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 ambiguity 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 ambiguity sets is often challenging. Existing works use Bayesian credible regions as ambiguity 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 ambiguity sets. Our theoretical and empirical results show that our algorithm implicitly constructs much smaller ambiguity sets and learns less conservative robust policies.

1 Introduction

Batch Reinforcement Learning (Batch RL) [26] 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 [10, 13, 27, 40, 45, 47] to model uncertainty and optimize the policy to maximize the returns corresponding to the worst $\alpha$-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 ambiguity set (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 ambiguity set.

Since the percentile criterion is non-convex, constructing ambiguity sets itself is a challenging problem. Existing work uses Bayesian credible regions (BCR$_\alpha$) [40] as ambiguity sets. However, these ambiguity sets are often unnecessarily large [15, 40] and result in learning conservative robust
policies. Some recent work approximates ambiguity sets using various heuristics [3, 40], but we show that they remain too conservative. Thus, the question of the optimal ambiguity set, i.e., ambiguity sets that result in optimizing the tightest possible lower bound on the percentile criterion and less-conservative policies remains unanswered.

Our Contributions In this paper, we answer two important questions: a) Are Bayesian credible regions the optimal ambiguity sets for optimizing the percentile criterion? b) Can we obtain a less conservative solution to the percentile criterion than RMDPs with BCR ambiguity sets while retaining its percentile guarantees? Our theoretical findings show that Bayesian credible regions can grow significantly with the number of states 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 provide a Generalized VaR Value iteration algorithm and analyze its error bounds. We also show that there exist directions in which the Bayesian credible regions can grow unnecessarily large with the number of states in the MDP and possibly result in a conservative solution. On the other hand, the ambiguity sets implicitly optimized by the VaR Bellman operator tend to be smaller, i.e., they have a smaller asymptotic radius and are independent of the number of states (Section 4). Finally, we empirically demonstrate the efficacy of our framework in three domains (Section 5).

1.1 Related Work
Several works [13, 34, 40] propose different methods for solving the percentile criterion, as well as other robust measures for handling uncertainty in the transition probabilities estimates. Russel and Petrik [40] and Behzadian et al. [3] propose various heuristics for minimizing the size of the ambiguity sets constructed for the percentile-criterion. Russel and Petrik [40] propose a method that interlaces robust value iteration with ambiguity set size optimization. Behzadian et al. [3] propose an iterative algorithm that optimizes the weights of $\ell_1$ and $\ell_\infty$ 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 probabilities to mitigate the conservativeness of learned policies [4, 15, 19, 29, 30]. 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 [2, 20, 40, 53] 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 [15, 40], resulting in conservative policies.

Finally, a large number of works [12, 16, 36, 44, 49, 50, 52] have proposed RL algorithms that use measures like Conditional Value at Risk, Entropic risk measure amongst other risk measures. However, we note that these works use risk measures to obtain robustness guarantees against aleatoric uncertainty (system uncertainty) and not epistemic uncertainty (model uncertainty), which is the focus of our work. Since these works optimize a completely different objective, we do not compare our framework against theirs. Robust RL work [2, 14, 20, 28, 32, 55] proposes other robust measures for handling uncertainty in transition probabilities; 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) [37, 38]. An MDP is a tuple $\langle S, A, P, R, p_0, \gamma \rangle$ that consists of (a) a set of states $S = \{1, 2, \ldots, S\}$, (b) a set of actions $A = \{1, 2, \ldots, A\}$, (c) a deterministic reward function $R: S \times A \times S \rightarrow \mathbb{R}$, (d) a transition probability function $P: S \times A \rightarrow \Delta^S$, (e) an initial state distribution $p_0 \in \Delta^S$, where $\Delta^S$ represents the $S$-dimensional probability simplex, and (f) a
We assume a batch reinforcement learning setting where $\text{VaR}_\alpha$ is lower-semicontinuous. The optimization in (1) is equivalent to

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

where $\text{VaR}_\alpha$ of a bounded random variable $\tilde{X}$ with a CDF function $F : [0, 1]$ is defined as

$$\text{VaR}_\alpha[\tilde{X}] = \text{VaR}_\alpha[\tilde{X}] = \sup \{ t \in [0, 1] : \Pr[\tilde{X} \geq t] \geq 1 - \alpha \}.$$
We illustrate the conservativeness of BCR

We use the shorthand

Although RMDPs have been used to solve the percentile criterion [3], the quality of the robust policies To optimize the percentile criterion, an SA-rectangular ambiguity set $P$ is constructed such that it contains the true model with high probability. The optimal policy of a Robust MDP $\pi^*$ maximizes the returns of the worst model in the ambiguity set: $\pi^* = \arg\max_{\pi \in \Pi} \min_{P \in P} \rho(\pi, P)$. General RMDPs are NP-hard to solve [53], but they are tractable for broad classes of ambiguity sets. The simplest such type is the SA-rectangular ambiguity set [33, 53], defined as

$$P = \{ P \in (\Delta^S)^{S \times A} \mid p_{s,a} \in P_{s,a}, \forall s \in S, \forall a \in A \},$$

for a given $P_{s,a} \subseteq \Delta^S, s \in S, a \in A$. SA-rectangular ambiguity sets [3, 40] assume that the transition probabilities corresponding to each state-action pair are independent. Similarly to MDPs, the optimal robust value function $v^* \in \mathbb{R}^S$ for an SA-rectangular RMDP is the unique fixed point of the robust Bellman optimality operator $T: \mathbb{R}^S \to \mathbb{R}^S$ defined as

$$(Tv)_s = \max_{a \in A} \min_{p_{s,a} \in P_{s,a}} \mathbb{E}_{r_{s,a} + \gamma v} \left[ r_{s,a} + \psi(b) \right].$$

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

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

Although RMDPs have been used to 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 [40]. SA-rectangular ambiguity sets are most commonly studied; thus we focus our attention on SA-rectangular Robust MDPs. We investigate whether Bayesian credibly 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, 40]. Given a state $s$ and an action $a$, let $\psi_{s,a}$ represent the size of the BCR$_{\alpha}$ ambiguity sets; $\mathcal{P}^{\text{BCR}}_{s,a}$ and $\tilde{p}_{s,a}$ represent the mean transition probabilities. The set $\mathcal{P}^{\text{BCR}}_{s,a}$ is constructed as

$$\mathcal{P}^{\text{BCR}}_{s,a} = \mathcal{P}_{s,a}(b, \psi, q) = \{ p_{s,a} \in \Delta^S \mid \|p_{s,a} - \tilde{p}_{s,a}\|_q b \leq \psi_{s,a} \}, \tag{4}$$

where $q \in \{1, \infty\}$ represents the norm of the weighted ball in [4] and $b \in \mathbb{R}^S_+$ is a weight vector. Here, $b$ 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., $\Pr \left( \tilde{p}_{s,a} \in \mathcal{P}_{s,a}(b, \psi, q) \right) \geq 1 - \alpha$. We refer to BCR$_{\alpha}$ ambiguity sets with non-uniform weights as weighted BCR$_{\alpha}$ ambiguity sets. We refer to the Robust Bellman optimality operator with BCR$_{\alpha}$ ambiguity sets $T^{\text{BCR}}_{\alpha}$ as the BCR$_{\alpha}$ Bellman optimality operator, and to RMDPs with BCR ambiguity sets as BCR RMDPs. For any $\delta \in (0, 0.5)$, setting the confidence level $\alpha$ in $T^{\text{BCR}}_{\alpha}$ to $\gamma / \delta$ 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$_{\alpha}$ RMDPs can be sub-optimal for optimizing the percentile criterion.

We use the shorthand $w_{s,a}$ for any $s \in S, a \in A$ to denote the vector of values associated with value $v \in \mathbb{R}^S$ and the one-step transition from state $s$ and action $a$, i.e., $w_{s,a} = r_{s,a} + \gamma v$. We use $\rho_{s,a} \in \mathbb{R}^s$ and $\Sigma_{s,a} \in \mathbb{R}^{S \times S}$ for any $s \in S, a \in A$ to represent the empirical mean and covariance of transition probabilities $p_{s,a}$ estimated from $D$. We use tilde to indicate that it is a random variable. 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 $\theta$ and variance $1$. The $Z$-Minkowski norm $\|x\|_Z$ for a vector $x$ given some positive-definite matrix $Z$ is defined as $\|x\|_Z = \sqrt{x^T Z^{-1} x}$.

We illustrate the conservativeness of BCR$_{\alpha}$ ambiguity sets with the following example.
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. For the sake of simplicity, we assume that it is only possible to transition to state \( s_1, s_2 \) and \( s_3 \) from state \( s_0 \). The transition probability of \( p_{s_0,a_0} \) is uncertain and distributed as a Dirichlet distribution \( \tilde{p}_{s_0,a_0} \sim \text{Dir}(10, 10, 1) \) with mean \([0.48, 0.48, 0.04]\). The rewards for transitions from state 0 are given by \( r_{s_0,a_0} = [0.25, 0.25, -1] \).

We wish to optimize the percentile criterion with confidence level \( \delta = 0.2 \). Following the sampling procedure proposed by Russel and Petrik \[40\] to construct a uniformly weighted BCR\(_{\alpha}\) ambiguity set for \( p_{s_0,a_0} \) with 100 posterior samples, yields an ambiguity set \( \mathcal{P}^{\text{BCR}_{\alpha}}_{s_0,a_0} = \{ p \in \Delta^S \| p - \tilde{p}_{s_0,a_0} \|_1 \leq 0.277 \} \). In this case, the reward estimate against the worst model in the ambiguity set \( p = [0.50, 0.32, 0.18] \) is \( \rho^{\text{BCR}_{\alpha}} = 0.025 \). Since we have a single non-terminating state in the MDP, the percentile returns are given by \( \rho^{\text{VaR}_{\alpha}} = \text{VaR}_{0.2}[\tilde{p}^T_{s_0,a}r_{s_0,a}] \). Computing \( \rho^{\text{VaR}_{\alpha}} \) for Dirichlet distribution \( \text{Dir}(10, 10, 1) \), we get \( \rho^{\text{VaR}_{\alpha}} = 0.17 > \rho^{\text{BCR}_{\alpha}} \). Thus, this example shows that BCR\(_{\alpha}\) ambiguity sets can be unnecessarily large and, thus, result in conservative policies.

3 VaR Framework

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

\[
(T^{\text{VaR}_{\alpha}}v)_s = \max_{a \in A} \text{VaR}_{\alpha} \left[ p^T_{s,a}(r_{s,a} + \gamma v) \right].
\]

For each state \( s \), \( T^{\text{VaR}_{\alpha}} \) maximizes the value corresponding to the worst \( \alpha \)-percentile model. In contrast to the BCR\(_{\alpha}\) Bellman optimality operator \( T^{\text{BCR}_{\alpha}} \), computing \( T^{\text{VaR}_{\alpha}} \) 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.

Proposition 3.1 (Validity). The following properties of \( T^{\text{VaR}_{\alpha}} \) hold for all value functions \( u, v \in \mathbb{R}^S \).

1. The operator \( T^{\text{VaR}_{\alpha}} \) is contraction mapping on \( \mathbb{R}^S \): \( \| T^{\text{VaR}_{\alpha}}u - T^{\text{VaR}_{\alpha}}v \|_\infty \leq \gamma \| u - v \|_\infty \).
2. The operator \( T^{\text{VaR}_{\alpha}} \) is monotone: \( u \geq v \Rightarrow T^{\text{VaR}_{\alpha}}u \geq T^{\text{VaR}_{\alpha}}v \).
3. The equality \( T^{\text{VaR}_{\alpha}}v = v \) has a unique solution.

Figure 1: Figure (a) (left) compares the asymptotic radius of \( \Sigma^{-1} \)-Minkowski norm BCR\(_{\alpha}\) ambiguity sets to VaR\(_{\alpha}\) ambiguity sets, where \( \Sigma \) is the covariance matrix. The size of the BCR\(_{\alpha}\) ambiguity sets significantly grows with the number of states. Figure (b) and Figure (c) (right) compare the asymptotic forms of BCR\(_{\alpha}\) and VaR\(_{\alpha}\) ambiguity sets under high and low uncertainty in \( P \) at confidence level \( \delta = 0.2 \) respectively. The grey dots represent the transition probabilities samples from a Dirichlet distribution. The sizes of BCR\(_{\alpha}\) and VaR\(_{\alpha}\) ambiguity sets increase with an increase in the variance of the Dirichlet distribution, however, the VaR\(_{\alpha}\) ambiguity sets are consistently smaller than the BCR\(_{\alpha}\) ambiguity sets.
See Appendix B.1 for the proof. Proposition 3.1 formally proves that $\mathcal{V}_{\text{VaR}}$ is a valid Bellman operator, i.e., it is a contraction mapping, monotone, and has a unique fixed point $\hat{\mathcal{V}} = (\mathcal{V}_{\text{VaR}})\hat{\mathcal{V}}$. Here on, we will refer to the policy $\hat{\mathcal{V}}$ corresponding to the fixed point value $\hat{\mathcal{V}}$ as the VaR policy.

