praveen Agarwal, Mohamed Jleli, and Bessem Samet. *Banach contraction principle and applications*, pages 1–23. Springer Singapore, Singapore, 2018.

[2] Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 511–520. PMLR, 18–24 Jul 2021.

[3] Bahram Behzadian, Reazul Hasan Russel, Marek Petrik, and Chin Pang Ho. Optimizing percentile criterion using robust MDPs. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1009–1017, 2021.

[4] Bahram Behzadian, Marek Petrik, and Chin Pang Ho. Fast algorithms for L_∞-constrained s-rectangular robust MDPs. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[5] Dimitri P. Bertsekas. *Abstract dynamic programming*. Athena Scientific, 2013. ISBN 978-1-886529-42-7.

[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 02 2013.

[7] G. P. Box and G. C. Tiao. Bayesian inference in statistical analysis. *International Statistical Review*, 43:242, 1973.

[8] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[9] Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *International Conference on Neural Information Processing Systems*, pages 3509–3517. MIT Press, 2014.

[10] E. Delage and S. Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010.

[11] Esther Derman, Daniel Mankowitz, Timothy A Mann, and Shie Mannor. Soft-robust actor-critic policy-gradient. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[12] Esther Derman, Daniel J. Mankowitz, Timothy A. Mann, and Shie Mannor. A Bayesian approach to robust reinforcement learning. In *Proceedings of the 35th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

[13] Vineet Goyal and Julien Grand-Clement. Robust Markov decision process: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

[14] Julien Grand-Clement and Christian Kroer. First-order methods for Wasserstein distributionally robust MDP. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 2010–2019, 2021.

[15] Vishal Gupta. Near-optimal Bayesian ambiguity sets for distributionally robust optimization. *Management Science*, 56(9), 2019.

[16] C. A. R. Hoare. Algorithm 65: Find. *Communications of the ACM*, 4(7):321–322, 1961.

[17] Marc Kery and Michael Schaub. *Bayesian Population Analysis Using WinBUGS*. Elsevier, 2012.

[18] Sasche Lange, Thomas Gabel, and Martin Riedmiller. *Batch reinforcement learning*. Springer, 2012.

[19] Elita A. Lobo, Mohammad Ghavamzadeh, and Marek Petrik. Soft-robust algorithms for batch reinforcement learning, 2021.

[20] Daniel J. Mankowitz, Nir Levine, Rae Ung Jeong, Abbas Abdolmaleki, Jost Tobias Springenberg, Timothy Mann, Todd Hester, and Martin A. Riedmiller. Robust reinforcement learning for continuous control with model misspecification. In *Proceedings of the 2020 International Conference on Learning Representations (ICLR)*, 2020.

[21] Shie Mannor, Ofir Mebel, and Huan Xu. Lightning does not strike twice: Robust MDPs with coupled uncertainty. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 451–458, 2012.

[22] Shie Mannor, Ofir Mebel, and Huan Xu. Robust MDPs with K-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

[23] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18:1269 – 1283, 1990.

[24] Ariel Neufeld and Julian Sester. Robust q-learning algorithm for Markov decision processes under wasserstein uncertainty, 2023.

[25] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[26] Martin L. Puterman. *Markov Decision Processes: Discrete stochastic dynamic programming*. John Wiley and Sons, Inc., 2nd edition, 2005.

[27] R. Tyrrell Rockafellar. *Convex analysis*. Princeton University Press, 1970.

[28] R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–41, 2000.

[29] Reazul Hasan Russel and Marek Petrik. Beyond confidence regions: Tight Bayesian ambiguity sets for robust MDPs. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[30] Sergey Sarykalin, Gaia Serraino, and Stan Uryasev. *Value-at-risk vs. conditional value-at-risk in risk management and optimization*, chapter 13, pages 270–294. INFORMS, 2008.

[31] A. Shapiro, W. Tekaya, J. P. da Costa, and M. P. Soares. Risk-neutral and risk-averse stochastic dual dynamic programming method. *European Journal of Operational Research*, pages 375–391, 2013.

[32] Alexander L. Strehl and Michael L. Littman. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 856–863. Association for Computing Machinery, 2005. ISBN 1595931805.

[33] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An introduction*. A Bradford Book, 2018.

[34] A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[35] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[36] Huan Xu and Shie Mannor. The robustness-performance trade-off in Markov decision processes. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2006.

[37] Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 37(2):288–300, 2012.

[38] Paul Herbert Zipkin. *Foundations of inventory management*. McGraw-Hill, Boston, 2000.

A Additional theoretical results

Definition A.1 (Subvariance)**.** For any function $\boldsymbol{v} : \mathcal{S} \to \mathbb{R}$, a scalar c and $\gamma \in [0, 1)$, the operator $\mathfrak{T}$ satisfies the Translation subvariance property if

$$\mathfrak{T}(\boldsymbol{v} + c\mathbf{1}) = (\mathfrak{T}\boldsymbol{v}) + \gamma c\mathbf{1} \quad .$$

Definition A.2 (Monotonicity)**.** For any function $\boldsymbol{v} : \mathcal{S} \to \mathbb{R}$ and $\boldsymbol{u} : \mathcal{S} \to \mathbb{R}$, such that , $\boldsymbol{v} \preceq \boldsymbol{u}$ the operator $\mathfrak{T}$ satisfies the Monotonicity property if

$$\mathfrak{T}\boldsymbol{v} \preceq \mathfrak{T}\boldsymbol{u} \quad .$$

Lemma A.3 (Contraction Mapping (5))**.** *For any two bounded functions $\boldsymbol{u} : \mathcal{S} \to \mathbb{R}$, $\boldsymbol{v} : \mathcal{S} \to \mathbb{R}$, and $\gamma \in [0, 1)$, the operator $\mathfrak{T}$ is a contraction mapping if it satisfies Monotonicity and Translation subvariance properties. In particular, for all $\boldsymbol{u}, \boldsymbol{v}$ there holds*

$$\|\mathfrak{T}\boldsymbol{u} - \mathfrak{T}\boldsymbol{v}\|_\infty \leq \gamma \|\boldsymbol{u} - \boldsymbol{v}\|_\infty \ .$$

Furthermore, for any two bounded functions $\boldsymbol{u} : \mathcal{S} \to \mathbb{R}$, and $\boldsymbol{v} : \mathcal{S} \to \mathbb{R}$, the operator $\mathfrak{T}$ is a non-expansive mapping if it satisfies Monotonicity and Translation invariance properties. In particular, for all $\boldsymbol{u}, \boldsymbol{v}$ there holds

$$\|\mathfrak{T}\boldsymbol{u} - \mathfrak{T}\boldsymbol{v}\|_\infty \leq \|\boldsymbol{u} - \boldsymbol{v}\|_\infty \ .$$

The proof of Lemma A.3 follows directly from Proposition 2.1.3 in (5). We re-derive the proof for the sake of completeness.

Proof. Denote

$$c = \max_{s \in S} |\boldsymbol{u}(s) - \boldsymbol{v}(s)| \ .$$

Therefore for all $s \in \mathcal{S}$,

$$\boldsymbol{u}(s) - c \leq \boldsymbol{v}(s) \leq \boldsymbol{u}(s) + c \ .$$

Applying $\mathfrak{T}$ to these inequalities and using the Subvariance (Definition A.1) and Monotonicity (Definition A.2) properties, we obtain that for all $s \in \mathcal{S}$,

$$(\mathfrak{T}\boldsymbol{u})(s) - \gamma c \leq (\mathfrak{T}\boldsymbol{v})(s) \leq (\mathfrak{T}\boldsymbol{u})(s) + \gamma c \ .$$

It follows that for all $s \in \mathcal{S}$,

$$|(\mathfrak{T}\boldsymbol{v})(s) - (\mathfrak{T}\boldsymbol{u})(s)| \leq \gamma c \ ,$$

and therefore $\|\mathfrak{T}\boldsymbol{u} - \mathfrak{T}\boldsymbol{v}\|_\infty \leq \gamma c$, proving the stated result.

□

Proposition A.4 (Validity)**.** *Suppose that the reward for any state s and action a, and next state s' is independent of the next state s', i.e., $R(s, a, s') = R(s, a)$, where $R(s, a) \in \mathbb{R}$. Then, the following properties hold for all value functions $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{\mathcal{S}}$.*

1. *The operator $\mathcal{T}_{\mathrm{VaR}}$ is contraction mapping on $\mathbb{R}^{\mathcal{S}}$: $\|\mathcal{T}_{\mathrm{VaR}}\boldsymbol{u} - \mathcal{T}_{\mathrm{VaR}}\boldsymbol{v}\|_\infty \leq \gamma \|\boldsymbol{u} - \boldsymbol{v}\|_\infty$.*
2. *The operator $\mathcal{T}_{\mathrm{VaR}}$ is monotone: $\boldsymbol{u} \succeq \boldsymbol{v} \Rightarrow \mathcal{T}_{\mathrm{VaR}}\boldsymbol{u} \succeq \mathcal{T}_{\mathrm{VaR}}\boldsymbol{v}$.*
3. *The equality $\mathcal{T}_{\mathrm{VaR}}\hat{\boldsymbol{v}} = \hat{\boldsymbol{v}}$ has a unique solution.*

Proof. From Lemma A.3, we know that an operator is a contraction mapping if it satisfies Monotonicity and Subvariance property.

In this proof, we will show that the VaR Bellman operator $\mathcal{T}^{\pi}_{\mathrm{VaR}}$ satisfies Monotonicity and Subvariance property, and therefore, is a contraction mapping.

We will use shorthand $r_{s,a}$ to denote the reward corresponding to state s and action a, i.e., $r_{s,a} = R(s, a)$.

First, we show that $\mathcal{T}_{\mathrm{VaR}}$ satisfies Subvariance property. Consider any $c \in \mathbb{R}$ and state s. Then,

$$\begin{aligned}
(\mathcal{T}_{\mathrm{VaR}}(\boldsymbol{v} + c\mathbf{1}))(s) &= \max_{a \in \mathcal{A}} \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}(r_{s,a} + \gamma(\boldsymbol{v} + c\mathbf{1}))] \\
&\overset{(a)}{=} \max_{a \in \mathcal{A}} \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}(r_{s,a} + \gamma\boldsymbol{v} + \gamma c\mathbf{1})] \\
&\overset{(b)}{=} \max_{a \in \mathcal{A}} \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}(r_{s,a} + \gamma\boldsymbol{v}) + \gamma c] \\
&\overset{(c)}{=} \max_{a \in \mathcal{A}} \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}(r_{s,a} + \gamma\boldsymbol{v})] + \gamma c \\
&\overset{(d)}{=} \mathcal{T}_{\mathrm{VaR}}\boldsymbol{v} + \gamma c \ .
\end{aligned}$$

(a) follows from simple algebraic manipulations, (b) follows from $\gamma c \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\mathbf{1} = \gamma c$, (c) follows from the translational invariance property of VaR measure (30), and (d) follows the definition of the VaR Bellman operator $\mathcal{T}_{\mathrm{VaR}}^{\pi}$.

Next, we show that $\mathcal{T}_{\mathrm{VaR}}$ satisfies Monotonicity property.

Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be any two value functions such that $\boldsymbol{v} \preceq \boldsymbol{u}$. Consider any state $s \in \mathcal{S}$. Then,

$$\begin{aligned}
(\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v})(s) - (\mathcal{T}_{\mathrm{VaR}}\boldsymbol{u})(s) &= \max_{a \in \mathcal{A}} \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}(r_{s,a} + \gamma\boldsymbol{v})] - \max_{a \in \mathcal{A}} \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}(r_{s,a} + \gamma\boldsymbol{u})] \\
&\overset{(a)}{=} \max_{a \in \mathcal{A}} \mathrm{VaR}_\alpha[r_{s,a} + \gamma\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{v}] - \max_{a \in \mathcal{A}} \mathrm{VaR}_\alpha[r_{s,a} + \gamma\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{u}] \\
&\overset{(b)}{=} \max_{a \in \mathcal{A}}(r_{s,a} + \gamma\,\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{v}]) - \max_{a \in \mathcal{A}}(r_{s,a} + \gamma\,\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{u}]) \\
&\overset{(c)}{\leq} 0 \\
(\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v})(s) &\leq (\mathcal{T}_{\mathrm{VaR}}\boldsymbol{u})(s)
\end{aligned}$$

(a) follows from the fact that $r_{s,a}$ is independent of the next state s', (b) follows from the translational invariance property of the VaR measure (30), (c) follows from the fact that, for any action a, $\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{u}] \geq \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{v}]$ because $\boldsymbol{u} \succeq \boldsymbol{v}$ and the VaR measure satisfies monotonicity property (30). Thus, $\forall a \in \mathcal{A}$, $r_{s,a} + \gamma\,\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{v}] \leq r_{s,a} + \gamma\,\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{u}]$ and therefore, $\max_{a \in \mathcal{A}}(r_{s,a} + \gamma\,\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{v}]) \leq \max_{a \in \mathcal{A}}(r_{s,a} + \gamma\,\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{u}])$.

Thus, we prove that $\mathcal{T}_{\mathrm{VaR}}$ is a γ-contraction mapping and a monotone operator. Since $\mathcal{T}_{\mathrm{VaR}}$ is a contraction operator on a Banach space, the Banach fixed point theorem (1) implies that the operator $\mathcal{T}_{\mathrm{VaR}}$ has a unique solution $\hat{\boldsymbol{v}}$, i.e., $\mathcal{T}_{\mathrm{VaR}}\hat{\boldsymbol{v}} = \hat{\boldsymbol{v}}$. □

Proposition A.5. *Suppose that $\tilde{P}$ is normally distributed, i.e., for any state s and action a, $\tilde{\boldsymbol{p}}_{s,a} \sim \mathcal{N}(\bar{\boldsymbol{p}}_{s,a}, \boldsymbol{\Sigma}_{s,a})$.*

Then, $\mathcal{T}_{\mathrm{VaR}}$ for any state $s \in \mathcal{S}$ takes the form

$$(\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v})(s) = \max_{a \in \mathcal{A}} \left(\bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a} - \Phi^{-1}(1-\alpha)\sqrt{\boldsymbol{w}_{s,a}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}\right) \ .$$

Proof. Consider the VaR Bellman optimality operator defined for any state s and value function $\boldsymbol{v}$ as

$$(\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v})(s) = \max_{a \in \mathcal{A}} \mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}\right] \ . \tag{11}$$

From the theory of multivariate normal distributions (6), we know that, for any state s and action a, since $\bar{\boldsymbol{p}}_{s,a}$ is normally distributed $\mathcal{N}(\bar{\boldsymbol{p}}_{s,a}, \boldsymbol{\Sigma}_{s,a})$, $\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}$ is also normally distributed $\mathcal{N}(\bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}, \boldsymbol{w}_{s,a}^\mathsf{T}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a})$. To find the $\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}]$ for any state s and action a, it is sufficient

to find t such that $\Pr(\tilde{\boldsymbol{p}}_{s,a}^{\mathsf{T}}\boldsymbol{w}_{s,a} \geq t) = 1-\alpha$.

$$\Pr\left(\frac{(\tilde{\boldsymbol{p}}-\bar{\boldsymbol{p}}_{s,a})^{\mathsf{T}}\boldsymbol{w}_{s,a}}{\sqrt{\boldsymbol{w}_{s,a}^{\mathsf{T}}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}} > \frac{t-\bar{\boldsymbol{p}}_{s,a}^{\mathsf{T}}\boldsymbol{w}_{s,a}}{\sqrt{\boldsymbol{w}_{s,a}^{\mathsf{T}}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}}\right) = 1-\alpha$$

$$1-\Phi\left(\frac{t-\bar{\boldsymbol{p}}_{s,a}^{\mathsf{T}}\boldsymbol{w}_{s,a}}{\sqrt{\boldsymbol{w}_{s,a}^{\mathsf{T}}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}}\right) = 1-\alpha$$

$$\left(\frac{t-\bar{\boldsymbol{p}}_{s,a}^{\mathsf{T}}\boldsymbol{w}_{s,a}}{\sqrt{\boldsymbol{w}_{s,a}^{\mathsf{T}}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}}\right) = \Phi^{-1}(\alpha)$$

$$t = \bar{\boldsymbol{p}}_{s,a}^{\mathsf{T}}\boldsymbol{w}_{s,a} + \Phi^{-1}(\alpha)\sqrt{\boldsymbol{w}_{s,a}^{\mathsf{T}}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}$$

$$t = \bar{\boldsymbol{p}}_{s,a}^{\mathsf{T}}\boldsymbol{w}_{s,a} - \Phi^{-1}(1-\alpha)\sqrt{\boldsymbol{w}_{s,a}^{\mathsf{T}}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}} \ ,$$

The first equation follows from substracting $\bar{\boldsymbol{p}}_{s,a}^{\mathsf{T}}\boldsymbol{w}_{s,a}$ and dividing by $\sqrt{\boldsymbol{w}_{s,a}^{\mathsf{T}}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}$ on both sides. The second equality follows from the definition of CDF of $\mathcal{N}(0,1)$ and the third equality follows from simple algebraic manipulations.

Substituting the value of $t = \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^{\mathsf{T}}\boldsymbol{w}_{s,a}] = \bar{\boldsymbol{p}}_{s,a}^{\mathsf{T}}\boldsymbol{w}_{s,a} - \Phi^{-1}(1-\alpha)\sqrt{\boldsymbol{w}_{s,a}^{\mathsf{T}}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}$ in (11), we obtain the stated results. □

Proposition A.5 is useful when the Bernstein von Mises theorem holds for transition probability $\tilde{P}$, i.e., as the number of posterior samples N goes to ∞, the posterior distribution of $\tilde{\boldsymbol{p}}_{s,a}$ for any state s and action a converges to a multivariate normal centered at the maximum likelihood true transition model $\bar{\boldsymbol{p}}_{s,a}$ with covariance matrix given by the Fisher information matrix $I(\boldsymbol{p}_{s,a}^{\star})^{-1}/N$ of the true transition probabilities $\boldsymbol{p}_{s,a}^{\star}$ (34).

Proposition A.6. *Let $\hat{\boldsymbol{q}} \in \mathbb{R}^{S\times A}$ represent the optimal action-value function corresponding to the optimal* VaR *policy $\hat{\pi}$. For any $k \in \mathbb{Z}$, let $\boldsymbol{f}^k$ represent the empirical estimate of the kth action-value function $\boldsymbol{q}^k$ in Algorithm 3.1 computed with atmost ζ error, i.e., $\|\boldsymbol{f}^k - \boldsymbol{q}^k\|_\infty \leq \zeta$. If $k \geq \frac{\ln(\frac{R_{\max}}{(\varepsilon-\zeta)(1-\gamma)})}{1-\gamma}$ where $R_{\max} = \max_{s\in\mathcal{S},a\in\mathcal{A},s'\in\mathcal{S}} r(s,a,s')$ and $\varepsilon > \zeta$, then,*

$$\|\hat{\boldsymbol{q}} - \boldsymbol{f}^k\|_\infty \leq \varepsilon \ .$$

Proof. Let $\hat{\boldsymbol{f}}$ be the empirical fixed point of the empirical VaR Bellman operator $\hat{\mathcal{T}}_{\mathrm{VaR}}$ defined for any value function $\boldsymbol{v}$ and state s as

$$(\hat{\mathcal{T}}_{\mathrm{VaR}}\boldsymbol{v})(s) = \max_{a\in\mathcal{A}} \widehat{\mathrm{VaR}}_\alpha[\tilde{p}_{s,a}^{\mathsf{T}}\boldsymbol{w}_{s,a}] \ .$$

To prove the result, we assume that the empirical VaR Bellman operator $\hat{\mathcal{T}}_{\mathrm{VaR}}$ is also a contraction mapping like the VaR Bellman operator $\mathcal{T}_{\mathrm{VaR}}$. Then,

$$\begin{aligned}
\|\hat{\boldsymbol{q}} - \boldsymbol{f}^k\|_\infty &\overset{\text{(a)}}{=} \|\hat{\boldsymbol{q}} - \hat{\boldsymbol{f}} + \hat{\boldsymbol{f}} - \boldsymbol{f}^k\| \\
&\overset{\text{(b)}}{\leq} \|\hat{\boldsymbol{q}} - \hat{\boldsymbol{f}}\| + \|\hat{\boldsymbol{f}} - \boldsymbol{f}^k\|_\infty \\
&\overset{\text{(c)}}{\leq} \zeta + \|\hat{\mathcal{T}}_{\mathrm{VaR}}\boldsymbol{f} - \hat{\mathcal{T}}_{\mathrm{VaR}}\boldsymbol{f}^{k-1}\| \\
&\overset{\text{(d)}}{\leq} \zeta + \gamma\|\hat{\boldsymbol{f}} - \boldsymbol{f}^{k-1}\| \\
&\overset{\text{(e)}}{\leq} \gamma\|\hat{\mathcal{T}}_{\mathrm{VaR}}\hat{\boldsymbol{f}} - \hat{\mathcal{T}}_{\mathrm{VaR}}\boldsymbol{f}^{k-2}\| + \zeta \\
&\overset{\text{(f)}}{\leq} \gamma^k\frac{R_{\max}}{1-\gamma} + \zeta \ .
\end{aligned}$$

(a) follows from simply adding and subtracting $\hat{\boldsymbol{f}}$, (b) follows from triangle inequality, the first term of (c) follows from the assumption that $\|\boldsymbol{q}^k - \boldsymbol{f}^k\|_\infty \leq \zeta$ for any $k \in \mathbb{Z}$ and the second term follows from the properties of VaR_α Bellman operator, (d) follows from the contraction property of $\widehat{\mathcal{T}_{\mathrm{VaR}}}$, (e) follows from applying the same procedure as in (a) to step (d), and (f) follows from unrolling (e) over k time steps and $\|\hat{\boldsymbol{f}} - \boldsymbol{f}^0\|_\infty \leq {}^{R_{\max}}/_{1-\gamma}$ for $\boldsymbol{f}^0 = \mathbf{0}$.

We find k such that $\|\hat{\boldsymbol{q}} - \boldsymbol{f}^k\|_\infty \leq \varepsilon$,

$$\frac{\gamma^k}{1-\gamma}(R_{\max}) + \zeta \leq \varepsilon$$
$$k \geq \frac{\ln\left(\frac{R_{\max}}{(\varepsilon-\zeta)(1-\gamma)}\right)}{1-\gamma} \quad .$$

□

Given a policy π, a state $s \in \mathcal{S}$, and a transition model P, let

$$(\mathcal{T}_P^\pi \boldsymbol{v})(s) = \sum_{a \in \mathcal{A}} \pi(s,a) \boldsymbol{p}_{s,a}^\mathsf{T} \boldsymbol{w}_{s,a} \text{ and } (\mathcal{T}_{\mathrm{VaR}}^\pi \boldsymbol{v})(s) = \sum_{a \in \mathcal{A}} \pi(s,a)\, \mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T} \boldsymbol{w}_{s,a}\right] \quad ,$$

represent the Bellman evaluation operator corresponding to transition model P and the VaR Bellman evaluation operator, respectively.

Lemma A.7. *For any policy π, let $\hat{\boldsymbol{v}}^\pi$ and $\boldsymbol{v}^\pi$ be the fixed point of the* VaR *policy evaluation operator $\mathcal{T}_{\mathrm{VaR}}^\pi$ and Bellman policy evaluation operator $\mathcal{T}_P^\pi$. If the VaR Bellman policy evaluation operator $\mathcal{T}_{\mathrm{VaR}}^\pi$ dominates the Bellman policy evaluation operator $\mathcal{T}_P^\pi$ at $\hat{\boldsymbol{v}}^\pi$, i.e., $\mathcal{T}_{\mathrm{VaR}}^\pi \hat{\boldsymbol{v}}^\pi \preceq \mathcal{T}_P^\pi \hat{\boldsymbol{v}}^\pi$, then, the fixed point of the* VaR *Bellman evaluation operator $\mathcal{T}_{\mathrm{VaR}}^\pi$ dominates the fixed point of the Bellman evaluation operator $\mathcal{T}_P^\pi$, i.e., $\hat{\boldsymbol{v}}^\pi \preceq \boldsymbol{v}^\pi$.*

We note that in contrast to the Bellman policy evaluation operator $\mathcal{T}_P^\pi$, the VaR Bellman policy evaluation operator $\mathcal{T}_{\mathrm{VaR}}^\pi$ is a function of the random variable $\tilde{P}$ and is not dependent on the transition model P assumed in this setting.