We now show that the VaR Bellman optimality operator $\mathcal{V}_{\text{VaR}}$ optimizes a lower bound on the percentile criterion. Given a policy $\pi$, a state $s \in S$, and transition probabilities $P$, let

$$
(T_{\text{VaR}}\pi v)_s = \sum_{a \in A} \pi(s, a) \mathbb{E}_{P_{s,a}}[r_{s,a} + \gamma v],
$$

represent the Bellman evaluation operator for transition probabilities $P$. Furthermore, let

$$
(T_{\text{VaR}}\alpha v)_s = \sum_{a \in A} \pi(s, a) \mathbb{E}_{P_{s,a}}[\hat{p}_{s,a}^T (r_{s,a} + \gamma v)],
$$

represent the VaR Bellman evaluation operator for random transition probabilities $\hat{P}$. We use $\hat{v}^\alpha$ to denote the fixed point of $T_{\text{VaR}}\alpha$. Furthermore, we use $\hat{v}^\alpha$ to represent the random fixed point of $T_{\hat{P}}\alpha$, which is computed using a realization $\hat{P}$ of the posterior distribution of the transition probabilities $P$.

**Proposition 3.2** (Lower Bound Percentile Criteron). For any $\delta \in (0, 0.5)$, if we set the confidence level $\alpha = \delta/s$, then for every policy $\pi \in \Pi : \Pr P[\hat{v}^\alpha \leq \bar{\mathcal{V}}^\alpha \ s | D] \geq 1 - \delta$, where $P$ is a realization of the posterior distribution of the transition probabilities $P$ conditioned on observed transitions $D$.

See Appendix B.2 for the proof. Proposition 3.2 shows that for any policy $\pi$ and state $s$, the VaR value at state $s$, $\bar{\mathcal{V}}^\alpha(s)$ lower bounds the true value $\hat{v}^\alpha(s)$ with high confidence. Comparing Proposition 3.2 with the definition of the percentile-criterion in [1], we see that the percentile-criterion requires confidence guarantees only on the returns computed from the initial states, whereas the equation in Proposition 3.2 provides confidence guarantees on the value of every state. Therefore, for any policy $\pi$, the value $p_{0}^T \hat{v}^\alpha$ is a lower bound on the percentile-criterion objective $\mathbb{E}_P[p(\pi, P)]$. Since $T_{\text{VaR}}\alpha$ finds a policy $\pi$ that maximizes the value $p_{0}^T \hat{v}^\alpha$, it follows [37] that $T_{\text{VaR}}\alpha$ optimizes a lower bound on the percentile criterion in [1].

**Proposition 3.3.** Suppose that $\hat{p}_{s,a}$ for any state $s$ and action $a$, is a multivariate sub-Gaussian with mean $\bar{p}_{s,a}$ and covariance factor $\Sigma_{s,a}$, i.e., $\mathbb{E} \left[ \exp (\lambda (\hat{p}_{s,a} - \bar{p}_{s,a})^T w) \right] \leq \exp (\lambda^2 \Sigma_{s,a}w^T)/, \forall \lambda \in \mathbb{R}, \forall w \in \mathbb{R}^s$. Then, for any state $s \in S$, $T_{\text{VaR}}\alpha$ satisfies

$$
(T_{\text{VaR}}\alpha v)_s \geq \max_{a \in A} \left( \bar{p}_{s,a}^T w_{s,a} - 2 \ln (1/\alpha) \sqrt{w_{s,a} \Sigma_{s,a} \bar{p}_{s,a}} \right).
$$

As a special case, when $\hat{p}_{s,a}$ is normally distributed $\hat{p}_{s,a} \sim N(\bar{p}_{s,a}, \Sigma_{s,a})$, then $T_{\text{VaR}}\alpha$ for any state $s \in S$ can be expressed as

$$
(T_{\text{VaR}}\alpha v)_s = \max_{a \in A} \left( \bar{p}_{s,a}^T w_{s,a} - \frac{1}{\sqrt{2 \pi \Sigma_{s,a}}} \left( \bar{p}_{s,a}^T w_{s,a} - \bar{p}_{s,a} \right) \right).
$$

See Appendix B.3 for the proof. Proposition 3.3 shows that by assuming that the transition probabilities are sub-Gaussian, we can easily compute a lower bound of the VaR Bellman update $(T_{\text{VaR}}\alpha v)$ for a given value function using only the mean and the covariance matrix of $\hat{P}$. In the special case where $\hat{P}$ is normally distributed, we can compute the VaR Bellman optimality operator $T_{\text{VaR}}\alpha(v)$ exactly.

3.1 Performance Guarantees

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

**Theorem 3.4** (Performance). Let $\hat{v}$ be the fixed point of the VaR Bellman optimality operator $T_{\text{VaR}}\alpha$, and $\pi^*$ be the optimal policy in [1]. Let $\hat{\rho}^* = \mathbb{E}[\rho(\pi^*, P)]$ denote the optimal percentile returns and $\tilde{\rho} = \hat{P}^T \hat{v}$ denote the lower bound on the percentile returns computed using the Bellman operator $T_{\text{VaR}}\alpha$ with $\alpha = \delta/s$. Then for each $\delta \in (0, 0.5)$:

$$
\hat{\rho}^* - \rho \leq \frac{1}{1 - \gamma} \max_{s \in S} \max_{a \in A} \left( \mathbb{E}[R_{1 - \delta}] \right), \quad \text{ VaR}_{\alpha} \left[ \hat{p}_{s,a}^T w_{s,a} \right] - \mathbb{E}[\rho(\pi^*, P)] \right).
$$

(6)
We provide a detailed description of the VaR framework. The loss varies proportionally to the maximum difference between the \( \delta/s \) and \( 1 - (1 - \delta/s) \) percentile of the one-step Bellman update for the optimal robust value function \( \hat{v} \). As expected, the VaR framework performs better when the uncertainty in the transition probabilities is small.

**Theorem 3.5 (Asymptotic Performance).** Suppose that the normality assumptions on the posterior distribution \( \hat{P} \) in (\ref{eq:posterior}) are satisfied. For any \( \delta \in (0, 0.5) \), set \( \alpha = \delta/s \) in \( T_{\text{VaR}_\alpha} \). For any state \( s \) and action \( a \), let \( I(p_{s,a}^*) \) be the Fisher information matrix corresponding to the true transition probabilities \( p_{s,a} \). Furthermore, let \( \sigma_{\text{max}}^2 = \max_{s,a} \omega_{s,a}^T I(p_{s,a}^*)^{-1} \omega_{s,a} \) be the maximum over state-action pairs of the asymptotic variance of the returns estimate \( \hat{p}_{s,a}^T \omega_{s,a} \). Then the asymptotic performance of the VaR framework \( \hat{p} \) w.r.t. the optimal percentile returns \( \rho^* \) satisfies

\[
\lim_{N \to \infty} \sqrt{N} (\rho^* - \hat{\rho}) \leq \frac{1}{1 - \gamma} \left( 2\Phi^{-1}(1 - \delta/s)\sigma_{\text{max}} \right) \leq \frac{1}{1 - \gamma} \sqrt{8 \ln(\delta/s)} \sigma_{\text{max}}.
\]

See Appendix B.5 for the proof. Theorem 3.5 shows that almost surely, the asymptotic loss in performance of the VaR framework convergence to 0, i.e., \( \lim_{N \to \infty} (\rho^* - \hat{\rho}) = 0 \).

### 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 (\( T_{\text{VaR}_\alpha} \)) for any given policy \( \pi \) and state \( s \) with high confidence \( 1 - \zeta \).

**Algorithm 3.1: Generalized VaR Value Iteration Algorithm**

**Input:** Confidence \( \alpha \), Posterior distribution \( f \), target Bellman residual \( \varepsilon \)

**Output:** Robust policy \( \pi \), lower bound \( \nu \)

1. Initialize \( \pi \) with arbitrary \( \pi_0 \), robust value-function \( v \) with arbitrary \( v_0, k = 0 \);
2. Sample \( N \) models \( \{\hat{P}(\omega_1), \hat{P}(\omega_2), \ldots, \hat{P}(\omega_N)\} \) from posterior \( f \);
3. repeat
   4. for \( s \) ← 1 to \( S \) do
      5. Initialize \( q ← [] \);
      6. for \( a \) ← 1 to \( A \) do
         7. \( q[a] ← \text{VaR}_{\alpha}[\hat{p}_{s,a}^T (r_{s,a} + \gamma v_k)] \) OR;
         8. \( q[a] ← \hat{p}_{s,a}^T \omega_{s,a} - \Phi^{-1}(1 - \alpha)\sqrt{\omega_{s,a}^T \hat{\Sigma}_{s,a} \omega_{s,a}} \) (under normal Assumptions);
      9. end
      10. \( v_k(s) ← \max(q); \quad \pi_k(s) ← \arg \max(q); \)
   11. end
   12. \( k ← k + 1; \)
5. until \( \|v_k - v_{k-1}\|_\infty ≤ \varepsilon \);
6. return \( \pi_k, v_k \);

In each iteration of Algorithm 3.1, we compute the one-step VaR Bellman update \( T_{\text{VaR}_\alpha}(v) \) using the current value function \( v \). When \( \hat{P} \) is not normally distributed, we use the Quick Select algorithm \( \text{Quick Select} \) to efficiently compute the empirical estimate of the \( \alpha \)-percentile of returns for any state \( s \) and action \( a \), i.e., \( \text{VaR}_{\alpha}[\hat{p}_{s,a}^T (r_{s,a} + \gamma v_k)] \) in \( O(SAN) \) time (Proposition 3.6). On the other hand, when \( \hat{P} \) is normally distributed, we compute the VaR Bellman update \( T_{\text{VaR}_\alpha}(v) \) using the empirical estimate of mean \( \{\hat{p}_{s,a}^T \} \) \( s,a \in S,A \) and covariance \( \{\hat{\Sigma}_{s,a}\} \) \( s,a \in S,A \) of the transition probabilities derived from the data \( D \) (Proposition 3.5). We repeat these steps until convergence.

**Proposition 3.6 (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 \( O(SAN) \), where \( N \) is the number of samples of the posterior samples of \( \hat{P} \).

**Proposition 3.7 (Empirical Error Bound).** For any state \( s \), action \( a \) and value function \( v \), let \( \text{VaR}_{\alpha}[\hat{p}_{s,a}^T \omega_{s,a}] \) represent the empirical estimate of \( \alpha \)-percentile of returns \( \text{VaR}_{\alpha}[\hat{p}_{s,a}^T \omega_{s,a}] \) and \( \Phi_f \) represent the cumulative density function (CDF) of the random estimate of returns \( \hat{p}_{s,a}^T \omega_{s,a} \). Suppose that \( \Phi_f \) is differentiable at the point \( \text{VaR}_{\alpha}[\hat{p}_{s,a}^T \omega_{s,a}] \) and let \( m = \Phi'(\text{VaR}_{\alpha}[\hat{p}_{s,a}^T \omega_{s,a}]) \)
represents the density of estimate of returns at point \( \text{VaR}_\alpha [\hat{p}_{s,a}^T w_{s,a}] \). Let \( N^* \) be the number of posterior samples required to obtain empirical error \( \varepsilon \in \mathbb{R} \), with confidence \( 1 - \zeta \), i.e., \( \Pr \left[ \text{VaR}_\alpha [\hat{p}_{s,a}^T w_{s,a}] - \text{VaR}_\alpha [\hat{p}_{s,a}^T w_{s,a}] \geq \varepsilon \right] \leq \zeta \). Then, \( \lim_{\varepsilon \to 0} N^* \varepsilon^2 = \ln(2)/2m^2 \).

**Proposition 3.8.** Define the empirical \( \text{VaR} \) Bellman optimality operator \( \mathcal{T}_{\text{VaR}_a} \) for any value \( v \in \mathbb{R}^S \) and state \( s \in S \) (as \( (\mathcal{T}_{\text{VaR}_a} v)_s = \max_{a \in A} \text{VaR}_\alpha [\hat{p}_{s,a}^T w_{s,a}] \)). Let \( \hat{v} \in \mathbb{R}^S \) and \( \hat{f} \in \mathbb{R}^S \) represent the fixed points of \( \mathcal{T}_{\text{VaR}_a} \) and \( \mathcal{T}_{\hat{f}} \) respectively. Furthermore, for any \( k \in \mathbb{Z} \), let \( f^k \) represent the value function at the \( k \)th iteration in Algorithm 3.1. If \( k \geq (1 - \gamma)^{-1} \ln \left( \frac{R_{\max}}{\varepsilon (1 - \gamma)} \right) \) where \( R_{\max} = \max_{s \in S, a \in A} \max_{s', a' \in S} R(s, a, s') \), then,

\[
\| \hat{f} - f^k \|_\infty \leq \varepsilon .
\]

Moreover, the error in the value function at the \( k \)th iteration \( (f^k) \) of Algorithm 3.1 with respect to the optimal value function \( \hat{v} \) is given by

\[
\| \hat{v} - f^k \|_\infty \leq \varepsilon + \frac{1}{1 - \gamma} \| \mathcal{T}_{\text{VaR}_a} \hat{f} - \mathcal{T}_{\text{VaR}_a} \hat{f} \|_\infty .
\]

See Appendix B.9 for the proof.