Using the assumption $\mathcal{T}_{\mathrm{VaR}} \hat{\boldsymbol{v}}^\pi \preceq \mathcal{T}_P^\pi \hat{\boldsymbol{v}}^\pi$, and from $\hat{\boldsymbol{v}}^\pi = \mathcal{T}_{\mathrm{VaR}}^\pi \hat{\boldsymbol{v}}^\pi$ and $\boldsymbol{v}^\pi = \mathcal{T}_P^\pi \boldsymbol{v}^\pi$, we get by algebraic manipulations:

Proof.

$$\hat{\boldsymbol{v}}^\pi - \boldsymbol{v}^\pi = \mathcal{T}_{\mathrm{VaR}}^\pi \hat{\boldsymbol{v}}^\pi - \mathcal{T}_P^\pi \boldsymbol{v}^\pi \preceq \mathcal{T}_P^\pi \hat{\boldsymbol{v}}^\pi - \mathcal{T}_P^\pi \boldsymbol{v}^\pi \preceq \gamma P_\pi (\hat{\boldsymbol{v}}^\pi - \boldsymbol{v}^\pi) \quad .$$

Here P_π is the transition probability function corresponding to policy π. Subtracting $\gamma P_\pi(\hat{\boldsymbol{v}}^\pi - \boldsymbol{v}^\pi)$ from the above inequality gives,

$$(I - \gamma P^\pi)(\hat{\boldsymbol{v}}^\pi - \boldsymbol{v}^\pi) \preceq \mathbf{0} \quad .$$

where I is the identity matrix. $(I - \gamma P_\pi)^{-1}$ is monotone as can be seen from its Neumann series.

$$\hat{\boldsymbol{v}}^\pi - \boldsymbol{v}^\pi \preceq (I - \gamma P_\pi)^{-1} \mathbf{0} = \mathbf{0} \quad .$$

which proves the result. □

Proposition A.8. *For any policy $\boldsymbol{\pi}$, the* VaR *Bellman evaluation operator $\mathcal{T}_{\mathrm{VaR}}^\pi$ dominates the BCR Bellman evaluation operator $\mathcal{T}_{BCR}^\pi$ i.e., $\mathcal{T}_{\mathrm{VaR}}^\pi \boldsymbol{v} \succeq \mathcal{T}_{BCR}^\pi \boldsymbol{v}$. Similar results hold for policy optimization operators, i.e., $\mathcal{T}_{\mathrm{VaR}} \boldsymbol{v} \succeq \mathcal{T}_{BCR} \boldsymbol{v}$.*

Proof. Recall the VaR Bellman and BCR Bellman operators defined for any value function $\boldsymbol{v}$, policy $\pi \in \Pi_D$, confidence level α, and state s as

$$(\mathcal{T}_{\mathrm{VaR}}^\pi \boldsymbol{v})(s) = \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T} \boldsymbol{w}_{s,\pi(s)}]$$
$$(\mathcal{T}_{\mathrm{BCR}}^\pi \boldsymbol{v})(s) = \min_{\boldsymbol{p}_{s,\pi(s)} \in \mathcal{P}_{s,\pi(s)}} \tilde{\boldsymbol{p}}_{s,\pi(s)}^\mathsf{T} \boldsymbol{w}_{s,\pi(s)} \quad .$$

Suppose that for any value function $\boldsymbol{v}$, confidence level α, state s and action a, $\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}\boldsymbol{w}_{s,a}] < \min_{\boldsymbol{p}_{s,a}\in\mathcal{P}^{\mathrm{BCR}}_{s,a}} \tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a}$. We know that,

$$\begin{aligned}
&\Pr\left(\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a} \le \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}\boldsymbol{w}_{s,a}]\right) > \alpha \\
&\implies \Pr(\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a} > \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}\boldsymbol{w}_{s,a}]) < 1-\alpha \\
&\implies \Pr(\tilde{\boldsymbol{p}}_{s,a} \in \mathcal{P}^{\mathrm{BCR}}_{s,a}) < 1-\alpha \ ,
\end{aligned}$$

which is a contradiction since the BCR ambiguity set $\mathcal{P}^{\mathrm{BCR}}$ is constructed such that $\Pr[\tilde{P} \in \mathcal{P}^{\mathrm{BCR}}] \ge 1-\alpha$.

Since, we proved that $\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}\boldsymbol{w}_{s,a}] \ge \min_{\boldsymbol{p}_{s,a}\in\mathcal{P}_{s,a}} \tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a}$ for any action a, state s and value function $\boldsymbol{v}$, it follows that $(\mathcal{T}^\pi_{\mathrm{VaR}}\boldsymbol{v})(s) \ge (\mathcal{T}^\pi_{\mathrm{BCR}}\boldsymbol{v})(s)$ and $(\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v})(s) \ge (\mathcal{T}_{\mathrm{BCR}}\boldsymbol{v})(s)$ for all value functions $\boldsymbol{v}$ and state s.

□

B Proofs

B.1 Proof of Proposition 3.1

Proposition 3.1 (Lower Bound Percentile Criterion)**.** *For any $\delta \in (0,1)$, if we set the confidence level α in the operator $\mathcal{T}^\pi_{\mathrm{VaR}}$ to δ/S, then for every policy $\pi \in \Pi : \Pr_{\tilde{P}}[\hat{\boldsymbol{v}}^\pi \preceq \tilde{\boldsymbol{v}}^\pi|\mathcal{D}] \ge 1-\delta$, where $\tilde{P}$ is a realization of the posterior distribution of the transition model P conditioned on observed transitions $\mathcal{D}$.*

Proof. Let $\alpha = \delta/S$. Recall that for any policy π, state $s \in \mathcal{S}$, transition model $\tilde{P}$, the Bellman evaluation operator $\mathcal{T}^\pi$ and the VaR Bellman evaluation operator $\mathcal{T}^\pi_{\mathrm{VaR}}$ are defined as

$$(\mathcal{T}^\pi\boldsymbol{v})(s) = \sum_{a\in\mathcal{A}} \pi(s,a)\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a} \text{ and } (\mathcal{T}^\pi_{\mathrm{VaR}}\boldsymbol{v})(s) = \sum_{a\in\mathcal{A}} \pi(s,a)\,\mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\boldsymbol{w}_{s,a}\right] \ ,$$

respectively.

It is also important to note that the Bellman operator $\mathcal{T}^\pi$ is defined for a random transition model $\tilde{P}$.

Let $\hat{\boldsymbol{v}}^\pi$ be the fixed point of $\mathcal{T}^\pi_{\mathrm{VaR}}$ conditioned on observed transitions $\mathcal{D}$, and let $\tilde{\boldsymbol{v}}^\pi$ be a random variable that represents the fixed point of $\mathcal{T}^\pi$ for a given realization $\tilde{P}$ of the posterior distribution of the transition model P given $\mathcal{D}$. Then, applying Lemma A.7 to $\mathcal{T}^\pi_{\mathrm{VaR}}$ and $\mathcal{T}^\pi$, we have, $\hat{\boldsymbol{v}}^\pi \preceq \boldsymbol{v}^\pi$ implies

$$\mathcal{T}^\pi_{\mathrm{VaR}}\hat{\boldsymbol{v}}^\pi \preceq \mathcal{T}^\pi\hat{\boldsymbol{v}}^\pi$$

That is for each state s,

$$\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,\pi(s)}\hat{\boldsymbol{v}}^\pi] \le \tilde{\boldsymbol{p}}^\mathsf{T}_{s,\pi(s)}\hat{\boldsymbol{v}}^\pi \ . \tag{12}$$

Using the equation (12), we can bound the probability that the VaR value function lower bounds the true value. □

$$\Pr_{\tilde{P}}[\hat{\boldsymbol{v}}^\pi \preceq \tilde{\boldsymbol{v}}^\pi|\mathcal{D}] = \Pr_{\tilde{P}}\left[\forall s\in\mathcal{S}:\ \mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,\pi(s)}\hat{\boldsymbol{v}}^\pi\right] \le \tilde{\boldsymbol{p}}^\mathsf{T}_{s,\pi(s)}\hat{\boldsymbol{v}}^\pi\middle|\mathcal{D}\right] \ . \tag{13}$$

From the definition of VaR, we know that for any state s and action a,

$$\Pr_{\tilde{P}}\left[\mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\hat{\boldsymbol{v}}^\pi\right] \le \tilde{\boldsymbol{p}}^\mathsf{T}_{s,a}\hat{\boldsymbol{v}}^\pi\middle|\mathcal{D}\right] \ge 1-\alpha, \tag{14}$$

Therefore, using union bound and (13) in (14), we can write

$$\Pr_{\tilde{P}}[\hat{\boldsymbol{v}}^\pi \succ \tilde{\boldsymbol{v}}^\pi|\mathcal{D}] \le \sum_{s\in\mathcal{S}} \Pr_{\tilde{P}}\left[\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}^\mathsf{T}_{s,\pi(s)}\hat{\boldsymbol{v}}^\pi] > \tilde{\boldsymbol{p}}^\mathsf{T}_{s,\pi(s)}\hat{\boldsymbol{v}}^\pi\middle|\mathcal{D}\right] \ .$$

Thus,

$$\Pr[\hat{\boldsymbol{v}}^\pi \succ \tilde{\boldsymbol{v}}^\pi|\mathcal{D}] = \sum_{s\in\mathcal{S}} \frac{\delta}{S} = S\frac{\delta}{S} = \delta \ .$$

B.2 Proof of Proposition 3.2

Proposition 3.2. *Suppose that $\tilde{\boldsymbol{p}}_{s,a}$ for any state s and action a, is a multivariate sub-Gaussian with mean $\bar{\boldsymbol{p}}_{s,a}$ and covariance factor $\boldsymbol{\Sigma}_{s,a}$, i.e., $\mathbb{E}\left[\exp\left(\lambda(\tilde{\boldsymbol{p}}_{\boldsymbol{s},\boldsymbol{a}} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\boldsymbol{w}\right)\right] \leq \exp\left(\lambda^2 \boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}/2\right), \forall \lambda \in \mathbb{R}, \forall \boldsymbol{w} \in \mathbb{R}^S$. Then, for any state $s \in \mathcal{S}$, $\mathcal{T}_{\mathrm{VaR}}$ satisfies*

$$(\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v})(s) \geq \max_{a\in\mathcal{A}} \left(\bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a} - \sqrt{2\ln(1/\alpha)}\sqrt{\boldsymbol{w}_{s,a}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}\right) \quad .$$

Proof.

$$\begin{aligned}
(\mathcal{T}_{\mathrm{VaR}}^{\pi}\boldsymbol{v})_s &= \max_{a\in\mathcal{A}} \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}] \\
&\overset{(a)}{=} \max_{a\in\mathcal{A}} \inf\left\{t \middle| \Pr\left(\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a} \leq t\right) > \alpha\right\} \\
&\overset{(b)}{=} \max_{a\in\mathcal{A}} \inf\left\{t \middle| \Pr\left((\tilde{\boldsymbol{p}}_{s,a} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\boldsymbol{w}_{s,a} \leq (t - \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a})\right) > \alpha\right\} \\
&\overset{(c)}{=} \max_{a\in\mathcal{A}} \inf\left\{t \middle| \Pr\left(\exp((\tilde{\boldsymbol{p}}_{s,a} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\boldsymbol{w}_{s,a}) \leq \exp(t - \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a})\right) > \alpha\right\} \\
&\overset{(d)}{\geq} \max_{a\in\mathcal{A}} \inf\left\{t \middle| \inf_{\lambda>0} \frac{\exp(\lambda^2 \boldsymbol{w}_{s,a}^\mathsf{T}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}/2)}{\exp(\lambda(t - \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}))} > \alpha\right\} \\
&\overset{(e)}{=} \max_{a\in\mathcal{A}} \inf\left\{t \middle| \exp\left(\frac{-(t - \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a})^2}{2\boldsymbol{w}_{s,a}^\mathsf{T}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}\right) > \alpha\right\} \\
&\overset{(f)}{=} \max_{a\in\mathcal{A}} \inf\left\{t \middle| (t - \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a})^2 < -2\ln(\alpha)\boldsymbol{w}_{s,a}^\mathsf{T}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}\right\} \\
&\overset{(g)}{=} \max_{a\in\mathcal{A}} \inf\left\{t \middle| (t - \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}) \in \left(-\sqrt{2\ln(1/\alpha)}\sqrt{\boldsymbol{w}_{s,a}^\mathsf{T}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}, \sqrt{2\ln(1/\alpha)}\sqrt{\boldsymbol{w}_{s,a}^\mathsf{T}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}\right)\right\} \\
&\overset{(h)}{=} \max_{a\in\mathcal{A}} \left(\bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a} - \sqrt{2\ln(1/\alpha)}\sqrt{\boldsymbol{w}_{s,a}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}\right) \quad .
\end{aligned}$$

Equality (a) follows from the definition of VaR, (b) follows from subtracting $\bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}$ on both sides, (c) follows from taking exponential on both sides, (d) follows from using the upper-bound given by Chernoff bound for a sub-Gaussian distribution (6) i.e., $\Pr\left(\exp((\tilde{\boldsymbol{p}}_{s,a} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\boldsymbol{w}_{s,a}) \leq \exp(t - \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a})\right) \leq \inf_{\lambda>0} \frac{\exp(\lambda^2 \boldsymbol{w}_{s,a}^\mathsf{T}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}/2)}{\exp(\lambda(t - \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}))}$, (e) follows from solving for λ and getting $\lambda = \frac{(t - \bar{\boldsymbol{p}}_{s,a})}{\boldsymbol{w}_{s,a}^\mathsf{T}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}}$, (f) follows from taking $\ln$ on both sides, (g) follows from simple algebraic manipulations and (h) follows from taking the infimum of the solution interval of t.

Solving for t in step (g), we get $t = \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}] = \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a} - \sqrt{2\ln(1/\alpha)}\sqrt{\boldsymbol{w}_{s,a}\boldsymbol{\Sigma}_{s,a}\boldsymbol{w}_{s,a}})$ which proves the stated result. □

B.3 Proof of Theorem 3.3

Theorem 3.3 (Performance)**.** *Let $\hat{\boldsymbol{v}}$ be the fixed point of the* VaR *Bellman optimality operator $\mathcal{T}_{\mathrm{VaR}}$, and $\pi^\star$ be the optimal policy in* (1)*. Let $\rho^\star = \mathrm{VaR}_\alpha\left[\rho(\pi^\star, \tilde{P})\right]$ denote the optimal percentile returns and $\hat{\rho} = \boldsymbol{p}_0^\mathsf{T}\hat{\boldsymbol{v}}$ denote the lower bound on the percentile returns computed using the Bellman operator $\mathcal{T}_{\mathrm{VaR}}$. For any $\delta \in (0,1)$, we set the confidence level $\alpha = \delta/(2SA)$ in $\mathcal{T}_{\mathrm{VaR}}$. Then, with probability at least $1-\delta$, the performance loss with respect to $\rho^\star$ is*

$$\rho^\star - \hat{\rho} \leq \frac{1}{1-\gamma} \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} (\mathrm{VaR}_{1-\alpha}[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\hat{\boldsymbol{w}}_{s,a}] - \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\hat{\boldsymbol{w}}_{s,a}]) \quad . \tag{6}$$

Proof. We prove this theorem in three parts.

Let $\hat{\boldsymbol{v}} \in \mathbb{R}^S$ be the fixed point of the VaR_α Bellman operator $\mathcal{T}_{\mathrm{VaR}}$, i.e., $\hat{\boldsymbol{v}} = \mathcal{T}_{\mathrm{VaR}}\hat{\boldsymbol{v}}$. Recall that the VaR_α returns is given by $\hat{\rho} = p_0^\mathsf{T}\hat{\boldsymbol{v}}$. Furthermore, let $\mathcal{T}_P^\pi$ represent the Bellman evaluation operator

for a given policy $\pi \in \Pi$ and a transition probability model P. Then, $\mathcal{T}_P^\pi$ is defined for each $s \in \mathcal{S}$ as

$$(\mathcal{T}_P^\pi \boldsymbol{v})_s = \boldsymbol{p}_{s,\pi(s)}^\mathsf{T} \boldsymbol{w}_{s,\pi(s)} \ ,$$

where $\boldsymbol{w}_{s,\pi(s)} = \boldsymbol{r}_{s,\pi(s)} + \gamma \cdot \boldsymbol{v}$.

It is well known that the Bellman operator $\mathcal{T}_P^\pi$ is a contraction mapping, monotone, and has a unique fixed point. Let $\tilde{\pi} \in \arg\max_{\pi \in \Pi} \ \rho(\pi, \tilde{P})$. Let $\tilde{\boldsymbol{v}}$ be the unique fixed point of $\mathcal{T}_{\tilde{P}}^{\tilde{\pi}}$:

$$\tilde{\boldsymbol{v}}^{\tilde{\pi}} = \mathcal{T}_{\tilde{P}}^{\tilde{\pi}} \tilde{\boldsymbol{v}}^{\tilde{\pi}} \ .$$

Note that it is well-known that:

$$\boldsymbol{p}_0^\mathsf{T} \tilde{\boldsymbol{v}} = \rho(\tilde{\pi}, \tilde{P}) \ .$$

First, we show that the lower bound on the percentile returns computed by the VaR Bellman operator $\mathcal{T}_{\mathrm{VaR}}$, i.e., ${\boldsymbol{p}_0}^\mathsf{T} \hat{\boldsymbol{v}}$ is less than than the returns corresponding to policy $\tilde{\pi}$, i.e., ${\boldsymbol{p}_0}^\mathsf{T} \tilde{\boldsymbol{v}}$ with high confidence $1 - \delta$.

We can write $\hat{\rho} = \boldsymbol{p}_0^\mathsf{T} \hat{\boldsymbol{v}}$ as

$$\begin{aligned} p_0^\mathsf{T} \hat{\boldsymbol{v}} &\overset{\text{(a)}}{\leq} \mathrm{VaR}_\alpha[\rho(\hat{\pi}, \tilde{P})] \\ &\overset{\text{(b)}}{\leq} \rho(\hat{\pi}, \tilde{P}) \\ &\overset{\text{(c)}}{\leq} \rho(\tilde{\pi}, \tilde{P}) = \boldsymbol{p}_0^\mathsf{T} \tilde{\boldsymbol{v}} \ . \end{aligned}$$

Inequality (a) follows because the VaR Bellman optimality operator optimizes a lower bound on the percentile criterion with high confidence $1 - \delta$. Inequality (b) follows from the definition of VaR. Inequality (c) follows because $\tilde{\pi}$ is optimal for $\tilde{P}$. The above equation implies

$$0 \leq \boldsymbol{p}_0^\mathsf{T} (\hat{\boldsymbol{v}} - \tilde{\boldsymbol{v}}) \leq \|\hat{\boldsymbol{v}} - \tilde{\boldsymbol{v}}\|_\infty$$

Now, we establish the probabilistic bound based on bounding the Bellman residual. We will use the following result to establish an upper bound on $\|\hat{\boldsymbol{v}} - \tilde{\boldsymbol{v}}\|_\infty$.

$$\begin{aligned} (\mathcal{T}_{\tilde{P}}^{\tilde{\pi}} \hat{\boldsymbol{v}} - \hat{\boldsymbol{v}})_s &\overset{\text{(a)}}{=} (\mathcal{T}_{\tilde{P}}^{\tilde{\pi}} \hat{\boldsymbol{v}} - \mathcal{T}_{\mathrm{VaR}} \hat{\boldsymbol{v}})_s \\ &\overset{\text{(def)}}{=} \tilde{\boldsymbol{p}}_{s,\tilde{\pi}(s)}^\mathsf{T} \hat{\boldsymbol{z}}_{s,\tilde{\pi}(s)} - \mathrm{VaR}_\alpha \left[\tilde{\boldsymbol{p}}_{s,\hat{\pi}(s)}^\mathsf{T} \hat{\boldsymbol{z}}_{s,\hat{\pi}(s)} \right] \\ &\overset{\text{(b)}}{\leq} \tilde{\boldsymbol{p}}_{s,\tilde{\pi}(s)}^\mathsf{T} \hat{\boldsymbol{z}}_{s,\tilde{\pi}(s)} - \mathrm{VaR}_\alpha \left[\tilde{\boldsymbol{p}}_{s,\tilde{\pi}(s)}^\mathsf{T} \hat{\boldsymbol{z}}_{s,\tilde{\pi}(s)} \right] \\ &\overset{\text{(c)}}{\leq} \max_{a \in \mathcal{A}} \left(\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T} \hat{\boldsymbol{z}}_{s,a} - \mathrm{VaR}_\alpha \left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T} \hat{\boldsymbol{z}}_{s,a} \right] \right) \\ &\overset{\text{(d)}}{\leq} \max_{a \in \mathcal{A}} \left(\mathrm{VaR}_{1-\alpha} \left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T} \hat{\boldsymbol{z}}_{s,a} \right] - \mathrm{VaR}_\alpha \left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T} \hat{\boldsymbol{z}}_{s,a} \right] \right) \qquad \text{with probability } 1 - 2\alpha A \ . \end{aligned} \tag{15}$$

(a) follows from $\hat{\boldsymbol{v}}$ being the fixed point of $\mathcal{T}_{\mathrm{VaR}}$, (b) follows from the optimality of $\hat{\pi}$, (c) follows from simple algebraic manipulations, and (d) follows from $\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T} \hat{\boldsymbol{z}}_{s,a} \geq \mathrm{VaR}_{1-\alpha}[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T} \hat{\boldsymbol{z}}_{s,a}]$ with probability α, $\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T} \hat{\boldsymbol{z}}_{s,a} \leq \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T} \hat{\boldsymbol{z}}_{s,a}]$ with probability α, and finally, taking a union bound over the set of actions $\mathcal{A}$ yields the given probabilistic bound.

For any state s, let $\hat{\boldsymbol{c}}_s = \max_{a \in \mathcal{A}} \left(\mathrm{VaR}_{1-\alpha} \left[\boldsymbol{p}_{s,a}^\mathsf{T} \hat{\boldsymbol{w}}_{s,a} \right] - \mathrm{VaR}_\alpha \left[\boldsymbol{p}_{s,a}^\mathsf{T} \hat{\boldsymbol{w}}_{s,a} \right] \right)$. Applying the inequality in (15) to all states, we get

$$(\mathcal{T}_{\tilde{P}}^{\tilde{\pi}} \hat{\boldsymbol{v}} - \hat{\boldsymbol{v}}) \preceq \hat{\boldsymbol{c}} \ .$$

We can now use the standard dynamic programming bounding technique to bound $\|\tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}}\|_\infty$ as follows:

$$\begin{aligned} 0 &\overset{\text{(a)}}{\preceq} \tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}} \\ &\overset{\text{(b)}}{\preceq} \tilde{\boldsymbol{v}} - \mathcal{T}_{\tilde{P}}^{\tilde{\pi}} \hat{\boldsymbol{v}} + \mathcal{T}_{\tilde{P}}^{\tilde{\pi}} \hat{\boldsymbol{v}} - \hat{\boldsymbol{v}} \\ &\overset{\text{(c)}}{\preceq} \tilde{\boldsymbol{v}} - \mathcal{T}_{\tilde{P}}^{\tilde{\pi}} \hat{\boldsymbol{v}} + \hat{\boldsymbol{c}} \\ \tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}} &\preceq \mathcal{T}_{\tilde{P}}^{\tilde{\pi}} \tilde{\boldsymbol{v}} - \mathcal{T}_{\tilde{P}}^{\tilde{\pi}} \hat{\boldsymbol{v}} + \hat{\boldsymbol{c}} \ . \end{aligned}$$

We have (a) because $\hat{\boldsymbol{v}} \preceq \tilde{\boldsymbol{v}}$ because $\mathcal{T}_{\mathrm{VaR}}\tilde{\boldsymbol{v}} \preceq \tilde{\boldsymbol{v}}$ and thus, $\tilde{\boldsymbol{v}} \succeq \mathcal{T}_{\mathrm{VaR}}\mathcal{T}_{\mathrm{VaR}}\tilde{\boldsymbol{v}} \succeq \cdots \succeq \mathcal{T}_{\mathrm{VaR}}\ldots\mathcal{T}_{\mathrm{VaR}}\tilde{\boldsymbol{v}} \succeq \hat{\boldsymbol{v}}$ because $\hat{\boldsymbol{v}}$ is the fixed point of $\mathcal{T}_{\mathrm{VaR}}$ and $\mathcal{T}_{\mathrm{VaR}}$ is monotone. We have (b) from simply adding and subtracting $\mathcal{T}_{\tilde{P}}^{\tilde{\pi}}\hat{\boldsymbol{v}}$ and (c) follows from (15).

Applying the ℓ_∞ norm on both sides (i.e., taking the max over all states $s \in \mathcal{S}$), we get

$$\begin{aligned}
\|\tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}}\|_\infty &\leq \|\mathcal{T}_{\tilde{P}}^{\tilde{\pi}}\tilde{\boldsymbol{v}} - \mathcal{T}_{\tilde{P}}^{\tilde{\pi}}\hat{\boldsymbol{v}} + \hat{\boldsymbol{c}}\|_\infty \\
&\overset{\text{(a)}}{\leq} \gamma\|\tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}}\|_\infty + \|\hat{\boldsymbol{c}}\|_\infty \\
\|\tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}}\|_\infty &\overset{\text{(b)}}{\leq} \frac{\max_{s\in\mathcal{S}} \hat{\boldsymbol{c}}_s}{1-\gamma} \quad \text{with probability } (1 - 2\alpha SA) \ .
\end{aligned}$$

Inequality (a) follows by the triangle inequality, and inequality (b) follows from applying the results in equation (15) and taking a union bound over all states in $\mathcal{S}$.