### 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\( \alpha \) Robust MDPs. First, we derive the robust form of the VaR framework and show that in contrast to the BCR\( \alpha \) 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 compare the asymptotic radii of the BCR\( \alpha \) ambiguity sets and the VaR ambiguity sets implicitly constructed by \( \mathcal{T}_{\text{VaR}_a} \). For any given confidence level \( \alpha \), the radius of the VaR ambiguity sets are asymptotically smaller than that of BCR ambiguity sets (Theorem 4.2). Precisely, the ratio of the radii 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 \( N(0, 1) \). This implies that there exist directions in which the BCR ambiguity sets are at least \( \Omega(\sqrt{S}) \) larger than VaR ambiguity sets. Thus, we prove that VaR framework is better suited for optimizing the percentile criterion than RMDPs with BCR\( \alpha \) ambiguity sets.

For any value function \( v \), define the VaR ambiguity set \( \mathcal{P}^{\text{VaR}, v} \) as

\[
\mathcal{P}^{\text{VaR}, v} = \times_{s \in S} \mathcal{P}^{\text{VaR}, v}_{s,a} \quad \text{where} \quad \mathcal{P}^{\text{VaR}, v}_{s,a} = \{ p_{s,a} \in \Delta^S | p_{s,a}^T v \geq \text{VaR}_\alpha [\hat{p}_{s,a}^T v] \}. \tag{7}
\]

**Proposition 4.1 (Equivalence).** The optimal VaR policy \( \hat{\pi} \in \Pi^D \) solves

\[
\max_{\pi \in \Pi^D} \min_{p \in \mathcal{P}^{\text{VaR}, v}} \rho(\pi, P), \tag{8}
\]

where \( \hat{v} \in \mathbb{R}^S \) is the fixed point of the VaR Bellman operator \( \mathcal{T}_{\text{VaR}_a} \cdot \hat{v} = (\mathcal{T}_{\text{VaR}_a} \hat{v}) \).

See Appendix B.9 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-space \( \{ p_{s,a} \in \mathbb{R}^S : p_{s,a}^T v^\pi \geq \text{VaR}_\alpha [\hat{p}_{s,a}^T v^\pi] \} \) dependent on the value function \( v^\pi \) of the current policy \( \pi \). In contrast, BCR\( \alpha \) ambiguity sets are independent of any policy or value function and are constructed such that they provide high-confidence guarantees on returns of all policies simultaneously. As a result, BCR\( \alpha \) ambiguity sets tend to be unnecessarily large.

We now compute the ratio of the asymptotic radii of BCR\( \alpha \) ambiguity sets and VaR ambiguity sets.

**Theorem 4.2 (Asymptotic Radii of VaR Ambiguity Sets).** For any state \( s \) and action \( a \), let \( I(p_{s,a}^*) \) be the Fisher information corresponding to \( p_{s,a}^* \). Suppose that the normality assumptions on the posterior distribution \( \hat{p} \) in (4) are satisfied. Then,

\[
\lim_{N \to \infty} \sqrt{N} (p_{s,a}^{\text{VaR}, v} - p_{s,a}^*) = \{ p_{s,a} \in \Delta^S | \| p_{s,a} - p_{s,a}^* \|_{I(p_{s,a}^*)^{-1}} \leq \Phi^{-1}(1 - \alpha) \} - p_{s,a}^* . \tag{9}
\]
We now empirically analyze the robustness of the VaR framework in three different domains. Specifically, we investigate if the VaR framework learns robust policies that are less conservative than BCR Robust MDPs.

**Riverswim**: The Riverswim MDP [46] 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 [25] 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 [56] models the classical inventory management problem and consists of 30 states and 30 actions. States represent the inventory level and actions represent the amount of inventory to be purchased. The sale price, holding cost, and purchase costs are 3.99, 0.03, and 2.219. The demand is normally distributed with mean=3.99 and standard deviation 1.0.

**Implementation details**: For each domain in our experiments, we sample a dataset $D$ 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 $D$, assuming Dirichlet priors over the model parameters. Using MCMC sampling, we construct two datasets $D_1$ and $D_2$ containing $M$ and $K$ transition probability models, respectively.

We construct $L$ train datasets by randomly sampling 80% of the models from $D_1$ each time. We use $D_2$ as our test dataset. For any given confidence level $\delta$, we train one RL agent per train dataset and method. For evaluation, we consider two instances of the VaR framework: one (denoted by VaRN) that assumes that $P$ is a multivariate normal, and another (denoted by VaR) that does not assume any structure over $P$. We use seven baseline methods for evaluating the robustness of our framework. They are: Naive Hoeffding [35], Optimized Hoeffding (Opt Hoeffding) [3], Soft-Robust [5], and BCR Robust MDPs with weighted $\ell_1$ ambiguity sets ($WBCR_\ell_1$) [3], weighted $\ell_\infty$ ambiguity sets ($WBCR_\ell_\infty$) [3], unweighted $\ell_1$ ambiguity sets ($BCR_\ell_1$) [40] and unweighted $\ell_\infty$ ambiguity sets ($BCR_\ell_\infty$) ambiguity [40] sets. See Appendix C in the appendix for more details.
We report the 95% confidence interval of the robust performance (δ-percentile of expected returns) of the VaR framework on the test dataset with that of other baselines for different values of δ.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Riverswim</th>
<th>Inventory</th>
<th>Population-Small</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>VaR</td>
<td>68.54 ± 5.08</td>
<td>437.95 ± 0.74</td>
<td>-3102.48 ± 429.7</td>
<td>-4576.87 ± 147.3</td>
</tr>
<tr>
<td>VaRN</td>
<td>67.27 ± 0.0</td>
<td>452.78 ± 0.02</td>
<td>-4005.53 ± 8.76</td>
<td>-4570.17 ± 38.84</td>
</tr>
<tr>
<td>BCR ℓ₁</td>
<td>67.27 ± 0.0</td>
<td>369.67 ± 0.0</td>
<td>-5614.95 ± 80.28</td>
<td>-6013.21 ± 1177.94</td>
</tr>
<tr>
<td>BCR ℓ∞</td>
<td>67.27 ± 0.0</td>
<td>199.41 ± 39.02</td>
<td>-7908.92 ± 41.6</td>
<td>-9033.7 ± 84.28</td>
</tr>
<tr>
<td>WBCR ℓ₁</td>
<td>67.9 ± 3.82</td>
<td>454.1 ± 4.16</td>
<td>-5290.38 ± 1084.26</td>
<td>-5408.01 ± 225.2</td>
</tr>
<tr>
<td>WBCR ℓ∞</td>
<td>67.27 ± 0.0</td>
<td>199.4 ± 39.02</td>
<td>-7712.43 ± 55.96</td>
<td>-8377.64 ± 126.24</td>
</tr>
<tr>
<td>Soft-Robust</td>
<td>61.79 ± 1.46</td>
<td>460.6 ± 0.0</td>
<td>-3647.18 ± 94.62</td>
<td>-6932.86 ± 154.16</td>
</tr>
<tr>
<td>Naive Hoeffding</td>
<td>51.52 ± 6.06</td>
<td>-0.0 ± 0.0</td>
<td>-8647.7 ± 59.5</td>
<td>-9127.14 ± 140.98</td>
</tr>
<tr>
<td>Opt Hoeffding</td>
<td>50.76 ± 4.56</td>
<td>-0.0 ± 0.0</td>
<td>-8640.48 ± 2.34</td>
<td>-9163.64 ± 13.62</td>
</tr>
</tbody>
</table>

Table 1: shows the 95% confidence interval of the robust (percentile) returns achieved by VaR, VaRN, BCR ℓ₁, BCR ℓ∞, WBCR ℓ₁, WBCR, Soft Robust, Worst RMDP, Naive Hoeffding and Opt Hoeffding agents at δ = 0.05 in Riverswim, Inventory, Population-Small, and Population domain.

**Experimental Results** Table 1 summarizes the performance of the VaR framework and the baselines for confidence level δ = 0.05 (Table 2 and Table 3 in the appendix summarizes the results for δ = 0.15 and δ = 0.3 respectively.). We observe that for confidence level δ = 0.05, the VaR framework outperforms the baseline methods in terms of mean robust performance in most domains. On the other hand, for δ = 0.15, the VaR framework outperforms baselines in Population and Population-Small domains and has comparable performance to the Soft-Robust method in the Inventory domain. However, at δ = 0.3 we observe that the VaR framework has lower robust performance relative to the the Soft-Robust method in Riverswim and Population domains. We conjecture that this is because the Soft-Robust method optimizes the policy to maximize the mean of expected returns and is therefore able to perform better in cases where lower levels of robustness are required. However, we note that in contrast to our method, the Soft-Robust method does not provide probabilistic guarantees against worst-case scenarios.

Furthermore, as expected, we find that in many cases, the robust performance of BCR Robust MDPs with span-optimized (weighted) ambiguity sets (WBCR ℓ₁, WBCR ℓ∞) is relatively higher than the robust performance of Robust MDPs with unweighted BCR_α ambiguity sets (BCR ℓ₁, BCR ℓ∞). However, we find that even Robust MDPs with span-optimized BCR_α ambiguity sets are generally unable to outperform the robust performance of our VaR_α framework.

Figure 2 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 Conclusion and Future Work**

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 [19,29,30]. However, due to the non-convex nature of the percentile-criterion [3,40], 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 [11,27] 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 will be valuable to rigorously test the VaR_α framework in large RL domains and compare its performance against other Robust RL methods [14,19,30].

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 ambiguity sets whose asymptotic radius is smaller than that of any Bayesian credible region, and therefore, 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.
Acknowledgments

The work in the paper was supported, in part, by NSF grants 2144601 and 1815275.

References


A  Additional theoretical results

Definition A.1 (Translation Subvariance). The operator $\mathcal{T} : \mathbb{R}^S \to \mathbb{R}^S$ satisfies the translation subvariance property if for all vector $v \in \mathbb{R}^S$ and scalar $c$, there exists $\gamma \in [0, 1)$ that satisfies

$$\mathcal{T}(v + c1) = (\mathcal{T}v) + \gamma c1 .$$

Definition A.2 (Monotonicity). The operator $\mathcal{T} : \mathbb{R}^S \to \mathbb{R}^S$ satisfies the monotonicity property if for all $v \in \mathbb{R}^S$ and $u \in \mathbb{R}^S$ such that $v \preceq u$, $\mathcal{T}$ satisfies

$$\mathcal{T}v \preceq \mathcal{T}u .$$

Lemma A.3 (Contraction Mapping [6]). The operator $\mathcal{T} : \mathbb{R}^S \to \mathbb{R}^S$ is a contraction mapping if it satisfies monotonicity and translation subvariance properties, i.e., there exists $\gamma \in [0, 1)$ such that for all $u, v \in \mathbb{R}^S$, $\mathcal{T}$ satisfies

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

The proof of Lemma A.3 follows directly from Proposition 2.1.3 in [6]. We re-derive the proof for the sake of completeness.

Proof. Denote

$$c = \max_{s \in S} |u_s - v_s| .$$

Therefore for all $s \in S$,

$$u_s - c \leq v_s \leq u_s + c .$$

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

$$(\mathcal{T}u)_s - \gamma c \leq (\mathcal{T}v)_s \leq (\mathcal{T}u)_s + \gamma c .$$

It follows that for all $s \in S$,

$$|(\mathcal{T}v)_s - (\mathcal{T}u)_s| \leq \gamma c ,$$

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

Given a policy $\pi$, a state $s \in S$, and transition probabilities $P$, let

$$(T_{\pi P} v)_a = \sum_{a \in A} \pi(s, a) P^T_{s, a} w_{s, a}$$

and

$$(T_{\pi \text{VaR}_\alpha} v)_a = \sum_{a \in A} \pi(s, a) \text{VaR}_\alpha \left( P^T_{s, a} w_{s, a} \right) ,$$

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

Lemma A.4. For any policy $\pi$, let $\hat{v}^\pi$ and $v^\pi$ be the fixed point of the VaR policy evaluation operator $T_{\text{VaR}_\alpha}$ and Bellman policy evaluation operator $T_{\pi P}$. If the VaR Bellman policy evaluation operator $T_{\text{VaR}_\alpha}$ dominates the Bellman policy evaluation operator $T_{\pi P}$ at $\tilde{v}^\pi$, i.e., $T_{\text{VaR}_\alpha} \tilde{v}^\pi \preceq T_{\pi P} \tilde{v}^\pi$, then, the fixed point of the VaR Bellman evaluation operator $T_{\text{VaR}_\alpha}$ dominates the fixed point of the Bellman evaluation operator $T_{\pi P}$, i.e., $\hat{v}^\pi \preceq v^\pi$.