Thus, setting $\alpha = \delta/2SA$, we get $\Pr\left[\|\tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}}\|_\infty \leq \frac{\max_{s\in\mathcal{S}} \hat{\boldsymbol{c}}_s}{1-\gamma}\right] \leq 1 - \delta$.

Finally, to prove the bound on $\rho^\star$ and $\hat{\rho} = \boldsymbol{p}_0^\mathsf{T}\hat{\boldsymbol{v}}$, we need to show $\rho^\star \leq \eta$, where $\eta = \hat{\rho} + \frac{\max_{s\in\mathcal{S}} \hat{\boldsymbol{c}}_s}{1-\gamma}$. Suppose that the contradiction $\rho^\star > \eta$ holds true. Realize that $\rho^\star$ is optimal in (1), and therefore, must satisfy

$$\Pr\left[\max_{\pi\in\Pi} \rho(\pi, \tilde{P}) \geq \rho^\star\right] \geq 1 - \delta \ .$$

Recall from the statement of the theorem that

$$\Pr\left[\max_{\pi\in\Pi} \rho(\pi, \tilde{P}) \geq \eta\right] \leq \delta \ .$$

We can now derive a contradiction.

$$\delta \geq \Pr\left[\max_{\pi\in\Pi} \rho(\pi, \tilde{P}) \overset{\text{(a)}}{\geq} \eta\right] \geq \Pr\left[\max_{\pi\in\Pi} \rho(\pi, \tilde{P}) \geq \rho^\star\right] \geq 1 - \delta \ .$$

where inequality (a) follows from the assumption $\rho^\star \geq \eta$. Thus, we get $\delta \geq 1 - \delta$, which is a contradiction for $\delta < 0.5$. The lower bound $0 \leq \rho^\star - \hat{\rho}$ follows from the optimality of $\rho^\star$ and Proposition 3.1, which proves the proposition. □

B.4 Proof of Theorem 3.4

Theorem 3.4 (Asymptotic Performance)**.** *For any $\delta \in (0, 1)$, set $\alpha = \delta/(2SA)$ in $\mathcal{T}_{\mathrm{VaR}}$. Let $I(\boldsymbol{p}_{s,a}^\star)^{-1}$ for any state s and action a, be the Fisher Information matrix corresponding to the true transition probabilities $\boldsymbol{p}_{s,a}^\star$. Furthermore, let $\sigma_{\max} = \max_{s\in\mathcal{S},a\in\mathcal{A}} \sqrt{\hat{\boldsymbol{w}}_{s,a}^\mathsf{T} I(\boldsymbol{p}_{s,a}^\star)^{-1}\hat{\boldsymbol{w}}_{s,a}}$ represent the maximum asymptotic standard deviation of the returns estimate $\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\hat{\boldsymbol{w}}_{s,a}$ for any state-action pair (s, a). Then, with probability at least $1 - \delta$, the asymptotic performance of the* VaR *framework $\hat{\rho}$ w.r.t. the optimal percentile returns $\rho^\star$ satisfies*

$$\lim_{N\to\infty} \sqrt{N}(\rho^\star - \hat{\rho}) \leq \frac{1}{1-\gamma}\left(2\Phi^{-1}(1-\alpha)\sigma_{\max}\right) \leq \frac{1}{1-\gamma}\sqrt{8\ln(1/\alpha)}\sigma_{\max} \ .$$

Proof. To prove this theorem, we assume that Berstein von Mises theorem (34) holds for $\tilde{P}$, i.e., the posterior distribution of transition probabilities $\tilde{\boldsymbol{p}}_{s,a}$ for any state s and action a converges in the limit to a multivariate normal $\mathcal{N}(\bar{\boldsymbol{p}}_{s,a}, I(\boldsymbol{p}_{s,a}^\star))$ centered at the maximum likelihood estimator of the true transition model $\boldsymbol{p}_{s,a}^\star$ with covariance matrix given by the Fisher information matrix $I(\boldsymbol{p}_{s,a}^\star)/N$ of the true transition probabilities $\boldsymbol{p}_{s,a}^\star$.

Therefore, for any state s and action a, $\lim_{N\to\infty} \sqrt{N}(\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a} - \bar{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}) \rightsquigarrow \mathcal{N}(0, \boldsymbol{w}_{s,a}^\mathsf{T} I(\boldsymbol{p}_{s,a}^\star)\boldsymbol{w}_{s,a})$.

We know from Proposition A.5, that Value at risk of a univariate normal random variable $X \sim \mathcal{N}(\mu, \sigma)$ can be written as $\mathrm{VaR}_\alpha[X] = \mu + \Phi^{-1}(\alpha)\sigma$. Therefore, applying this result to the R.H.S

of Equation (6) gives

$$
\begin{aligned}
\lim_{N\to\infty} \sqrt{N}(\rho^\star - \hat{\rho}) \leq \lim_{N\to\infty} \frac{1}{1-\gamma} \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \Big(& \sqrt{N}\bar{\boldsymbol{p}}_{s,a}^\mathsf{T} \boldsymbol{w}_{s,a} + \Phi^{-1}(1-\alpha)\sqrt{\boldsymbol{w}_{s,a}^\mathsf{T} I({\boldsymbol{p}^\star}_{s,a})^{-1} \boldsymbol{w}_{s,a}} \\
& - \Big(\sqrt{N}\bar{\boldsymbol{p}}_{s,a}^\mathsf{T} \boldsymbol{w}_{s,a} + \Phi^{-1}(\alpha)\sqrt{\boldsymbol{w}_{s,a}^\mathsf{T} I({\boldsymbol{p}^\star}_{s,a})^{-1} \boldsymbol{w}_{s,a}}\Big)\Big) \\
= \frac{1}{1-\gamma} \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \Big(& \Phi^{-1}(1-\alpha)\sqrt{\boldsymbol{w}_{s,a}^\mathsf{T} I({\boldsymbol{p}^\star}_{s,a})^{-1} \boldsymbol{w}_{s,a}} \\
& -\Phi^{-1}(\alpha)\sqrt{\boldsymbol{w}_{s,a}^\mathsf{T} I({\boldsymbol{p}^\star}_{s,a})^{-1} \boldsymbol{w}_{s,a}}\Big) \\
= \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \Big(& \Phi^{-1}(1-\alpha)\sqrt{\boldsymbol{w}_{s,a}^\mathsf{T} I({\boldsymbol{p}^\star}_{s,a})^{-1} \boldsymbol{w}_{s,a}} \\
& -\Phi^{-1}(\alpha)\sqrt{\boldsymbol{w}_{s,a}^\mathsf{T} I({\boldsymbol{p}^\star}_{s,a})^{-1} \boldsymbol{w}_{s,a}}\Big) \\
= \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \Big(& (\Phi^{-1}(1-\alpha) - \Phi^{-1}(\alpha))\sqrt{\boldsymbol{w}_{s,a}^\mathsf{T} I({\boldsymbol{p}^\star}_{s,a})^{-1} \boldsymbol{w}_{s,a}}\Big) \\
= \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \Big(& 2\Phi^{-1}(1-\alpha)\sqrt{\boldsymbol{w}_{s,a}^\mathsf{T} I({\boldsymbol{p}^\star}_{s,a})^{-1} \boldsymbol{w}_{s,a}}\Big) \quad .
\end{aligned}
\tag{16}
$$

We prove the second inequality in Theorem 3.4, by leveraging the sub-Gaussian bounds for a standard normal distribution $\mathcal{N}(0,\sigma)$ to show that $\Phi^{-1}(1-\alpha) \leq \sqrt{2\ln(1/\alpha)}$.

We know that for a standard normal distribution $X \sim \mathcal{N}(0,\sigma)$ where $\sigma \in \mathbb{R}$, the follow sub-Gaussian bounds holds true (6).

$$
\Pr\left(-\sqrt{2\ln(2/\alpha)}\sigma \leq X \leq \sqrt{2\ln(2/\alpha)}\sigma\right) \geq 1-\alpha \quad . \tag{17}
$$

It is well known that for a standard normal distribution $\mathcal{N}(0,\sigma)$, the following equation holds.

$$
\Pr\left(-\Phi^{-1}(1-\alpha/2)\sigma \leq X \leq \Phi^{-1}(1-\alpha/2)\sigma\right) = 1-\alpha \tag{18}
$$

Comparing equation (17) and (18), we get

$$
\begin{aligned}
\Phi^{-1}(1-\alpha/2) &\leq \sqrt{2\ln(2/\alpha)} \\
\equiv \Phi^{-1}(1-\alpha) &\leq \sqrt{2\ln(1/\alpha)} \quad .
\end{aligned}
\tag{19}
$$

Substituting equation (19) in equation (16), proves the second inequality of the theorem.

□

B.5 Proof of Proposition 3.5

Proposition 3.5 (Time Complexity)**.** *The time complexity of a single iteration of the loop in line 3 of the* VaR *Value Iteration (Algorithm 3.1) is in $\mathcal{O}(SAN)$, where N is the number of samples of the posterior samples of $\tilde{P}$.*

Proof. The proposition follows from the fact that any quantile of an array of real values can be computed in linear time using the Quick Select algorithm (16) and a single iteration of the loop in line 3 of Algorithm 3.1 computes quantile of returns SA times. □

B.6 Proof of Proposition 3.6

Proposition 3.6 (Empirical Error Bound)**.** *For any state s, action a and value function $\boldsymbol{v}$, let $\widehat{\mathrm{VaR}}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}]$ represent the empirical estimate of α-percentile of returns $\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}]$ and Φ_f represent the cumulative density function (CDF) of the random estimate of returns $\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}$. Suppose that Φ_f is differentiable at the point $\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}]$ and let $m = \Phi'(\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}])$*

represents the density of estimate of returns at point $\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}]$. *Let* $N^\star$ *be the number of posterior samples required to obtain empirical error* $\varepsilon \in \mathbb{R}$, *with confidence* $1-\zeta$, *i.e.,* $\Pr\left[|\widehat{\mathrm{VaR}}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}] - \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}]| > \varepsilon\right] \leq \zeta$. *Then,* $\lim_{\varepsilon\to 0} N^\star\varepsilon^2 = {}^{\ln(2/\zeta)}/_{2m^2}$.

Proof. To prove this theorem, we first compute the derivative of the inverse of the CDF ${}^{\partial\Phi^{-1}}/_{\partial\alpha}$ as follows. From the definition of the cdf Φ_f and VaR, we know that, for any $\alpha \in (0, 0.5)$, $\Phi_f^{-1}(\alpha) = \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}]$.

From the inverse-function theorem, we get,

$$\begin{aligned}(\Phi_f^{-1}(\alpha))' &= \frac{1}{\Phi_f'(\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}])} \\ &= \frac{1}{\Phi_f'(\Phi_f^{-1}(\alpha))} \\ &= \frac{1}{m} \ .\end{aligned}$$

Equipped with the above result, we can now proceed to prove the main result.

To prove the result, we need to find $N^\star$ such that

$$\begin{aligned}&\Pr\left[\mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}] - \varepsilon \leq \widehat{\mathrm{VaR}}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}] \leq \mathrm{VaR}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}] + \varepsilon\right] \geq 1-\zeta \\ &\overset{(a)}{=} \Pr\left[\Phi_f^{-1}(\alpha - \varepsilon m) \leq \widehat{\mathrm{VaR}}_\alpha[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}] \leq \Phi_f^{-1}(\alpha + \varepsilon m)\right] \geq 1-\zeta \ .\end{aligned} \tag{20}$$

Equation (a) follows from applying a first order Taylor expansion to Φ_f^{-1} around the point α. We apply the following results to obtain a bound on $N^\star$.

Let $\hat{F}$ and F represent the empirical CDF and the true CDF of a random variable $\tilde{Z}$. Suppose that the empirical estimate of the CDF $\hat{F}$ is estimated using $N^\star$ samples from the true distribution of $\tilde{Z}$ and $0 < \zeta < 1$ represents the desired level of confidence guarantees, Then, from DWK inequality (23), we know that

$$\Pr\left(\|\hat{F} - F\|_\infty \geq \sqrt{{}^{\ln(2/\zeta)}/_{2N^\star}}\right) \leq \zeta$$

The above equation implies that

$$\Pr\left(\exists p \in (0,1) : F^{-1}(p) < \hat{F}^{-1}(p - l_t) \text{or } F^{-1}(p) > \hat{F}^{-1}(p + u_t)\right) \leq \zeta \tag{21}$$

where $l_t = u_t = \sqrt{{}^{\ln(2/\zeta)}/_{2N^\star}}$.

Thus, applying equation (21) to (20), i.e., $l_t = \sqrt{{}^{\ln(2/\zeta)}/_{2N^\star}} = \varepsilon m$ gives $N^\star = {}^{\ln(2/\zeta)}/_{2\varepsilon^2 m^2}$.

□

B.7 Proof of Proposition 4.1

Proposition 4.1 (Equivalence)**.** *Let* $\hat{\boldsymbol{v}}$ *be the fixed point of the* VaR *Bellman optimality operator* $\mathcal{T}_{\mathrm{VaR}}$, *i.e,* $\hat{\boldsymbol{v}} = (\mathcal{T}_{\mathrm{VaR}}^\pi \hat{\boldsymbol{v}})$. *Let* Π^D *be the set of all deterministic policies. Then,* $\hat{\boldsymbol{v}}$ *is the value function of policy* $\hat{\pi}$ *where* $\hat{\pi}$ *is an optimal solution of the following problem,*

$$\arg\max_{\pi\in\Pi^D} \min_{P\in\mathcal{P}^{\mathrm{VaR},\boldsymbol{v}^\pi}} \rho(\pi, P) \ . \tag{8}$$

Proof. Consider the VaR Bellman optimality operator.

$$\forall s \in \mathcal{S}, \quad \boldsymbol{v} \in \mathbb{R}^{\mathcal{S}},\ (\mathcal{T}_{\mathrm{VaR}}^\pi \boldsymbol{v})_s = \max_{a\in\mathcal{A}} \mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}\right] \ . \tag{22}$$

Notice that $\mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}\right]$ is convex in $\tilde{\boldsymbol{p}}_{s,a}$. Thus, using the definition of support functions of convex functions (8), we can write Equation (22) as

$$\begin{aligned}&\forall s \in \mathcal{S}, \quad \boldsymbol{v} \in \mathbb{R}^{\mathcal{S}}, (\mathcal{T}_{\mathrm{VaR}}^\pi \boldsymbol{v})_s = \max_{a\in\mathcal{A}} \min_{P\in\mathcal{P}_{s,a}^{\mathrm{VaR},\boldsymbol{v}}} \boldsymbol{p}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a} \ , \\ &\qquad\qquad \text{where } \mathcal{P}_{s,a}^{\mathrm{VaR},\boldsymbol{v}} = \left\{\boldsymbol{p}_{s,a} \in \Delta^S \middle| \boldsymbol{p}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a} \geq \mathrm{VaR}_\alpha\left[\boldsymbol{p}_{s,a}^\mathsf{T}\boldsymbol{w}_{s,a}\right]\right\} \ .\end{aligned} \tag{23}$$

Equation (23) represents the VaR Bellman operator $\mathcal{T}^{\pi}_{\mathrm{VaR}}$ (policy evaluation) as a Robust Bellman operator with a special ambiguity set that is dependent on the value function $\boldsymbol{v}$. Thus, from Equation 23 and theory of value iteration in RL, we get that the fixed point of VaR Bellman optimality operator $\mathcal{T}_{\mathrm{VaR}}$ (policy optimization) is the value function of the optimal policy that solves the following optimization problem.

$$\begin{aligned}
&\arg\max_{\pi\in\Pi^D} \min_{P\in\mathcal{P}^{\mathrm{VaR},\boldsymbol{v}^\pi}} \rho(\pi,P)\\
&\mathcal{P}^{\mathrm{VaR},\boldsymbol{v}^\pi} = \bigtimes \mathcal{P}^{\mathrm{VaR},\boldsymbol{v}^\pi}_{s,a}, \quad \mathcal{P}^{\mathrm{VaR},\boldsymbol{v}^\pi}_{s,a} = \left\{\boldsymbol{p}_{s,a}\in\Delta^S \middle| \boldsymbol{p}^{\mathsf{T}}_{s,a}\boldsymbol{v}^\pi \geq \mathrm{VaR}_\alpha\left[\tilde{\boldsymbol{p}}^{\mathsf{T}}_{s,a}\boldsymbol{v}^\pi\right]\right\} \quad .
\end{aligned}$$

□

B.8 Proof of Proposition 4.2

Proposition 4.2. *For any policy π, the fixed point of the* VaR *policy evaluation operator $\mathcal{T}^{\pi}_{\mathrm{VaR}}$ dominates the fixed point of the Bellman evaluation operator $\mathcal{T}^{\pi}_{BCR}$, i.e., $\mathcal{T}^{\pi}_{BCR}\cdots\mathcal{T}^{\pi}_{BCR}\boldsymbol{v} \preceq \mathcal{T}^{\pi}_{\mathrm{VaR}}\cdots\mathcal{T}^{\pi}_{\mathrm{VaR}}\boldsymbol{v}$ for any $\boldsymbol{v}$. Similar results hold for policy optimization operators $\mathcal{T}_{\mathrm{VaR}}$ and $\mathcal{T}_{BCR}$, i.e., $\mathcal{T}_{BCR}\cdots\mathcal{T}_{BCR}\boldsymbol{v} \preceq \mathcal{T}_{\mathrm{VaR}}\cdots\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v}$ for any $\boldsymbol{v}$.*

We prove the first part of the proposition using induction. First, we verify that $(\mathcal{T}^{\pi}_{\mathrm{VaR}}\boldsymbol{v}) \succeq (\mathcal{T}^{\pi}_{\mathrm{BCR}}\boldsymbol{v})$ fo any $\boldsymbol{v}$.

Proof.

$$\mathcal{T}^{\pi}_{\mathrm{VaR}}\boldsymbol{v} \overset{(a)}{\succeq} \mathcal{T}^{\pi}_{\mathrm{BCR}}\boldsymbol{v} \quad \forall \boldsymbol{v}\in\mathbb{R}^S$$

(a) follows from the definition of $\mathcal{T}^{\pi}_{\mathrm{VaR}}$ and $\mathcal{T}^{\pi}_{\mathrm{BCR}}$ since the constraints of BCR ambiguity sets encompass the constraints of the ambiguity sets implicitly constructed by $\mathcal{T}^{\pi}_{\mathrm{VaR}}$ (Proposition A.8).

Suppose that for for any $\boldsymbol{v}\in\mathbb{R}^S$ and $k\in\mathcal{Z}$, applying $\mathcal{T}^{\pi}_{\mathrm{VaR}}$ and $\mathcal{T}^{\pi}_{\mathrm{BCR}}$ k times to $\boldsymbol{v}$ yields, $(\mathcal{T}^{\pi}_{\mathrm{VaR}})^k\boldsymbol{v} \succeq (\mathcal{T}^{\pi}_{\mathrm{BCR}})^k\boldsymbol{v}$. Then,

$$\begin{aligned}
(\mathcal{T}^{\pi}_{\mathrm{VaR}})^{k+1}\boldsymbol{v} &\overset{(a)}{\succeq} \mathcal{T}^{\pi}_{\mathrm{VaR}}(\mathcal{T}^{\pi}_{\mathrm{BCR}})^k\boldsymbol{v}\\
(\mathcal{T}^{\pi}_{\mathrm{VaR}})^{k+1}\boldsymbol{v} &\overset{(b)}{\succeq} \mathcal{T}^{\pi}_{\mathrm{BCR}}(\mathcal{T}^{\pi}_{\mathrm{BCR}})^k\boldsymbol{v}\\
(\mathcal{T}^{\pi}_{\mathrm{VaR}})^{k+1}\boldsymbol{v} &\succeq (\mathcal{T}^{\pi}_{\mathrm{BCR}})^{k+1}\boldsymbol{v} \quad .
\end{aligned}$$

(a) follows from the monotonicity property of $\mathcal{T}^{\pi}_{\mathrm{VaR}}$, (b) follows from the fact that $\mathcal{T}^{\pi}_{\mathrm{VaR}}\boldsymbol{w} \succeq \mathcal{T}^{\pi}_{\mathrm{BCR}}\boldsymbol{w}$ for any $\boldsymbol{w}$ which in turn, follows from the definition of $\mathcal{T}^{\pi}_{\mathrm{VaR}}$ and $\mathcal{T}^{\pi}_{\mathrm{BCR}}$. Therefore, $\mathcal{T}^{\pi}_{\mathrm{VaR}}(\mathcal{T}^{\pi}_{\mathrm{BCR}})^k\boldsymbol{v} \succeq \mathcal{T}^{\pi}_{\mathrm{BCR}}(\mathcal{T}^{\pi}_{\mathrm{BCR}})^k\boldsymbol{v}$. Thus, by induction, it follows that $\mathcal{T}^{\pi}_{\mathrm{VaR}}\cdots\mathcal{T}^{\pi}_{\mathrm{VaR}}\boldsymbol{v} \succeq \mathcal{T}^{\pi}_{\mathrm{BCR}}\cdots\mathcal{T}^{\pi}_{\mathrm{BCR}}\boldsymbol{v}$ for any $\boldsymbol{v}$.

Now, we prove the second part of the proposition. Let , $\hat{\boldsymbol{v}}$ and $\boldsymbol{v}^{\mathrm{BCR}}$ be the fixed points of $\mathcal{T}_{\mathrm{VaR}}$ and $\mathcal{T}_{\mathrm{BCR}}$ respectively. Supposed that $\boldsymbol{v}^{\mathrm{BCR}} \succ \hat{\boldsymbol{v}}$, then

$$\begin{aligned}
\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v}^{\mathrm{BCR}} &\overset{(a)}{\succ} \mathcal{T}_{\mathrm{VaR}}\hat{\boldsymbol{v}}\\
\mathcal{T}_{\mathrm{VaR}}\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v}^{\mathrm{BCR}} &\overset{(b)}{\succ} \mathcal{T}_{\mathrm{VaR}}\mathcal{T}_{\mathrm{VaR}}\hat{\boldsymbol{v}}\\
\mathcal{T}_{\mathrm{VaR}}\ldots\mathcal{T}_{\mathrm{VaR}}\mathcal{T}_{\mathrm{VaR}}\boldsymbol{v}^{\mathrm{BCR}} &\overset{(c)}{\succ} \mathcal{T}_{\mathrm{VaR}}\ldots\mathcal{T}_{\mathrm{VaR}}\mathcal{T}_{\mathrm{VaR}}\hat{\boldsymbol{v}}\\
\hat{\boldsymbol{v}} &\overset{(d)}{\succ} \hat{\boldsymbol{v}} \quad .
\end{aligned}$$

(a), (b), (c) follows from the monotonicity property of $\mathcal{T}_{\mathrm{VaR}}$ and (d) follows from the fact that $\hat{\boldsymbol{v}}$ is the unique fixed point of $\mathcal{T}_{\mathrm{VaR}}$. The last equation is a contradiction. therefore, it must be that $\boldsymbol{v}^{\mathrm{BCR}} \preceq \hat{\boldsymbol{v}}$.

□

B.9 Proof of Theorem 4.3

Theorem 4.3 (Asymptotic Size of VaR Ambiguity Sets)**.** *Let $\bar{P} = (\bar{\boldsymbol{p}}_{s,a})_{s\in\mathcal{S},a\in\mathcal{A}}$ be the Maximum Likelihood estimate of transition probabilities computed from data $\mathcal{D}$ and $\Sigma = (I(\boldsymbol{p}^{\star}_{s,a})^{-1})_{s\in\mathcal{S},a\in\mathcal{A}}$ be the corresponding covariance matrix. Then, as $N \to \infty$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$,*

$$\mathcal{P}^{\mathrm{VaR}}_{s,a} = \left\{ \boldsymbol{p}_{s,a} \in \Delta^{\mathcal{S}} \middle| \frac{1}{\sqrt{N}} \|\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a}\|_{\boldsymbol{\Sigma}^{-1}_{s,a}} \leq \frac{\Phi^{-1}(1-\alpha)}{\sqrt{N}} \right\} \quad . \tag{9}$$

To prove this theorem, we first establish some primary results.

We note that to prove this theorem, we assume that Bernstein von Mises theorem holds for transition probabilities $\tilde{P}$.