We note that in contrast to the Bellman policy evaluation operator $T_{\pi P}$, the VaR Bellman policy evaluation operator $T_{\text{VaR}_\alpha}$ is a function of the random variable $\hat{P}$ and is not dependent on the transition probabilities $P$ assumed in this setting.

Using the assumption $T_{\text{VaR}_\alpha} \hat{v}^\pi \preceq T_{\pi P} \hat{v}^\pi$, and from $\hat{v}^\pi = T_{\text{VaR}_\alpha} \hat{v}^\pi$ and $v^\pi = T_{\pi P} v^\pi$, we get by algebraic manipulations:
Proof.

\[ \hat{v}^\pi - v^\pi = T^\pi_{\text{VaR}, \alpha} \hat{v}^\pi - T^\pi_{\text{VaR}, \alpha} v^\pi \preceq T^\pi_{\text{VaR}, \alpha} \hat{v}^\pi - T^\pi_{\text{VaR}, \alpha} v^\pi \preceq \gamma P_\pi (\hat{v}^\pi - v^\pi) . \]

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

\[ (I - \gamma P_\pi) (\hat{v}^\pi - v^\pi) \preceq 0 . \]

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

\[ \hat{v}^\pi - v^\pi \preceq (I - \gamma P_\pi)^{-1} 0 = 0 . \]

which proves the result. \( \square \)

B Proofs

B.1 Proof of Proposition 3.1

**Proposition 3.1 (Validity).** The following properties of \( \mathcal{T}_{\text{VaR}, \alpha} \) hold for all value functions \( u, v \in \mathbb{R}^S \).

1. The operator \( \mathcal{T}_{\text{VaR}, \alpha} \) is contraction mapping on \( \mathbb{R}^S \): \( \| \mathcal{T}_{\text{VaR}, \alpha} u - \mathcal{T}_{\text{VaR}, \alpha} v \|_\infty \leq \gamma \| u - v \|_\infty \).  
2. The operator \( \mathcal{T}_{\text{VaR}, \alpha} \) is monotone: \( u \succeq v \Rightarrow \mathcal{T}_{\text{VaR}, \alpha} u \succeq \mathcal{T}_{\text{VaR}, \alpha} v \).
3. The equality \( \mathcal{T}_{\text{VaR}, \alpha} \hat{v} = \hat{v} \) 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 \( \mathcal{T}_{\text{VaR}, \alpha} \) satisfies monotonicity and subvariance property, and therefore, is a contraction mapping.

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

First, we show that \( \mathcal{T}_{\text{VaR}, \alpha} \) satisfies translation subvariance property. Consider any \( c \in \mathbb{R} \) and state \( s \). Then,

\[ (\mathcal{T}_{\text{VaR}, \alpha} (v + c \mathbf{1}))_s = \max_{a \in \mathcal{A}} \text{VaR}_\alpha [p^T_{s,a} (r_{s,a} + \gamma (v + c \mathbf{1})] \]

\[ \overset{(a)}{=} \max_{a \in \mathcal{A}} \text{VaR}_\alpha [p^T_{s,a} (r_{s,a} + \gamma v + \gamma c \mathbf{1})] \]

\[ \overset{(b)}{=} \max_{a \in \mathcal{A}} \text{VaR}_\alpha [p^T_{s,a} (r_{s,a} + \gamma v) + \gamma c] \]

\[ \overset{(c)}{=} \max_{a \in \mathcal{A}} \text{VaR}_\alpha [p^T_{s,a} (r_{s,a} + \gamma v)] + \gamma c \]

\[ \overset{(d)}{=} (\mathcal{T}_{\text{VaR}, \alpha} v)_s + \gamma c . \]

(a) follows from simple algebraic manipulations, (b) follows from \( \gamma c p^T_{s,a} \mathbf{1} = \gamma c \), (c) follows from the translational invariance property of VaR measure \([41]\), and (d) follows the definition of the VaR Bellman operator \( \mathcal{T}_{\text{VaR}, \alpha} \).

Next, we show that \( \mathcal{T}_{\text{VaR}, \alpha} \) satisfies the monotonicity property.

Let \( u \) and \( v \) be any two value functions such that \( v \preceq u \). Consider any state \( s \in S \). Then,

\[ (\mathcal{T}_{\text{VaR}, \alpha} v)_s - (\mathcal{T}_{\text{VaR}, \alpha} u)_s = \max_{a \in \mathcal{A}} \text{VaR}_\alpha [p^T_{s,a} (r_{s,a} + \gamma v)] - \max_{a \in \mathcal{A}} \text{VaR}_\alpha [p^T_{s,a} (r_{s,a} + \gamma u)] \]

\[ \overset{(a)}{\leq} \max_{a \in \mathcal{A}} \{ \text{VaR}_\alpha [p^T_{s,a} (r_{s,a} + \gamma v)] - \text{VaR}_\alpha [p^T_{s,a} (r_{s,a} + \gamma u)]\} \]

\[ \overset{(b)}{\leq} 0 \]

\[ (\mathcal{T}_{\text{VaR}, \alpha} v)_s \leq (\mathcal{T}_{\text{VaR}, \alpha} u)_s \]
Thus, we prove that $T_{\text{VaR}_\alpha}$ is a $\gamma$-contraction mapping and a monotone operator. Since $T_{\text{VaR}_\alpha}$ is a contraction operator on a Banach space and from the Banach fixed point theorem \cite{Banach1922}, it follows that the operator $T_{\text{VaR}_\alpha}$ has a unique solution $v$, i.e., $T_{\text{VaR}_\alpha}v = v$.

\section*{B.2 Proof of Proposition 3.2}

**Proposition 3.2** (Lower Bound Percentile Criterion). For any $\delta \in (0, 0.5)$, if we set the confidence level $\alpha$ in the operator $T_{\text{VaR}_\alpha}$ to $\delta/s$, then for every policy $\pi \in \Pi$ : $\Pr_P [\hat{v}^\pi \leq \hat{v}^\pi | D] \geq 1 - \delta$, where $\hat{P}$ is a realization of the posterior distribution of the transition probabilities $P$ conditioned on observed transitions $D$.

**Proof.** Let $\alpha = \delta/s$. Recall that for any policy $\pi$, state $s \in S$, transition probabilities $P$, the Bellman evaluation operator $T_{\hat{P}}$ and the VaR Bellman evaluation operator $T_{\text{VaR}_\alpha}$, are defined as

\[(T_{\hat{P}}v)_s = \sum_{a \in A} \pi(s,a)P_{s,a}^Tw_{s,a}\] and \[(T_{\text{VaR}_\alpha}v)_s = \sum_{a \in A} \pi(s,a) \text{VaR}_\alpha [p_{s,a}^T \omega_{s,a}] ,\]

respectively.

Let $\hat{v}^\pi$ be the fixed point of $T_{\text{VaR}_\alpha}$ conditioned on observed transitions $D$, and let $\hat{v}^\pi$ be a random variable that represents the fixed point of $T_{\hat{P}}$ for a given realization $\hat{P}$ of the posterior distribution of the transition probabilities $P$ given $D$. Then, applying Lemma A.4 to $T_{\text{VaR}_\alpha}$ and $T_{\hat{P}}$, we have, $\hat{v}^\pi \leq v^\pi$ implies

\[T_{\text{VaR}_\alpha} \hat{v}^\pi \leq T_{\hat{P}} \hat{v}^\pi\]

That is for each state $s$,

\[\text{VaR}_\alpha [\hat{p}_{s,\pi(s)}^T \hat{v}^\pi] \leq \hat{p}_{s,\pi(s)}^T \hat{v}^\pi .\] (11)

Using the equation (11), we can bound the probability that the VaR value function lower bounds the true value.

\[\Pr_{\hat{P}} [\hat{v}^\pi \leq \hat{v}^\pi | D] = \Pr_{\hat{P}} \left[ \forall s \in S : \text{VaR}_\alpha [\hat{p}_{s,\pi(s)}^T \hat{v}^\pi] \leq \hat{p}_{s,\pi(s)}^T \hat{v}^\pi | D \right] .\] (12)

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

\[\Pr_{\hat{P}} [\text{VaR}_\alpha [\hat{p}_{s,a}^T \hat{v}^\pi] \leq \hat{p}_{s,a}^T \hat{v}^\pi | D] \geq 1 - \alpha,\] (13)

Therefore, using union bound and (12) in (13), we can write

\[\Pr_{\hat{P}} [\hat{v}^\pi \geq \hat{v}^\pi | D] \leq \sum_{s \in S} \Pr_{\hat{P}} [\text{VaR}_\alpha [\hat{p}_{s,\pi(s)}^T \hat{v}^\pi] > \hat{p}_{s,\pi(s)}^T \hat{v}^\pi | D] .\]

Thus,

\[\Pr [\hat{v}^\pi \geq \hat{v}^\pi | D] = \sum_{s \in S} \frac{\delta}{s} = S\frac{\delta}{S} = \delta .\]
B.3 Proof of Proposition 3.3

**Proposition 3.3.** Suppose that \( \tilde{p}_{s,a} \) for any state \( s \) and action \( a \), is a multivariate sub-Gaussian with mean \( \bar{p}_{s,a} \) and covariance factor \( \Sigma_{s,a} \), i.e., \( \mathbb{E} \left[ \exp \left( \lambda (\tilde{p}_{s,a} - \bar{p}_{s,a})^T w \right) \right] \leq \exp \left( \lambda^2 w^T \Sigma_{s,a} w / 2 \right) \), \( \forall \lambda \in \mathbb{R}, \forall w \in \mathbb{R}^s \). Then, for any state \( s \in S \), \( T_{\text{VaR}} \) satisfies

\[
(T_{\text{VaR}} v)_s = \max_{a \in A} \left( \tilde{p}_{s,a}^T w_{s,a} - \sqrt{2 \ln(1/\alpha)} \sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}} \right).
\]

As a special case, when \( \tilde{p}_{s,a} \) is normally distributed \( \tilde{p}_{s,a} \sim N(\bar{p}_{s,a}, \Sigma_{s,a}) \), then \( T_{\text{VaR}} \) for any state \( s \in S \) can be expressed as

\[
(T_{\text{VaR}} v)_s = \max_{a \in A} \left( \bar{p}_{s,a}^T w_{s,a} - \phi^{-1}(1-\alpha) \sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}} \right).
\]

**Proof.**

\[
(T_{\text{VaR}} v)_s = \max_{a \in A} \text{VaR}_\alpha[\tilde{p}_{s,a}^T w_{s,a}]
\]

\( a \in A \)

\[
= \max_{a \in A} \sup_t \left\{ t : \Pr \left( \tilde{p}_{s,a}^T w_{s,a} \geq t \right) \geq 1 - \alpha \right\}
\]

\( b \)

\[
= \max_{a \in A} \inf_t \left\{ t : \Pr \left( (\tilde{p}_{s,a} - \bar{p}_{s,a})^T w_{s,a} > (t - \tilde{p}_{s,a}^T w_{s,a}) \right) < 1 - \alpha \right\}
\]

\( c \)

\[
= \max_{a \in A} \inf_t \left\{ t : \Pr \left( \exp((\tilde{p}_{s,a} - \bar{p}_{s,a})^T w_{s,a}) > \exp(t - \tilde{p}_{s,a}^T w_{s,a}) \right) < 1 - \alpha \right\}
\]

\( d \)

\[
= \max_{a \in A} \left\{ t : 1 - \exp \left( \frac{-t + \tilde{p}_{s,a}^T w_{s,a}}{2w_{s,a}^T \Sigma_{s,a} w_{s,a}} \right) < 1 - \alpha \right\}
\]

\( e \)

\[
= \max_{a \in A} \left\{ t : (t - \tilde{p}_{s,a}^T w_{s,a}) < -2 \ln(\alpha) w_{s,a}^T \Sigma_{s,a} w_{s,a} \right\}
\]

\( f \)

\[
= \max_{a \in A} \left\{ t : (t - \tilde{p}_{s,a}^T w_{s,a}) \in \left( -2 \ln(1/\alpha) \sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}}, 2 \ln(1/\alpha) \sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}} \right) \right\}
\]

\( g \)

\[
= \max_{a \in A} \left( \tilde{p}_{s,a}^T w_{s,a} - 2 \ln(1/\alpha) \sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}} \right).
\]

Equality \((a)\) follows from the definition of \( \text{VaR} \). \((b)\) follows from subtracting \( \tilde{p}_{s,a}^T w_{s,a} \) on both sides. \((c)\) follows from taking exponential on both sides. \((d)\) follows from applying the upper-bound given by Chernoff bound for a sub-Gaussian distribution \([7]\) i.e.,

\[
\Pr \left( \exp((\tilde{p}_{s,a} - \bar{p}_{s,a})^T w_{s,a}) \right) \leq \exp \left( \frac{-t + \tilde{p}_{s,a}^T w_{s,a}}{2w_{s,a}^T \Sigma_{s,a} w_{s,a}} \right),
\]

\((e)\) follows from subtracting 1 on both sides and then, taking ln on both sides. \((f)\) follows from simple algebraic manipulations and \((g)\) follows from taking the infimum of the solution interval of \( t \).