For any state s, action a and value function $\boldsymbol{v}$, consider the 1-step VaR_α Bellman update

$$f_{s,a}(\boldsymbol{v}) = \mathrm{VaR}_\alpha \left[\tilde{\boldsymbol{p}}^{\mathsf{T}}_{s,a} (\boldsymbol{r}_{s,a} + \gamma \boldsymbol{v}) \right] , \tag{24}$$

Furthermore, let $\boldsymbol{w}_{s,a} = \boldsymbol{r}_{s,a} + \gamma \boldsymbol{v}$. If the posterior of transition probability $\tilde{\boldsymbol{p}}_{s,a}$ has mean $\bar{\boldsymbol{p}}_{s,a} \in \mathbb{R}^S$ and covariance matrix $\boldsymbol{\Sigma}_{s,a} = I(\boldsymbol{p}^{\star}{}_{s,a})^{-1} \in \mathbb{R}^{S\times S}$ and Bernstein von Mises theorem (34) holds for posterior of transition probability $\tilde{\boldsymbol{p}}_{s,a}$, then as $N \to \infty$, the returns $\tilde{\boldsymbol{p}}^{\mathsf{T}}_{s,a} \boldsymbol{w}_{s,a} \sim \mathcal{N}\left(\bar{\boldsymbol{p}}^{\mathsf{T}}_{s,a} \boldsymbol{w}_{s,a}, \frac{\boldsymbol{w}^{\mathsf{T}}_{s,a} I(\boldsymbol{p}^{\star}_{s,a}) \boldsymbol{w}_{s,a}}{N} \right)$. Then, we can write equation (24) as

$$f_{s,a}(\boldsymbol{v}) = \left(\bar{\boldsymbol{p}}^{\mathsf{T}}_{s,a} \boldsymbol{w}_{s,a} - \Phi^{-1}(1-\alpha) \|\boldsymbol{w}_{s,a}\|_{\boldsymbol{\Sigma}^{-1}_{s,a}} \right) , \tag{25}$$

where $\Phi^{-1}(1-\alpha)$ represents the $1-\alpha$ percentile of standard normal distribution. (25) follows from the analytical form of VaR of a normal random variable, i.e., for any normal random variable $\tilde{Y}$ with mean μ, variance matrix σ and confidence level α, $\mathrm{VaR}_\alpha[\tilde{Y}] = \mu - \Phi^{-1}(1-\alpha)\sigma$.

Since $f_{s,a}(\boldsymbol{v})$ is convex in $\boldsymbol{v}$ when $\tilde{\boldsymbol{p}}_{s,a}$ is normally distributed, we can use the definition of support functions of a closed convex set (8) to construct a unique ambiguity set $\mathcal{P}^{\mathrm{VaR}}_{s,a}$ of the form

$$\mathcal{P}^{\mathrm{VaR}}_{s,a} = \left\{ \boldsymbol{p}_{s,a} \middle| \bar{\boldsymbol{p}}_{s,a} \leq \bar{\boldsymbol{p}}^{\mathsf{T}}_{s,a} \boldsymbol{w}_{s,a} - \Phi^{-1}(1-\alpha) \|\boldsymbol{w}_{s,a}\|_{\boldsymbol{\Sigma}^{-1}_{s,a}}, \forall \boldsymbol{v} \in \mathbb{R}^S \right\} \quad . \tag{26}$$

The above equation implies that

$$\forall \boldsymbol{v}, \min_{\boldsymbol{p}_{s,a} \in \mathcal{P}^{\mathrm{VaR}}_{s,a}} \boldsymbol{p}^{\mathsf{T}}_{s,a} \boldsymbol{w}_{s,a} = \mathrm{VaR}_\alpha \left[\tilde{\boldsymbol{p}}^{\mathsf{T}}_{s,a} \boldsymbol{w}_{s,a} \right] \quad .$$

Using basic algebraic manipulations as shown in Proposition B.1, we can alternatively express the ambiguity set $\mathcal{P}^{\mathrm{VaR}}_{s,a}$ in the form of a semi-ellipsoidal ball with radius $\Phi^{-1}(1-\alpha)$.

$$\mathcal{P}^{\mathrm{VaR}}_{s,a} = \left\{ \boldsymbol{p}_{s,a} \middle| \frac{1}{\sqrt{N}} \|\bar{\boldsymbol{p}}_{s,a} - \boldsymbol{p}_{s,a}\|_{\boldsymbol{\Sigma}^{-1}_{s,a}} \leq \frac{\Phi^{-1}(1-\alpha)}{\sqrt{N}} \right\} \quad . \tag{27}$$

We note that although $\boldsymbol{\Sigma}_{s,a}$ are invertible, we can proceed as described in Section 3.1 in (15) to find an approximation of their inverses.

Proposition B.1. *Consider the two ambiguity sets given in equations* (26) *and* (27)*. These two representations are equivalent.*

Proof. We begin with equation (26),

$$\begin{aligned}
&\boldsymbol{p}^{\mathsf{T}}_{s,a} \boldsymbol{w}_{s,a} \overset{\text{(a)}}{\leq} \bar{\boldsymbol{p}}^{\mathsf{T}}_{s,a} \boldsymbol{w}_{s,a} - \Phi^{-1}(1-\alpha) \|\boldsymbol{w}_{s,a}\|_{\boldsymbol{\Sigma}^{-1}_{s,a}} \\
&\boldsymbol{p}^{\mathsf{T}}_{s,a} \boldsymbol{w}_{s,a} - \bar{\boldsymbol{p}}^{\mathsf{T}}_{s,a} \boldsymbol{w}_{s,a} \overset{\text{(b)}}{\leq} -\Phi^{-1}(1-\alpha) \|\boldsymbol{w}_{s,a}\|_{\boldsymbol{\Sigma}^{-1}_{s,a}} \\
&\boldsymbol{w}^{\mathsf{T}}_{s,a} (\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})^{\mathsf{T}} \boldsymbol{w}_{s,a} \overset{\text{(c)}}{\leq} \Phi^{-1}(1-\alpha)^2 \boldsymbol{w}^{\mathsf{T}}_{s,a} \boldsymbol{\Sigma}_{s,a} \boldsymbol{w}_{s,a} \quad .
\end{aligned}$$

The first and second equation follow from the definition of $\mathcal{P}^{\mathrm{VaR}}_{s,a}$ and simple algebraic manipulations. The third equation follows by squaring on both sides.

Equation (c) is obtained by simply squaring Equation (b) on both sides. Using the basic properties of semi-positive definite matrices, we can write the above equation as

$$\begin{aligned}
\left(\Phi^{-1}(1-\alpha)^2\boldsymbol{\Sigma}_{s,a} - (\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\right) &\succeq 0 \\
(\boldsymbol{\Sigma}_{s,a})^\mathsf{T}\left(\Phi^{-1}(1-\alpha)^2\boldsymbol{\Sigma}_{s,a} - (\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\right)\boldsymbol{\Sigma}_{s,a} &\succeq 0 \\
(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})(\boldsymbol{\Sigma}_{s,a}^{-1})^\mathsf{T}\left((\Phi^{-1}(1-\alpha))^2\boldsymbol{\Sigma}_{s,a} - (P_{s,a} - \bar{\boldsymbol{p}}_{s,a})(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\right)\boldsymbol{\Sigma}_{s,a}^{-1}(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a}) &\overset{(1)}{\succeq} 0 \\
\Phi^{-1}(1-\alpha)^2(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\boldsymbol{\Sigma}_{s,a}^{-1}(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a}) - \left((\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\boldsymbol{\Sigma}_{s,a}(P_{s,a} - \bar{\boldsymbol{p}}_{s,a})\right)^2 &\succeq 0
\end{aligned}$$

Equation (1) holds since $\boldsymbol{U}^\mathsf{T}\boldsymbol{M}\boldsymbol{U} \succeq 0 \; \forall \, \boldsymbol{U}, \boldsymbol{M} \succeq 0$.

$$\begin{aligned}
(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})^\mathsf{T}\boldsymbol{\Sigma}_{s,a}(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a}) &\leq \Phi^{-1}(1-\alpha)^2 \\
\|(\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a})\|_{\boldsymbol{\Sigma}_{s,a}^{-1}} &\leq \Phi^{-1}(1-\alpha) \;\; .
\end{aligned}$$

The proof for the other direction is simply the reverse of this proof and hence we omit it. □

B.10 Proof of Theorem 4.4

Theorem 4.4 (Asymptotic size of Bayesian Credible Regions)**.** *For any state s and action a, let $\mathcal{P}_{s,a}^{BCR}$ be any Bayesian credible region and $\bar{\boldsymbol{p}}_{s,a}$ be the maximum likelihood estimator based on data $\mathcal{D}$. Furthermore, $\xi < \sqrt{\chi^2_{S,1-\alpha}}/\Phi^{-1}(1-\alpha)$. Then, as $N \to \infty$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$,*

$$\mathcal{P}_{s,a}^{BCR} - \bar{\boldsymbol{p}}_{s,a} \not\subseteq \xi(\mathcal{P}_{s,a}^{\mathrm{VaR}} - \bar{\boldsymbol{p}}_{s,a}) \;\; . \tag{10}$$

This theorem follows directly from Theorem 10 in (15). For the sake of clarity, we re-derive the theorem below.

We note that to prove this theorem, we assume that Bernstein von Mises theorem holds for transition probabilities $\tilde{P}$.

Lemma B.2. *For any positive semidefinite matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and any $\boldsymbol{v}$, the following inequality holds true.*

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

This proof of this lemma can be found in Lemma C.2 in (15). We will use this lemma to prove the theorem.

For any state-action pair (s, a), we will use the shorthand $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$ to represent the mean $\bar{\boldsymbol{p}}_{s,a}$ and the covariance matrix $\boldsymbol{\Sigma}_{s,a}$ of transition probabilities $\tilde{\boldsymbol{p}}_{s,a}$ for any state s and action a, such that $\bar{\boldsymbol{p}}_{s,a}$ and $\boldsymbol{\Sigma}_{s,a}$ are estimated using N samples from the posterior distribution of transition probabilities f. We will also use the shorthand $\tilde{\boldsymbol{p}}$ to represent $\tilde{\boldsymbol{p}}_{s,a}$ and $\boldsymbol{p}^\star$ to represent $\boldsymbol{p}^\star_{s,a}$.

Define $\mathcal{P}(\mathcal{D}, \tau)$ for any $\tau \in \mathbb{R}$ and dataset $\mathcal{D}$, as

$$\mathcal{P}(\mathcal{D}, \tau) = \left\{\boldsymbol{p} \in \Delta^S \middle| \frac{1}{N}\|\boldsymbol{p} - \boldsymbol{\mu}_N\|_{\boldsymbol{\Sigma}_N} \leq \tau\right\}.$$

Notice that $\mathcal{P}(\mathcal{D}, {}^{\Phi^{-1}(1-\alpha)}/\sqrt{N})$ is the asymptotic ambiguity set $\mathcal{P}^{\mathrm{VaR}}$ in Theorem 4.3. Let $\mathcal{P}(\mathcal{D})$ represent a Bayesian credible region for any dataset $\mathcal{D}$. Consider any $\mathcal{D}$ such that $\mathcal{P}(\mathcal{D}) - \boldsymbol{\mu}_N \subseteq \xi(\mathcal{P}(\mathcal{D}, {}^{\Phi^{-1}(1-\alpha)}/\sqrt{N}) - \boldsymbol{\mu}_N)$. Since $\mathcal{P}(\mathcal{D})$ is a credible region,

$$\begin{aligned}
1 - \varepsilon &\leq \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}(\tilde{\boldsymbol{p}} - \boldsymbol{\mu}_N \in \mathcal{P}(\mathcal{D}) - \boldsymbol{\mu}_N) \\
&\leq \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}(\tilde{\boldsymbol{p}} - \boldsymbol{\mu}_N \in \xi\left(\mathcal{P}(\mathcal{D}, {}^{\Phi^{-1}(1-\alpha)}/\sqrt{N}) - \boldsymbol{\mu}_N\right) \\
&= \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}\left(N^{-1}(\tilde{\boldsymbol{p}} - \boldsymbol{\mu}_N)^\mathsf{T}\boldsymbol{\Sigma}_N^{-1}(\tilde{\boldsymbol{p}} - \boldsymbol{\mu}_N) \leq {}^{\xi^2\Phi^{-1}(1-\alpha)^2}/N\right) \;\; .
\end{aligned}$$