Solving for \( t \) in step \((g)\), we get \( t = \text{VaR}_\alpha[\tilde{p}_{s,a}^T w_{s,a}] = \tilde{p}_{s,a}^T w_{s,a} - 2 \ln(1/\alpha) \sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}} \) which proves the stated result.

Next, we derive the \( \text{VaR} \) Bellman optimality update \( T_{\text{VaR}}(v) \) when \( \tilde{p}_{s,a} \forall s \in S, a \in A \) is normally distributed.

Consider the \( \text{VaR} \) Bellman optimality operator defined for any state \( s \) and value function \( v \) as

\[
(T_{\text{VaR}} v)_s = \max_{a \in A} \text{VaR}_\alpha \left[ \tilde{p}_{s,a}^T w_{s,a} \right].
\]

From the theory of multivariate Gaussian distributions \([7]\), we know that, for any state \( s \) and action \( a \), if \( \tilde{p}_{s,a} \) is Gaussian distributed \( N(\bar{p}_{s,a}, \Sigma_{s,a}) \), then \( \tilde{p}_{s,a}^T w_{s,a} \) is also Gaussian distributed \( N(\bar{p}_{s,a}^T w_{s,a}, \bar{p}_{s,a}^T \Sigma_{s,a} w_{s,a}) \). To find the \( \text{VaR}_\alpha[\tilde{p}_{s,a}^T w_{s,a}] \) for any state \( s \) and action \( a \), it is sufficient
to find \( t \) such that \( \Pr(\hat{p}^T_{s,a} w_{s,a} > t) = 1 - \alpha \).

\[
\Pr\left( \frac{(\hat{p} - \tilde{p}_{s,a})^T w_{s,a}}{\sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}}} > \frac{t - \tilde{p}_{s,a}^T w_{s,a}}{\sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}}} \right) = 1 - \alpha
\]

\[
1 - \Phi\left( \frac{t - \tilde{p}_{s,a}^T w_{s,a}}{\sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}}} \right) = 1 - \alpha
\]

\[
t = \tilde{p}_{s,a}^T w_{s,a} + \Phi^{-1}(\alpha) \sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}}
\]

The first equation follows from substituting \( \tilde{p}_{s,a}^T w_{s,a} \) and dividing by \( \sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}} \) 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 \( t = \text{VaR}_\alpha[\hat{p}_{s,a}^T w_{s,a}] = \tilde{p}_{s,a}^T w_{s,a} - \Phi^{-1}(1 - \alpha) \sqrt{w_{s,a}^T \Sigma_{s,a} w_{s,a}} \) in (14), we obtain the stated results.

The normal form of the VaR Bellman optimality operator \( \mathcal{T}_{\text{VaR} \alpha} \) is useful to analyze the asymptotic properties of the VaR Bellman operator.

### B.4 Proof of Theorem 3.4

**Theorem 3.4 (Performance).** Let \( \hat{v} \) be the fixed point of the VaR Bellman optimality operator \( \mathcal{T}_{\text{VaR} \alpha} \) and \( \pi^* \) be the optimal policy in (1). Let \( \rho^* = \text{VaR}_\delta[\rho(\pi^*, \mathcal{P})] \) denote the optimal percentile returns and \( \hat{\rho} = \hat{p}_0^T \hat{v} \) denote the lower bound on the percentile returns computed using the Bellman operator \( \mathcal{T}_{\text{VaR} \alpha} \) with \( \alpha = \delta/s \). Then for each \( \delta \in (0, 0.5) \):

\[
\rho^* - \hat{\rho} \leq \frac{1}{1 - \gamma} \max_{s \in S} \max_{a \in A} \left( \text{VaR}_{1-\frac{\gamma}{1-\delta}}[\hat{p}_{s,a}^T w_{s,a}] - \text{VaR}_\alpha[\hat{p}_{s,a}^T w_{s,a}] \right).
\]  

**Proof.** We denote the optimal policy that optimizes the \( \delta \)-percentile criterion by \( \pi^* \), i.e., \( \pi^* \in \arg \max_{\pi \in \Pi} \text{VaR}_\delta[\rho(\pi, \mathcal{P})] \).

Recall that \( \hat{v} \in \mathbb{R}^S \) is the fixed point of the VaR \( \alpha \) Bellman optimality operator \( \mathcal{T}_{\text{VaR} \alpha} \), i.e., \( \hat{v} = \mathcal{T}_{\text{VaR} \alpha} \hat{v} \) and \( \hat{\rho} = p_0^T \hat{v} \) is the returns of the corresponding policy. The operator \( \mathcal{T}_P \) represents the Bellman evaluation operator for a given policy \( \pi \in \Pi \) and a transition probability model \( \mathcal{P} \). The Bellman evaluation operator \( \mathcal{T}_P \) for any \( s \in S \) and value \( v \in \mathbb{R}^S \) is defined as

\[
(\mathcal{T}_P v)_s = p_{s,\pi(s)}^T w_{s,\pi(s)} + \gamma \cdot v.
\]

where \( w_{s,\pi(s)} = r_{s,\pi(s)} + \gamma \cdot v \).

It is known that the Bellman operator \( \mathcal{T}_P \) is a \( \gamma \)-contraction mapping, monotone, and has a unique fixed point. We will use \( \hat{v}^\pi = \mathcal{T}_P \hat{v}^\pi \) to represent the random unique fixed point of the Bellman operator \( \mathcal{T}_P^\pi \) defined for a random realization of transition probabilities \( \hat{P} \). Furthermore, we will use \( p_0^T \hat{v}^\pi = \rho(\pi^*, \mathcal{P}) \) to denote the random expected returns corresponding to policy \( \pi^* \) and random realization of the transition probabilities \( \mathcal{P} \).

Suppose that \( c \in \mathbb{R}_+ \) upper-bounds the difference in performance of the VaR \( \alpha \) policy \( \hat{\pi} \) and optimal percentile-criterion policy \( \pi^* \), i.e., \( \rho^* - \hat{\rho} = c \iff \rho^* = \hat{\rho} - c \).
From the definition of VaR and \( \pi^* \), we know that
\[
\rho^* = \text{VaR}_\alpha(\rho(\pi^*, \hat{P})) = \sup\{ t : \Pr[\rho(\pi^*, \hat{P}) \geq t] \geq 1 - \delta \}. 
\]
Since \( \hat{\rho} + c \) upper bounds \( \rho^* \), we can write
\[
\Pr[\rho(\pi^*, \hat{P}) \leq \hat{\rho} + c] \geq \delta \iff \Pr[\rho(\pi^*, \hat{P}) - \hat{\rho} \leq c] \geq \delta.
\]
The above equation suggests that the error in the performance of the \( \text{VaR}_\alpha \) policy is upper-bounded by \( c \) if \( \Pr[\rho(\pi^*, \hat{P}) - \hat{\rho} \leq c] \) holds with at least \( \delta \) probability.

To derive a lower bound on \( c \), we proceed as follows.

We begin by showing that \( \rho(\pi^*, \hat{P}) - \hat{\rho} \leq \| \tilde{v}^{\pi^*} - \hat{v} \|_\infty \).
\[
\rho(\pi^*, \hat{P}) - \hat{\rho} = \rho(\pi^*, \hat{P}) - p_0^T \hat{v} \quad \text{from the definition of } \hat{\rho} \text{ and } \rho(\pi^*, \hat{P}) \\
\leq \| p_0^T \tilde{v}^{\pi^*} - p_0^T \hat{v} \|_1 \quad \text{from } \| x \|_1 \leq x, \quad \| x \|_\infty \\
\leq \| p_0 \|_1 \| \tilde{v}^{\pi^*} - \hat{v} \|_\infty \quad \text{from Hölder’s Inequality} \\
\leq \| \tilde{v}^{\pi^*} - \hat{v} \|_\infty \quad \| p_0 \|_1 = 1. \tag{16}
\]

Next, we bound \( \| \tilde{v}^{\pi^*} - \hat{v} \|_\infty \) to obtain a lower-bound on \( c \).
\[
\| \tilde{v}^{\pi^*} - \hat{v} \|_\infty \leq \frac{1}{1 - \gamma} \| T^{\pi^*}_\alpha \hat{v} - \text{VaR}_\alpha \hat{v} \|_\infty. \tag{17}
\]

Let \( \varepsilon > 0 \). Then, combining (16) and (17), we get
\[
\rho(\pi^*, \hat{P}) - \hat{\rho} \leq \frac{1}{1 - \gamma} \| T^{\pi^*}_\alpha \hat{v} - \text{VaR}_\alpha \hat{v} \|_\infty
\]
\[
\leq \frac{1}{1 - \gamma} \max_{s \in S} \left( (T^{\pi^*}_\alpha \hat{v})_s - (\text{VaR}_\alpha \hat{v})_s \right) \tag{18}
\]
\[
\leq \frac{1}{1 - \gamma} \max_{s \in S} \left( \hat{p}_{s,\pi^*(s)}^T \hat{w}_{s,\pi^*(s)} - \text{VaR}_\alpha \hat{p}_{s,\pi^*(s)}^T \hat{w}_{s,\pi^*(s)} \right) \text{ Bellman operators, and } (c) \text{ follows from the fact that } \hat{\pi}(s) \text{ maximizes } \max_{a \in A} \text{VaR}_\alpha \left[ \hat{p}_{s,a}^T \hat{w}_{s,a} \right] \text{ and thus, } \text{VaR}_\alpha \left[ \hat{p}_{s,\hat{\pi}(s)}^T \hat{w}_{s,\hat{\pi}(s)} \right] \geq \text{VaR}_\alpha \left[ \hat{p}_{s,\pi^*(s)}^T \hat{w}_{s,\pi^*(s)} \right].
\]
Therefore, for any \( \varepsilon > 0 \), we can write
\[
\Pr \left[ \rho(\pi^*, \hat{P}) - \hat{\rho} \leq \frac{1}{1 - \gamma} \max_{s \in S} \left( \text{VaR}_\alpha \left[ \hat{p}_{s,\pi^*(s)}^T \hat{w}_{s,\pi^*(s)} \right] - \text{VaR}_\alpha \left[ \hat{p}_{s,\pi^*(s)}^T \hat{w}_{s,\pi^*(s)} \right] \right) \geq \delta \right] \\
\Pr \left[ \rho(\pi^*, \hat{P}) - \hat{\rho} \leq \frac{1}{1 - \gamma} \max_{s \in S, a \in A} \left( \text{VaR}_\alpha \left[ \hat{p}_{s,a}^T \hat{w}_{s,a} \right] - \text{VaR}_\alpha \left[ \hat{p}_{s,a}^T \hat{w}_{s,a} \right] \right) \geq \delta \right]. \tag{19}
\]
The first equation in the above follows from $\Pr \left[ \hat{p}_{s,a} w_{s,a} > \text{VaR}_{1 - \frac{1}{S} - \frac{\delta}{S}} \left( \hat{p}_{s,a}^T w_{s,a} \right) \right] < \frac{1 - \delta}{S}$ for any state $s$ and action $a$ and applying union bound over all states yields $\Pr \left[ \hat{p}_{s,a} w_{s,a} > \text{VaR}_{1 - \frac{1}{S} - \frac{\delta}{S}} \left( \hat{p}_{s,a}^T w_{s,a} \right) \forall s \in S \right] < 1 - \delta$. Thus, we get $\Pr \left[ \hat{p}_{s,a} w_{s,a} \leq \text{VaR}_{1 - \frac{1}{S} - \frac{\delta}{S}} \left( \hat{p}_{s,a}^T w_{s,a} \right) \right] \forall s \in S \geq \delta$. The second equation follows by simply replacing $\pi^*(s)$ with the worst-case action, i.e., action that maximizes the upper bound.

Comparing (15) and (19), we get

$$\rho^* - \hat{\rho} \leq \frac{1}{1 - \gamma} \max_{s \in S} \max_{a \in A} \left( \inf_{\varepsilon > 0} \text{VaR}_{1 - \frac{1}{S} - \frac{\delta}{S}} \left( \hat{p}_{s,a}^T w_{s,a} \right) - \text{VaR}_{a} \left( \hat{p}_{s,a}^T w_{s,a} \right) \right).$$

Note that VaR of returns is upper semicontinuous and thus, this implies that the infimum in the above equation is achieved at $\varepsilon = 0$.