Since $\xi \leq \frac{\sqrt{\chi^2_{S,1-\alpha}}}{\Phi^{-1}(1-\alpha)}$, there exists $\delta \geq 0$, such that $\xi^2 \Phi^{-1}(1-\alpha)^2 \leq \chi^2_{S,1-\delta-\alpha} - \delta$. Fix such δ. Then,

$$
\begin{aligned}
&\Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}\left(N^{-1}(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N)^\mathsf{T}\boldsymbol{\Sigma}_N^{-1}(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N) \leq {}^{\xi^2\Phi^{-1}(1-\alpha)^2}/_N\right) \\
&\quad\leq \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}\left(N^{-1}(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N)^\mathsf{T}\boldsymbol{\Sigma}_N^{-1}(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N) \leq {}^{(\chi^2_{S,1-\delta-\alpha}-\delta)}/_N\right) \\
&\quad\leq \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}\left(N^{-1}(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N)^\mathsf{T}I(\boldsymbol{p}^\star_{s,a})(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N) \leq {}^{(\chi^2_{S,1-\delta-\alpha})}/_N\right) \\
&\quad+ \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}\left(N^{-1}(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N)^\mathsf{T}(N^{-1}\boldsymbol{\Sigma}_N^{-1} - I(\boldsymbol{p}^\star))(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N) \leq {}^{-\delta}/_N\right) \\
&\quad\leq \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}(\|\sqrt{N}(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N)\|^2_{I(\boldsymbol{p}^\star)^{-1}} \leq \chi^2_{S,1-\alpha-\delta}) + Z(\mathcal{D}) \\
&\quad\leq \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}(\|\zeta\|^2_{I(\boldsymbol{p}^\star)^{-1}} \leq \chi^2_{S,1-\alpha-\delta}) + R_N(\mathcal{D}) + Z(\mathcal{D}) \ .
\end{aligned}
$$

where $\zeta \sim \mathcal{N}(0, I(\boldsymbol{p}^\star)^{-1})$, $R_N(\mathcal{D}) = \sup_{\mathcal{A}} |\Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}(\sqrt{N}(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N)) - \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}(\zeta \in \mathcal{A})|$ denotes the total variational distance for the realization $\mathcal{D}$, where $\mathcal{A}$ is a measurable subset of Δ^S (see Bernstein von Mises theorem in Theorem 6 in (15)), $I(\boldsymbol{p}^\star)$ is the Fisher information matrix of $\Pr_{\mathcal{D}|\boldsymbol{p}^\star}$, and $Z(\mathcal{D}) = \Pr((\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N)^\mathsf{T}(N^{-1}\Sigma_N^{-1} - I(\boldsymbol{p}^\star_{s,a}))(\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N)| > \frac{\delta}{N})$. The first probability is atmost $1-\alpha-\delta$. Thus, for any $\mathcal{D}$ such that the theorem statement does not hold, $R_N(\mathcal{D}) + S(\mathcal{D}) > \delta$.

Fix $t > 0$ such that $\Pr(\|\zeta\|_{I(\boldsymbol{p}^\star)^{-1}} > t) \leq 0.5\delta$. From, the second inequality in Lemma B.2,

$$
\begin{aligned}
Z(\mathcal{D}) &\leq \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}(\|\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N\|_{I(\boldsymbol{p}^\star)^{-1}}\sqrt{\|N^{-1}\boldsymbol{\Sigma}_N^{-1} - I(\boldsymbol{p}^\star)\|_F} > \sqrt{\delta/N}) \\
&\leq \Pr_{\tilde{\boldsymbol{p}}|\mathcal{D}}(\|\tilde{\boldsymbol{p}}-\boldsymbol{\mu}_N\|_{I(\boldsymbol{p}^\star)^{-1}} > {}^{t}/_{\sqrt{N}}) + \mathbb{I}(\sqrt{\|N^{-1}\boldsymbol{\Sigma}_N^{-1} - I(\boldsymbol{p}^\star)\|_F} > {}^{\sqrt{\delta}}/_t) \\
&\leq \Pr(\|\zeta\|_{I(\boldsymbol{p}^\star)^{-1}} > t) + R(\mathcal{D}) + \mathbb{I}(\sqrt{\|N^{-1}\boldsymbol{\Sigma}_N^{-1} - I(\boldsymbol{p}^\star)\|_F} > {}^{\sqrt{\delta}}/_t) \\
&0.5\delta + R_N(\mathcal{D}) + \mathbb{I}(\sqrt{\|N^{-1}\boldsymbol{\Sigma}_N^{-1} - I(\boldsymbol{p}^\star)\|_F} > {}^{\sqrt{\delta}}/_t) \ .
\end{aligned}
$$

Thus, for any randomly drawn $\tilde{\mathcal{D}}$,

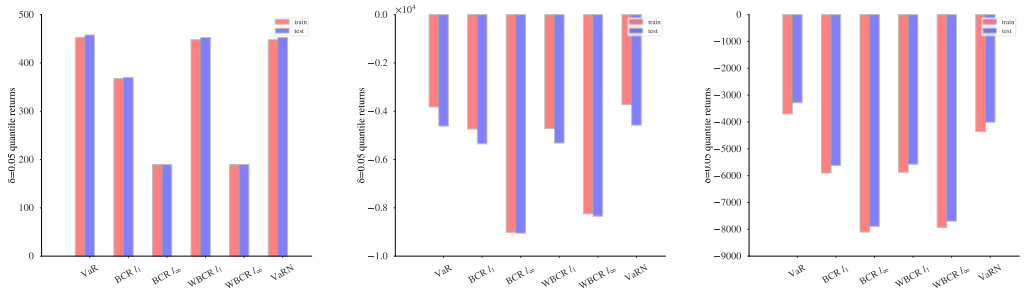
$$
\begin{aligned}
&\Pr_{\tilde{\mathcal{D}}}(\text{Theorem does not hold}) \leq \Pr_{\tilde{\mathcal{D}}}(R_N(\tilde{\mathcal{D}}) + Z(\tilde{\mathcal{D}}) > \delta) \\
&\leq \Pr_{\tilde{\mathcal{D}}}(2R_N(\tilde{\mathcal{D}}) + \mathbb{I}(\sqrt{\|N^{-1}\boldsymbol{\Sigma}_N^{-1} - I(\boldsymbol{p}^\star)\|_F} > {}^{\sqrt{\delta}}/_t) > 0.5\delta) \\
&\leq \Pr_{\tilde{\mathcal{D}}}(2R_N(\tilde{\mathcal{D}}) > 0.25\delta) + \Pr(\mathbb{I}(\sqrt{\|N^{-1}\boldsymbol{\Sigma}_N^{-1} - I(\boldsymbol{p}^\star)\|_F} > {}^{\sqrt{\delta}}/_t) > 0.25\delta) \ .
\end{aligned}
$$

From the assumption that the Bernstein-Von-Mises theorem holds for $\tilde{P}$, the first probability tends to zero, and since $N\boldsymbol{\Sigma}_N \to I(\boldsymbol{p}^\star)^{-1}$ for a large number of samples, the second probability tends to zero as well. This proves the theorem.

C Experiments

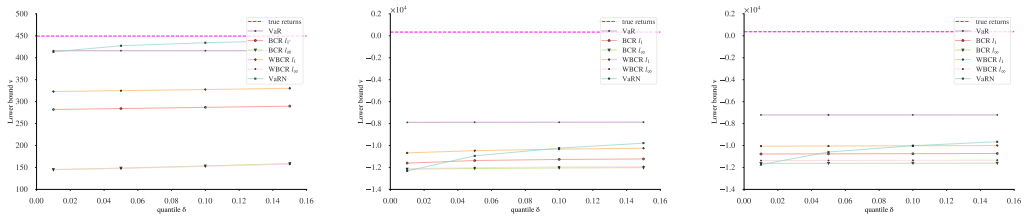
Domain	*VaR*	*VaRN*	*BCR* (ℓ_1)	*BCR* (ℓ_∞)	*WBCR* (ℓ_1)	*WBCR* (ℓ_∞)
Riverswim	74.69	74.69	74.69	74.69	74.69	74.69
Population	**-3994.32**	**-3789.34**	-4809.7	-8695.62	-4802.13	-7937.44
Population-Small	**-2480.6**	**-3193.98**	-4648.93	-6981.7	-4622.99	-6760.08
Inventory	**469.6**	**467.6**	375.02	189.86	464.58	189.86

Table 2: Summary of robust returns at $\delta = 0.10$ achieved by *VaR*, *VaRN*, *BCR* ℓ_1, *BCR* ℓ_∞, *WBCR* ℓ_1 and *WBCR* ℓ_∞ in Riverswim, Inventory, Population-Small, and Population domain. *VaR* and *VarN* achieves the highest robust returns in most of the domains.



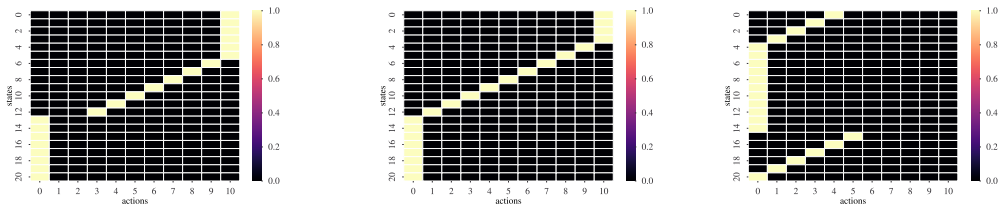
(a) Inventory Domain (b) Population Domain (c) Population-Small Domain

Figure 3: Comparison of test and train robust returns achieved by *VaR*, *VaRN*, *BCR* ℓ_1, *BCR* ℓ_∞, *WBCR* ℓ_1 and *WBCR* ℓ_∞ at confidence level $\delta = 0.05$ in Inventory, Population-Small and Population domain. VaR framework achieves the highest robust returns in all the domains on test and train datasets.



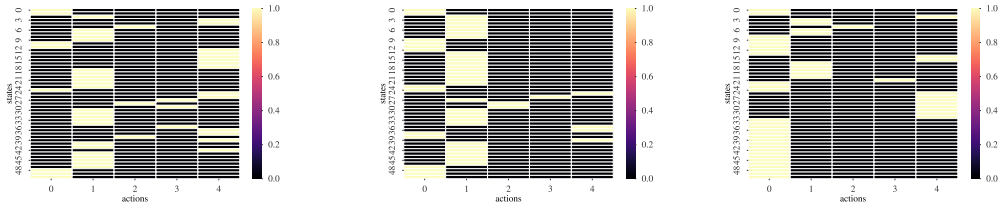
(a) Inventory Domain (b) Population Domain (c) Population-Small Domain

Figure 4: Comparison of robust lower bound value achieved by *VaR*, *VaRN*, *BCR* ℓ_1, *BCR* ℓ_∞, *WBCR* ℓ_1 and *WBCR* ℓ_∞ ambiguity sets for different confidence levels δ in Inventory, Population-Small and Population domain. VaR_α achieves the highest robust returns in all the domains.



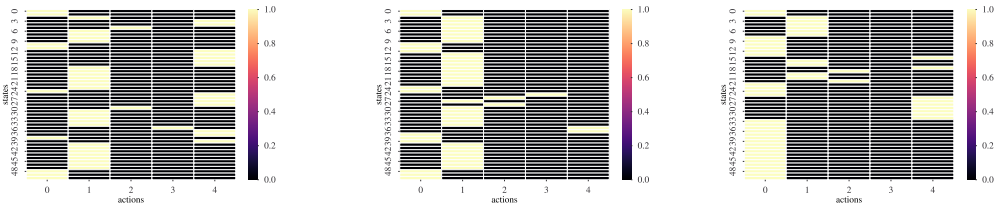
(a) VaR Policy (b) WBCR ℓ_1 policy (c) WBCR ℓ_∞ policy

Figure 5: Comparison of robust policies corresponding to VaR, WBCR ℓ_1, WBCR ℓ_∞ ambiguity sets at confidence level $\delta = 0.1$ in Inventory domain.



(a) VaR Policy (b) WBCR ℓ_1 policy (c) WBCR ℓ_∞ policy

Figure 6: Comparison of robust policies corresponding to VaR, WBCR ℓ_1, WBCR ℓ_∞ ambiguity sets at confidence level $\delta = 0.1$ in Population-Small domain.



(a) VaR Policy (b) WBCR ℓ_1 policy (c) WBCR ℓ_∞ policy

Figure 7: Comparison of robust policies corresponding to VaR, WBCR ℓ_1, WBCR ℓ_∞ ambiguity sets at confidence level $\delta = 0.1$ in Population domain.

D Implementation Details

Hyperparameters for Riverswim Domain	
Number of train models	200
Number of test models	200

Hyperparameters for Inventory Domain	
Number of train models	100
Number of test models	100

Hyperparameters for Population-Small Domain	
Number of train models	50
Number of test models	50

Hyperparameters for Population Domain	
Number of train models	1000
Number of test models	1000

D.1 Code

We have provided the code in the supplementary materials. Since the dataset for the population domain is very large, we were unable to add it to the supplementary materials. We will make the dataset publicly available after the paper is published.

D.2 Machine Specifications

We ran all the experiments on MacBook Air (M2 2022) with 16GB Memory and 8 cores. The total computational time $\sim$3 hours.