Therefore, we can write

$$\rho^* - \hat{\rho} \leq \frac{1}{1 - \gamma} \max_{s \in S} \max_{a \in A} \left( \text{VaR}_{1 - \frac{1}{S} - \frac{\delta}{S}} \left( \hat{p}_{s,a}^T w_{s,a} \right) - \text{VaR}_{a} \left( \hat{p}_{s,a}^T w_{s,a} \right) \right).$$

\[\Box\]

B.5 Proof of Theorem 3.5

**Theorem 3.5** (Asymptotic Performance). Suppose that the normality assumptions on the posterior distribution $\hat{P}$ in (2) are satisfied. For any $\delta \in (0, 0.5)$, set $\alpha = \delta/s$ in $\text{VaR}_{\alpha}$. For any state $s$ and action $a$, let $I(p_{s,a}^*)$ be the Fisher information matrix corresponding to the true transition probabilities $p_{s,a}$. Furthermore, let $\sigma_{\max}^2 = \max_{s \in S, a \in A} w_{s,a}^T I(p_{s,a}^*)^{-1} w_{s,a}$ be the maximum over state-action pairs of the asymptotic variance of the returns estimate $\hat{p}_{s,a}^T w_{s,a}$. Then the asymptotic performance of the VaR framework $\hat{\rho}$ w.r.t. the optimal percentile returns $\rho^*$ satisfies

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

**Proof.** As noted in Section 2, we assume that as $N \to \infty$, the posterior distribution of transition probabilities at any state $s$ and action $a$ $(\hat{p}_{s,a})$ converges in the limit to a multivariate Gaussian distribution with mean $p_{s,a}$ and covariance matrix $I(p_{s,a}^*)^{-1}/N$. Hence, we can write

$$\lim_{N \to \infty} \sqrt{N}(\hat{p}_{s,a}^T w_{s,a} - p_{s,a}^* T w_{s,a}) \Rightarrow N(0, w_{s,a}^T I(p_{s,a}^*)^{-1} w_{s,a}).$$

We know from Proposition 3.3 that $\text{VaR}_{\alpha}$ of a univariate Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ can be written as $\text{VaR}_{\alpha}[X] = \mu + \Phi^{-1}(\alpha)\sigma$. Therefore, applying this result to the R.H.S. of Equation (6), we get

$$\lim_{N \to \infty} \sqrt{N}(\rho^* - \hat{\rho}) \leq \frac{a}{1 - \gamma} \max_{s \in S} \max_{a \in A} \left( \sqrt{N} p_{s,a}^* T w_{s,a} + \Phi^{-1} \left( 1 - \frac{\delta}{S} \right) \sqrt{w_{s,a}^T I(p_{s,a}^*)^{-1} w_{s,a}} \right)$$

$$- \left( \sqrt{N} p_{s,a}^* T w_{s,a} + \Phi^{-1} \left( \frac{\delta}{S} \right) \sqrt{w_{s,a}^T I(p_{s,a}^*)^{-1} w_{s,a}} \right)$$

$$\leq \frac{b}{1 - \gamma} \max_{s \in S} \max_{a \in A} \left( \Phi^{-1} \left( 1 - \frac{\delta}{S} \right) \sqrt{w_{s,a}^T I(p_{s,a}^*)^{-1} w_{s,a}} \right)$$

$$- \Phi^{-1} \left( \frac{\delta}{S} \right) \sqrt{w_{s,a}^T I(p_{s,a}^*)^{-1} w_{s,a}}$$

$$\leq \frac{c}{1 - \gamma} \max_{s \in S} \max_{a \in A} \left( \Phi^{-1} \left( 1 - \frac{\delta}{S} \right) - \Phi^{-1} \left( \frac{\delta}{S} \right) \right) \sqrt{w_{s,a}^T I(p_{s,a}^*)^{-1} w_{s,a}}$$

$$\leq \frac{d}{1 - \gamma} \max_{s \in S} \max_{a \in A} \left( 2 \Phi^{-1} \left( 1 - \frac{\delta}{S} \right) \sqrt{w_{s,a}^T I(p_{s,a}^*)^{-1} w_{s,a}} \right).$$

(20)
We prove the second inequality in Theorem 3.5 by leveraging the sub-Gaussian bounds for a standard normal distribution \( \mathcal{N}(0, \sigma^2) \) to show that \( \forall \alpha' \in (0, 0.5) \), \( \Phi^{-1}(1 - \alpha') \leq \sqrt{2 \ln(1/\alpha')} \).

We know that for a standard normal distribution \( X \sim \mathcal{N}(0, \sigma^2) \) where \( \sigma \in \mathbb{R} \), the following sub-Gaussian bounds holds \([7]\):

\[
\Pr \left( -\sqrt{2 \ln(2/\alpha)} \sigma \leq X \leq \sqrt{2 \ln(2/\alpha)} \sigma \right) \geq 1 - \alpha' .
\] (21)

By definition, for a standard normal distribution \( \mathcal{N}(0, \sigma^2) \), it holds

\[
\Pr \left( -\Phi^{-1}(1 - \alpha'/2) \sigma \leq X \leq \Phi^{-1}(1 - \alpha'/2) \sigma \right) = 1 - \alpha' .
\] (22)

Comparing equation (21) and (22), we get

\[
\Phi^{-1}(1 - \alpha'/2) \leq \sqrt{2 \ln(2/\alpha')} \implies \Phi^{-1}(1 - \alpha') \leq \sqrt{2 \ln(1/\alpha')} .
\] (23)

Using the result in (23) in equation (20), proves the second inequality of the theorem. \( \square \)

### B.6 Proof of Proposition 3.6

Proposition 3.6 (Time Complexity). The time complexity of a single iteration of the loop in line 3 of the VaR Value Iteration (Algorithm 3.1) is \( \mathcal{O}(SAN) \), where \( \mathcal{N} \) is the number of samples of the posterior samples of \( P \).

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 [23] and a single iteration of the loop in line 3 of Algorithm 3.1 computes quantile of returns \( S \times A \) times. \( \square \)

### B.7 Proof of Proposition 3.7

Proposition 3.7 (Empirical Error Bound). For any state \( s \), action \( a \) and value function \( v \), let \( \hat{\text{VaR}}_a[\hat{p}^{T}_{s,a}w_{s,a}] \) represent the empirical estimate of \( \alpha \)-percentile of returns \( \text{VaR}_a[p^{T}_{s,a}w_{s,a}] \) and \( \Phi_f \) represent the cumulative density function (CDF) of the random estimate of returns \( p^{T}_{s,a}w_{s,a} \). Suppose that \( \Phi_f \) is differentiable at the point \( \text{VaR}_a[p^{T}_{s,a}w_{s,a}] \) and let \( m = \Phi'(\text{VaR}_a[p^{T}_{s,a}w_{s,a}]) \) represents the density of estimate of returns at point \( \text{VaR}_a[p^{T}_{s,a}w_{s,a}] \). Let \( N^* \) be the number of posterior samples required to obtain empirical error \( \varepsilon \in \mathbb{R} \), with confidence \( 1 - \zeta \), i.e.,

\[
\Pr \left[ |\hat{\text{VaR}}_a[\hat{p}^{T}_{s,a}w_{s,a}] - \text{VaR}_a[p^{T}_{s,a}w_{s,a}]| > \varepsilon \right] \leq \zeta .
\]

Then, \( \lim_{\varepsilon \to 0} N^* \varepsilon^2 = \ln(2/(2\zeta))/2m^2 \).

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

From the inverse-function theorem, we get,

\[
(\Phi_f^{-1}(\alpha))' = \frac{1}{\Phi_f'([\hat{p}^{T}_{s,a}w_{s,a}])}
= \frac{1}{\Phi_f'([\Phi_f^{-1}(\alpha)])}
= \frac{1}{m} .
\]

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

To prove the result, we need to find \( N^* \) such that

\[
\Pr \left[ \text{VaR}_a[p^{T}_{s,a}w_{s,a}] - \varepsilon \leq \hat{\text{VaR}}_a[\hat{p}^{T}_{s,a}w_{s,a}] \leq \text{VaR}_a[p^{T}_{s,a}w_{s,a}] + \varepsilon \right] \geq 1 - \zeta
\]

\[
\overset{(\omega)}{=} \Pr \left[ \Phi_f^{-1}(\alpha - \varepsilon m) \leq \hat{\text{VaR}}_a[p^{T}_{s,a}w_{s,a}] \leq \Phi_f^{-1}(\alpha + \varepsilon m) \right] \geq 1 - \zeta .
\] (24)
Equation (a) follows from applying a first order Taylor expansion to $\Phi_f^{-1}$ around the point $\alpha$. We apply the following results to obtain a bound on $N^*$.

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 $N^*$ samples from the true distribution of $\tilde{Z}$ and $0 < \zeta < 1$ represents the desired level of confidence guarantees, Then, from DWK inequality \[31\], we know that

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

The above equation implies that

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

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

Thus, applying equation (25) to (24), i.e., $l_t = \sqrt{\ln(2/\zeta)/2N^*}$ gives $N^* = \ln(2/\zeta)/2\pi^2m^2$.

\[\Box\]

### B.8 Proof of Proposition 3.8

**Proposition 3.8.** Define the empirical VaR Bellman optimality operator $T_{VaR_a}$ for any value $v \in \mathbb{R}^S$ and state $s \in S$ as $(T_{VaR_a}v)_s = \max_{a \in A} \text{VaR}_a \left[ \beta_{s,a}^T \omega_{s,a} \right]$. Let $\hat{v} \in \mathbb{R}^S$ and $\hat{f} \in \mathbb{R}^S$ represent the the fixed points of $T_{VaR_a}$ and $T_{VaR_a}$ respectively. Furthermore, for any $k \in \mathbb{Z}$, let $f^k$ represent the value function at the $k^{th}$ iteration in Algorithm 3.3. If $k \geq (1 - \gamma)^{-1} \ln \left( \frac{R_{max}}{\epsilon(1 - \gamma)} \right)$

where $R_{max} = \max_{s \in S, a \in A, a' \in S} R(s, a, a')$, then,

$$\| \hat{f} - f^k \|_\infty \leq \epsilon.$$  

Moreover, the error in the value function at the $k^{th}$ iteration $(f^k)$ of Algorithm 3.3 with respect to the optimal value function $\hat{v}$ is given by

$$\| \hat{v} - f^k \|_\infty \leq \epsilon + \frac{1}{1 - \gamma} \| T_{VaR_a} \hat{f} - T_{VaR_a} \hat{f} \|_\infty.$$  

**Proof.** We begin by noting that similar to the VaR$_a$ Bellman operator $T_{VaR_a}$, the empirical VaR$_a$ Bellman operator $T_{VaR_a}$ is a contraction mapping because of the monotonicity of the VaR$_a$ operator and since the empirical VaR$_a$ in the algorithm is always computed from a fixed set of transition probabilities sampled from the posterior of $\hat{P}$. Therefore, we can write

$$\| \hat{f} - f^k \|_\infty \leq \| T_{VaR_a} \hat{f} - T_{VaR_a} f^{k-1} \|_\infty \leq \gamma \| \hat{f} - f^{k-1} \|_\infty \leq \gamma^k \| T_{VaR_a} \hat{f} - T_{VaR_a} f^{k-2} \|_\infty \leq \gamma^k R_{max} \frac{1}{1 - \gamma}.$$  

(a) follows from the definition of the VaR Bellman operator $T_{VaR_a}$, (b) follows from the contraction property of the empirical VaR Bellman operator $T_{VaR_a}$, (c) follows from applying the same procedure as in (a) to step (b), and (d) follows from unrolling (c) over $k - 2$ time steps and using the fact that $\| \hat{f} - f^0 \|_\infty \leq R_{max}/1 - \gamma$ for $f^0 = 0$.

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

$$\frac{\gamma^k}{1 - \gamma} R_{max} \leq \epsilon$$  

$$k \geq \frac{\ln \left( \frac{R_{max}}{\epsilon(1 - \gamma)} \right)}{1 - \gamma}.$$  

(26)
Next, we bound the error between the optimal value function \( \hat{v} \) and the fixed point of the empirical VaR Bellman optimality operator \( \hat{f} \).

Suppose that \( \| \mathcal{T}_{\text{VaR}_\alpha} \hat{f} - \mathcal{T}_{\text{VaR}_\alpha} \hat{f} \|_\infty \leq \varepsilon \). Then,

\[
\| \hat{v} - \hat{f} \|_\infty \leq \frac{\| \mathcal{T}_{\text{VaR}_\alpha} \hat{f} - \mathcal{T}_{\text{VaR}_\alpha} \hat{f} \|_\infty}{1 - \gamma}.
\]  

The final result of the proposition follows by simply combining the result in (26) and (27) and applying triangle inequality,

\[
\| \hat{v} - \hat{f} \|_\infty \leq \| \hat{v} - \hat{f} \|_\infty + \| \hat{f} - \hat{f} \|_\infty.
\]

\[
\leq \frac{1}{1 - \gamma} \| \mathcal{T}_{\text{VaR}_\alpha} \hat{f} - \mathcal{T}_{\text{VaR}_\alpha} \hat{f} \|_\infty + \varepsilon.
\]

\( B.9 \) Proof of Proposition 4.1

**Proposition 4.1** (Equivalence). The optimal VaR policy \( \hat{\pi} \in \Pi^D \) solves

\[
\max_{\pi \in \Pi^D} \min_{P \in \mathcal{P}_{\text{VaR}, \hat{v}}} \rho(\pi, P),
\]

where \( \hat{v} \in \mathbb{R}^S \) is the fixed point of the VaR Bellman operator \( \mathcal{T}_{\text{VaR}_\alpha} \cdot \hat{v} = (\mathcal{T}_{\text{VaR}_\alpha} \hat{v}) \).

**Proof.** Recall that the VaR Bellman optimality operator defined for each \( s \in \mathcal{S} \) and \( v \in \mathbb{R}^S \) as

\[
(\mathcal{T}_{\text{VaR}_\alpha} v)_s = \max_{a \in \mathcal{A}} \text{VaR}_\alpha \left[ \tilde{p}_{s,a}^T w_{s,a} \right].
\]

Suppose that \( \hat{v} \) is the unique fixed point of \( \mathcal{T}_{\text{VaR}_\alpha} \) and \( \hat{w}_{s,a} = r_{s,a} + \gamma \cdot \hat{v} \) is the corresponding transition value for each \( s \in \mathcal{S} \) and \( a \in \mathcal{A} \). That is

\[
\hat{v} = \mathcal{T}_{\text{VaR}_\alpha} \hat{v}.
\]

Then, we define the following robust Bellman operator \( \mathcal{T} : \mathbb{R}^S \to \mathbb{R}^S \) for each \( s \in \mathcal{S} \) and \( v \in \mathbb{R}^S \) as

\[
(\mathcal{T} v)_s = \max_{a \in \mathcal{A}} \min_{P \in \mathcal{P}_{\text{VaR}, \hat{v}}} P^T w_{s,a}, \quad \text{where}
\]

\[
\mathcal{P}_{\text{VaR}, \hat{v}} = \{ P \in \Delta^\mathcal{A} | P^T \hat{w}_{s,a} \geq \text{VaR}_\alpha \left[ \tilde{p}_{s,a}^T \hat{w}_{s,a} \right] \}.
\]

We now show that \( \hat{v} \) is also the fixed point of the robust Bellman operator \( \mathcal{T} \). That is, to establish \( \mathcal{T} \hat{v} = \hat{v} \), we show for \( a \in \mathcal{A} \) that

\[
\min_{P \in \mathcal{P}_{\text{VaR}, \hat{v}}} P^T \hat{w}_{s,a} = \text{VaR}_\alpha \left[ \tilde{p}_{s,a}^T \hat{w}_{s,a} \right].
\]
The minimum above exists it minimizes a linear function on a compact set. The direction “≥” in \(30\) follows immediately from the constraint in the construction of the set \(p_{s,a}^{\text{Var}}\). The direction “≤” follows by linear program duality of the minimization over \(p\) and from the fact that \(\text{VaR}_{\alpha}(\hat{p}^{T}_{s,a}\hat{w}_{s,a}) \geq \min_{s' \in S} \hat{w}_{s,a,s'}\).

Using the equalities above, we now get that

\[
\hat{v} = \mathcal{I}_{\text{Var}_{\alpha}}(\hat{v}) = \max_{a \in A} \text{VaR}_{\alpha} \left[ \tilde{p}^{T}_{s,a} \hat{w}_{s,a} \right] = \max_{a \in A} \min_{p \in p_{s,a}^{\text{Var}}} p^{T} \hat{w}_{s,a} = \mathcal{I}_{\hat{v}}.
\]

Therefore, \(\hat{v}\) is the unique fixed point of the SA-rectangular robust Bellman operator [24] and \(\hat{v}\) is greedy with respect to the optimal robust value function of \(v\) and, therefore, is an optimal policy that solves [9].

B.10  Proof of Theorem 4.2

**Theorem 4.2** (Asymptotic Radii of VaR Ambiguity Sets). For any state \(s\) and action \(a\), let \(I(p_{s,a}^{*})\) be the Fisher information corresponding to \(p_{s,a}^{*}\). Suppose that the normality assumptions on the posterior distribution \(P\) in [2] are satisfied. Then,

\[
\lim_{N \to \infty} \sqrt{N}(P_{s,a}^{\text{VaR}} - p_{s,a}^{*}) = \begin{cases} p_{s,a} \in \Delta^S \left| \|p_{s,a} - p_{s,a}^{*}\|_{I(p_{s,a}^{*})} \leq \Phi^{-1}(1 - \alpha) \right) - p_{s,a}^{*} \end{cases}. \tag{9}
\]

For any state \(s\) and action \(a\), we define the 1-step Bellman update function \(f_{s,a} : \mathbb{R}^S \to \mathbb{R}\) such that

\[
f_{s,a}(v) = \text{VaR}_{\alpha} \left( \tilde{p}^{T}_{s,a}(r_{s,a} + \gamma v) \right).
\]

As noted in [2], we assume that as \(N \to \infty\), the posterior distribution of transition probabilities for any state \(s\) and action \(a\) (\(\tilde{p}_{s,a}\)) converges in the limit to a multivariate Gaussian distribution with mean \(p_{s,a}^{*}\) and covariance \(I(p_{s,a}^{*})/N\). Therefore, the 1-step returns \(\hat{p}^{T}_{s,a} w_{s,a}\) is asymptotically a univariate Gaussian random variable, i.e., \(\lim_{N \to \infty} \sqrt{N}(\hat{p}^{T}_{s,a} w_{s,a} - p_{s,a}^{*} w_{s,a}) \sim N(0, w_{s,a}^{T} I^{-1}(p_{s,a}^{*}) w_{s,a})\).

Then, we can write

\[
\lim_{N \to \infty} \sqrt{N} \left( f_{s,a}(v) - p_{s,a}^{*} w_{s,a} \right) = -\Phi^{-1}(1 - \alpha) \|w_{s,a}\|_{I(p_{s,a}^{*})}, \tag{31}
\]

where \(\Phi^{-1}\) represents the CDF inverse of a standard normal distribution. Equation (31) follows from the analytical form of VaR of a Gaussian random variable, i.e., for any Gaussian random variable \(Y \sim N(\mu, \sigma^2)\) with mean \(\mu\), variance \(\sigma^2\), and confidence level \(\alpha\), \(\text{VaR}_{\alpha}(Y) = \mu - \Phi^{-1}(1 - \alpha) \sigma\).

Since \(f_{s,a}\) is convex in \(v\) when \(\hat{p}_{s,a}\) is Multivariate Gaussian distributed, we can use the definition of support function of a closed convex set [9] to retrieve the unique ambiguity set \(\tilde{P}_{s,a}^{\text{VaR}}\).

Note that, in this case the support function is \(f_{s,a}\) and \(\tilde{P}_{s,a}^{\text{VaR}}\) is the closed convex set

\[
\forall v \in \mathbb{R}^S : \min_{p_{s,a} \in \tilde{P}_{s,a}^{\text{VaR}}} p^{T}_{s,a}(r_{s,a} + \gamma v) = f_{s,a}(v)
\]

\[
\forall v \in \mathbb{R}^S : \min_{p_{s,a} \in \tilde{P}_{s,a}^{\text{VaR}}} \sqrt{N} p^{T}_{s,a} w_{s,a} = \sqrt{N} f_{s,a}(v)
\]

\[
\forall v \in \mathbb{R}^S : \min_{p_{s,a} \in \tilde{P}_{s,a}^{\text{VaR}}} \sqrt{N}(p^{T}_{s,a} w_{s,a} - p_{s,a}^{*} w_{s,a}) = \sqrt{N}(f_{s,a}(v) - p_{s,a}^{*} w_{s,a})
\]

\[
\forall v \in \mathbb{R}^S, z_{s,a} \in \left( \sqrt{N}(\tilde{P}_{s,a}^{\text{VaR}} - p_{s,a}^{*}) \right) : z^{T}_{s,a} w_{s,a} \geq -\Phi^{-1}(1 - \alpha) \|w_{s,a}\|_{I(p_{s,a}^{*})}.
\]

Rearranging the terms in the above and setting \(z_{s,a} = p_{s,a} - p_{s,a}^{*}\), we get

\[
\lim_{N \to \infty} \sqrt{N}(P_{s,a}^{\text{VaR}} - p_{s,a}^{*}) = \left\{ p_{s,a} - p_{s,a}^{*} \mid \forall v \in \mathbb{R}^S, p_{s,a} \in \Delta^S, p^{T}_{s,a} w_{s,a} \geq p_{s,a}^{*} w_{s,a} - \Phi^{-1}(1 - \alpha) \|w_{s,a}\|_{I(p_{s,a}^{*})} \right\} \tag{32}.
\]
Using basic algebraic manipulations as shown in Proposition B.1, we can write the asymptotic form of ambiguity set $P^*_{s,a}$ as an ellipsoid

$$\lim_{N \to \infty} \sqrt{N} (P^*_{s,a} - p^*_{s,a}) = \left\{ p_{s,a} \in \Delta^S \mid \|p_{s,a} - p_{s,a}\|_I(p^*_{s,a}) \leq \Phi^{-1}(1 - \alpha) \right\} - p^*_{s,a}. \quad (33)$$

**Proposition B.1.** Consider the two ambiguity sets given in equations (32) and (33). These two representations are equivalent.

**Proof.** For any state $s$ and action $a$, as $N \to \infty$, $(p_{s,a} - p^*_{s,a}) \in \sqrt{N} (P^*_{s,a} - p^*_{s,a})$ satisfies

$$p^T_{s,a}w_{s,a} \geq p^*_{s,a}^T w_{s,a} - \Phi^{-1}(1 - \alpha)\|w_{s,a}\|_I(p^*_{s,a}) \quad (a)$$

$$w^T_{s,a}(p_{s,a} - p^*_{s,a}) \leq \Phi^{-1}(1 - \alpha)^2 w^T_{s,a} I(p^*_{s,a})^{-1} w_{s,a} \quad (b)$$

The first and second equation follow from the definition of $P^*_{s,a}$ and simple algebraic manipulations. The third equation follows by multiplying by $-1$ on both sides and squaring both sides. Using the basic properties of semi-positive definite matrices, we can write the above equation as

$$((\Phi^{-1}(1 - \alpha))^2 I(p^*_{s,a})^{-1} - (p_{s,a} - p^*_{s,a})(p_{s,a} - p^*_{s,a})^T) \succeq 0 \quad (a)$$

$$((\Phi^{-1}(1 - \alpha)^2 I(p^*_{s,a})^{-1} - (p_{s,a} - p^*_{s,a})(p_{s,a} - p^*_{s,a})^T) I(p^*_{s,a})) \succeq 0 \quad (b)$$

$$((\Phi^{-1}(1 - \alpha)^2 I(p^*_{s,a}) I(p^*_{s,a}) - (p_{s,a} - p^*_{s,a})(p_{s,a} - p^*_{s,a})^T) I(p^*_{s,a}) \succeq 0 \quad (c)$$

$$((p_{s,a} - p^*_{s,a})^T I(p^*_{s,a}) (p_{s,a} - p^*_{s,a}) I(p^*_{s,a}) I(p^*_{s,a}) - ((p_{s,a} - p^*_{s,a})(p_{s,a} - p^*_{s,a})^T I(p^*_{s,a})(p_{s,a} - p^*_{s,a}))^T) \succeq 0 \quad (d)$$

$$((\Phi^{-1}(1 - \alpha))^2 (p_{s,a} - p^*_{s,a})^T I(p^*_{s,a}) (p_{s,a} - p^*_{s,a}) - ((p_{s,a} - p^*_{s,a})(p_{s,a} - p^*_{s,a})^T I(p^*_{s,a})(p_{s,a} - p^*_{s,a}))^2 \succeq 0 \quad (e).$$

(a) holds because for any $A \in \mathbb{R}^{n \times n}$ if $x^T A x \geq 0 \forall x \in \mathbb{R}^n$, then $A \succeq 0$, (b) follows from $U^T M U \succeq 0 \forall U, M \succeq 0$, (c) follows from simple algebraic manipulations, (d) follows from $(p_{s,a} - p^*_{s,a})^T I(p^*_{s,a})(p_{s,a} - p^*_{s,a}) \succeq 0$ because $I(p^*_{s,a})^{-1} \succeq 0$, and (e) follows from simply rearranging all the terms in the above equation.

Therefore, we get

$$(p_{s,a} - p^*_{s,a})^T I(p^*_{s,a})(p_{s,a} - p^*_{s,a}) \leq (\Phi^{-1}(1 - \alpha))^2 \iff \|p_{s,a} - p^*_{s,a}\|_I(p^*_{s,a})^{-1} \leq \Phi^{-1}(1 - \alpha).$$

The proof for the other direction is simply the reverse of this proof and therefore we omit it. \(\square\)

**B.11 Proof of Theorem 4.3.**

**Theorem 4.3 (Asymptotic Radius of Bayesian Credible Regions).** For any state $s$ and action $a$, let $P^*_{s,a}$ represent any Bayesian credible region. Let $\xi < \sqrt{\frac{2}{1 - \alpha}}/\Phi^{-1}(1 - \alpha)$. Suppose that the normality assumptions on the posterior distribution of $P$ in (2) are satisfied. Then, $\forall s \in S, a \in A$.

$$\lim_{N \to \infty} \sqrt{N} (P^*_{s,a} - p^*_{s,a}) \not\subseteq \lim_{N \to \infty} \sqrt{N} (P^*_{s,a} - p^*_{s,a}). \quad (10)$$

We will prove this theorem by contradiction.

For any state $s$ and action $a$, suppose that $\lim_{N \to \infty} \sqrt{N} (P^*_{s,a} - p^*_{s,a}) \subseteq \lim_{N \to \infty} \sqrt{N} (P^*_{s,a} - p^*_{s,a})$. 25
Since \( \mathcal{P}_{s,a}^{\text{BCR}} \) is a credible region, it satisfies
\[
1 - \alpha \leq \Pr \left[ (\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*) \in \mathcal{P}_{s,a}^{\text{BCR}} \right]
\]
\[
= \Pr \left[ \lim_{N \to \infty} \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*) \in \lim_{N \to \infty} \sqrt{N}(\mathcal{P}_{s,a}^{\text{BCR}}) \right]
\]
\[
= \Pr \left[ \lim_{N \to \infty} \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*)^T \mathbf{I}(\mathbf{p}_{s,a}^*) \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*) \leq \xi^2(\Phi^{-1}(1 - \alpha))^2 \right].
\]
(a) follows from the definition of Bayesian credible regions, (b) follows from multiplying by \( \sqrt{N} \) on both sides and taking limit \( N \to \infty \), (c) follows from the assumption \( \lim_{N \to \infty} \sqrt{N}(\mathcal{P}_{s,a}^{\text{BCR}}) \subseteq \lim_{N \to \infty} \sqrt{N}(\mathcal{P}_{s,a}^{\text{VaR}} - \mathbf{p}_{s,a}^*) \), and (d) follows from results in (33) and squaring both sides.

Since \( \xi < \frac{\sqrt{\chi^2_{S,1-\beta}}}{\Phi^{-1}(1-\alpha)} \), there exists \( \beta > 0 \), such that \( \xi^2(\Phi^{-1}(1 - \alpha))^2 \leq \chi^2_{S,1-\beta} - \beta \). Fix such \( \beta \). Then we can write
\[
1 - \alpha \leq \Pr \left[ \lim_{N \to \infty} \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*)^T \mathbf{I}(\mathbf{p}_{s,a}^*) \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*) \leq \xi^2(\Phi^{-1}(1 - \alpha))^2 \right]
\]
\[
= \Pr \left[ \lim_{N \to \infty} \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*)^T \mathbf{I}(\mathbf{p}_{s,a}^*) \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*) \leq (\chi^2_{S,1-\beta} - \beta) \right]
\]
\[
\leq \Pr \left[ \lim_{N \to \infty} \left\| \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*) \right\|^2_{\mathbf{I}(\mathbf{p}_{s,a}^*)^{-1}} \leq (\chi^2_{S,1-\beta} - \beta) \right]
\]
\[
= \Pr \left[ \lim_{N \to \infty} \left\| \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*) \right\|^2_{\mathbf{I}(\mathbf{p}_{s,a}^*)^{-1}} \leq \chi^2_{S,1-\beta} \right]
\]
\[
= 1 - \beta - \alpha < 1 - \alpha \quad \text{Contradiction!}.
\]
(a) follows from choosing \( \beta > 0 \) such that \( \xi^2(\Phi^{-1}(1 - \alpha))^2 \leq \chi^2_{S,1-\beta} - \beta \), (b) follows from simple algebraic manipulations, c follows because of the monotonic property of CDF, and (d) follows because \( \lim_{N \to \infty} \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*) \sim \mathcal{N}(0, \mathbf{I}(\mathbf{p}_{s,a}^*)) \) and \( \lim_{N \to \infty} \left\| \sqrt{N}(\mathbf{p}_{s,a} - \mathbf{p}_{s,a}^*) \right\|^2_{\mathbf{I}(\mathbf{p}_{s,a}^*)^{-1}} \) is a standard Chi-squared random variable with degrees of freedom \( S \). Therefore, the probability in (d) can be atmost \( 1 - \beta - \alpha \).

Therefore, \( \xi \geq \frac{\sqrt{\chi^2_{S,1-\beta}}}{\Phi^{-1}(1-\alpha)} \).

### C Experiments

<table>
<thead>
<tr>
<th>Methods</th>
<th>Riverswim</th>
<th>Inventory</th>
<th>Population-Small</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>VaR</td>
<td>83.51 ± 11.76</td>
<td>474.98 ± 0.7</td>
<td>-1877.66 ± 104.64</td>
<td>-3338.26 ± 213.44</td>
</tr>
<tr>
<td>VaRN</td>
<td>92.32 ± 11.76</td>
<td>472.49 ± 2.24</td>
<td>-2271.13 ± 165.98</td>
<td>-3140.68 ± 122.72</td>
</tr>
<tr>
<td>BCR ( l_1 )</td>
<td>82.04 ± 8.82</td>
<td>377.73 ± 0.0</td>
<td>-3731.29 ± 122.4</td>
<td>-4363.46 ± 240.62</td>
</tr>
<tr>
<td>BCR ( l_{\infty} )</td>
<td>80.57 ± 0.0</td>
<td>199.82 ± 39.5</td>
<td>-6668.47 ± 43.1</td>
<td>-8118.68 ± 327.74</td>
</tr>
<tr>
<td>WBCR ( l_1 )</td>
<td>82.04 ± 8.82</td>
<td>470.39 ± 5.82</td>
<td>-3286.26 ± 1115.22</td>
<td>-4257.33 ± 194.4</td>
</tr>
<tr>
<td>WBCR ( l_{\infty} )</td>
<td>80.57 ± 0.0</td>
<td>199.82 ± 39.5</td>
<td>-6395.77 ± 88.1</td>
<td>-7547.52 ± 160.76</td>
</tr>
<tr>
<td>Soft-Robust</td>
<td>115.33 ± 1.16</td>
<td>477.81 ± 0.0</td>
<td>-2052.83 ± 78.2</td>
<td>-3869.46 ± 124.72</td>
</tr>
<tr>
<td>Naive Hoeffding</td>
<td>52.29 ± 9.18</td>
<td>-0.0 ± 0.0</td>
<td>-7778.21 ± 89.36</td>
<td>-8496.47 ± 205.54</td>
</tr>
<tr>
<td>Opt Hoeffding</td>
<td>51.15 ± 6.88</td>
<td>-0.0 ± 0.0</td>
<td>-7723.6 ± 4.82</td>
<td>-8583.83 ± 16.06</td>
</tr>
</tbody>
</table>

Table 2: This shows the 95% confidence interval of the robust (percentile) returns achieved by VaR, VaRN, BCR, \( l_1 \), BCR, \( l_{\infty} \), WBCR, \( l_1 \), WBCR, Soft Robust, Naive Hoeffding and Opt Hoeffding agents at \( \delta = 0.15 \) in Riverswim, Inventory, Population-Small, and Population domain.

### D Scalability of the VaR Framework

We define the VaR\( \alpha \) q-value function (Q-function) for any policy \( \pi \in \Pi_D \) as \( q^\pi : S \times A \to \mathbb{R} \) such that for any state \( s \) and action \( a \), \( q^\pi(s,a) = \text{VaR}_{\alpha}(q^{T}_{s,a}(r_{s,a} + \gamma v^\pi)) \), where \( v^\pi = T^\pi_{\text{VaR}_{\alpha}} \) is the
Table 3: shows the 95% confidence interval of the robust (percentile) returns achieved by VaR, VaRN, BCR \( \ell_1 \), BCR \( \ell_\infty \), WBCR \( l_1 \), WBCR \( l_\infty \), Soft Robust, Naive Hoeffding and Opt Hoeffding agents at \( \delta = 0.30 \) in Riverswim, Inventory, Population-Small, and Population domain. Bolded text indicates the instances in which the VaR framework outperforms the other baselines in terms of the mean robust performance.

Figure 2: Comparison of test and train robust returns achieved by VaR, VaRN, BCR \( \ell_1 \), BCR \( \ell_\infty \), WBCR \( l_1 \), WBCR \( l_\infty \), Soft Robust, Naive Hoeffding and Opt Hoeffding agents at confidence level \( \delta = 0.05 \) in Riverswim, Inventory, Population-Small and Population domain. VaR framework achieves the highest mean robust returns in most of the domains on test and train datasets.

fixed point of the \( \text{VaR}_\delta \) Bellman evaluation operator for policy \( \pi \). We will use \( \hat{q} \) to denote the q-value function corresponding to the optimal \( \text{VaR}_\delta \) policy \( \hat{\pi} \), i.e., \( \hat{\pi}(s) = \arg \max_{a \in A} \hat{q}(s, a) \).

Let \( \hat{q}_\theta \) and \( \tilde{q}_\theta \) denote the parameterized Q-value and corresponding target Q-value networks with parameters \( \theta \in \Theta \) and \( \bar{\theta} \in \bar{\Theta} \) respectively. Here \( \hat{q}_\theta \) and \( \tilde{q}_\theta \) may represent any function approximators with parameter space \( \Theta \).
The optimal Q-value network \(q^*_{\theta} \) can be learned by simply minimizing the empirical \( \text{VaR}_\alpha \) Bellman residual error \( J_{\text{VaR}_\alpha}(\theta) \), i.e.,

\[
\hat{q}^*_{\theta} = \arg\min_{\theta \in \Theta} J_{\text{VaR}_\alpha}(\theta) \\
= \arg\min_{\theta \in \Theta} \mathbb{E}_{(s,a) \sim D} \left[ (q_{\theta}(s,a) - \max_{a' \in A} \text{VaR}_\alpha[\mathbb{E}_{s' \sim \bar{P}_{s,a}} [r_{s,a} + \gamma \bar{q}_{\theta}(s', a')]])^2 \right],
\]

where data \( D \) is a list state-action \((s, a)\) tuples that is collected using the current policy on the mean model \( \bar{P} \). In this case, the target Q-value network parameters \( \bar{\theta} \) is periodically updated by overwriting the target q-value network parameter with the Q-value network parameters \( \theta \).

We can also using any of the existing Actor-Critic methods \cite{22, 42} to learn the optimal \( \text{VaR}_\alpha \) policy. In this case, instead of optimizing the policy to minimize the Bellman residual error, the algorithm will have to optimize the policy to minimize the \( \text{VaR}_\alpha \) Bellman residual error \( J_{\text{VaR}_\alpha}(\theta) \).

### E Implementation Details

#### Hyperparameters for Riverswim Domain

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of train models per dataset ((M))</td>
<td>80</td>
</tr>
<tr>
<td>Number of test models ((K))</td>
<td>700</td>
</tr>
<tr>
<td>Number of train datasets ((L))</td>
<td>10</td>
</tr>
</tbody>
</table>

#### Hyperparameters for Inventory Domain

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of train models per dataset ((M))</td>
<td>80</td>
</tr>
<tr>
<td>Number of test models ((K))</td>
<td>200</td>
</tr>
<tr>
<td>Number of train datasets ((L))</td>
<td>10</td>
</tr>
</tbody>
</table>

#### Hyperparameters for Population-Small Domain

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of train models per dataset ((M))</td>
<td>80</td>
</tr>
<tr>
<td>Number of test models ((K))</td>
<td>100</td>
</tr>
<tr>
<td>Number of train datasets ((L))</td>
<td>10</td>
</tr>
</tbody>
</table>

#### Hyperparameters for Population Domain

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of train models per dataset ((M))</td>
<td>800</td>
</tr>
<tr>
<td>Number of test models ((K))</td>
<td>1000</td>
</tr>
<tr>
<td>Number of train datasets ((L))</td>
<td>9</td>
</tr>
</tbody>
</table>

#### E.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. The code and datasets is made available at [https://github.com/elitalobo/VaRFramework.git](https://github.com/elitalobo/VaRFramework.git).

#### E.2 Machine Specifications

We concurrently ran all experiments using 120 threads on a CPU swarm cluster with 2GB memory per thread. The total computational time was \( \sim 3 \) hours.