A Bayesian Risk Approach to MDPs with Parameter Uncertainty

Yifan Lin
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
ylin429@gatech.edu

Yuxuan Ren
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
yren79@gatech.edu

Enlu Zhou
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
enlu.zhou@isye.gatech.edu

Abstract
We consider Markov Decision Processes (MDPs) where distributional parameters, such as transition probabilities, are unknown and estimated from data. The popular distributionally robust approach to addressing the parameter uncertainty can sometimes be overly conservative. In this paper, we propose a Bayesian risk approach to MDPs with parameter uncertainty, where a risk functional is applied in nested form to the expected discounted total cost with respect to the Bayesian posterior distributions of the unknown parameters in each time stage. The proposed approach provides more flexibility of risk attitudes towards parameter uncertainty and takes into account the availability of data in future time stages. For the finite-horizon MDPs, we show the dynamic programming equations can be solved efficiently with an upper confidence bound (UCB) based adaptive sampling algorithm. For the infinite-horizon MDPs, we propose a risk-adjusted Bellman operator and show the proposed operator is a contraction mapping that leads to the optimal value function to the Bayesian risk formulation. We demonstrate the empirical performance of our proposed algorithms in the finite-horizon case on an inventory control problem and a path planning problem.

1 Introduction
Markov decision process (MDP) is a paradigm for modeling sequential decision making under uncertainty. From a modeling perspective, some parameters of MDPs are unknown and need to be estimated from data over time. In this paper, we consider MDPs where transition probabilities and cost functions are not known initially. A natural question would be: given a set of data that is finite and often small initially, how does a decision maker find a robust policy that minimizes the expected total cost under the uncertain transition parameters and cost parameters? A possible approach that mitigates the (distributional) parameter uncertainty in MDPs lies in the framework of robust MDPs (Nilim and Ghaoui [2004], Iyengar [2005] and Wiesemann et al. [2013]). In robust MDPs, parameters are assumed to belong to a known set, which is called the ambiguity set. However, one of the issues of robust MDPs is its difficulty in incorporating probabilistic information of unknown parameters into the model (Mannor and Xu [2019]). As a result, Xu and Mannor [2010] extend the distributionally robust approach, which originated from (single-stage) stochastic optimization (e.g., Delage and Ye [2010]).
To summarize, the contributions of this paper are three folds. First, we propose a Bayesian risk optimization framework that reformulates the (single-stage) stochastic optimization problem with parameter uncertainty. They estimate the parameter uncertainty using a Bayesian posterior distribution, and impose a risk functional on the objective function with respect to the posterior distribution. Later [Wu et al., 2018] show the consistency and robustness against parameter uncertainty of the BRO framework.

In this paper, we extend the Bayesian risk approach in [Wu et al., 2018] to MDPs with parameter uncertainty. We model the uncertainty over the unknown parameters (transition probability and cost) via a Bayesian posterior distribution that is updated based on realization of the randomness at every time stage. We then impose a risk functional, taken with respect to the posterior distribution, on the expected (discounted) total cost in a nested form. Therefore, we seek an optimal policy that minimizes the risk-adjusted expected (discounted) total cost. Note that Thompson sampling or posterior sampling (PSRL) [Strens, 2000, Osband et al., 2013, Fonteneau et al., 2013, Gopalan and Mannor, 2015, Abbasi-Yadkori and Szepesvári, 2015, Osband and Van Roy, 2017, and Ouyang et al., 2017] applies a similar technique to MDPs or reinforcement learning (RL) through two steps: first, a single instance of the environment is sampled from the posterior distribution; second, PSRL seeks the policy that is optimal under the sampled environment over time. Different from PSRL, we keep updating the posterior distribution conditioned on the realized randomness in the transition and cost. In addition, we take into account the entire posterior distribution (via the risk functional) instead of only one single sample from the posterior distribution. The risk functionals we consider include the risk-neutral expectation and risk-averse ones such as value at risk (VaR, cf. Jorion, 1997) and conditional value at risk (CVaR, cf. Rockafellar and Uryasev, 2000).

Risk-averse decision making has been studied in MDPs and RL (e.g., Ruszczyński, 2010, Petrik and Subramanian, 2012, Carpin et al., 2016, Jiang and Powell, 2018, Ahmadi et al., 2020, and Rigter et al., 2021). In particular, Ruszczyński, 2010 considers the risk averse MDPs, with the aim of replacing the expectation (with respect to the state transition) by general risk measures. However, most of the aforementioned risk-averse MDP literature focus on the intrinsic stochastic uncertainty in the state transition, while the risk in our work is with respect to the parameter uncertainty. The risk-averse approach is also studied in the related literature of stochastic programming (e.g., Philpott et al., 2013, Shapiro, 2016, Dowson et al., 2020, Shawro, 2021, Fichler and Shapiro, 2021). The closest to our approach is probably Shapiro, 2021, who proposes a nested formulation of the risk-averse multistage stochastic programming and shows that this nested form explicitly specifies the dynamics of the considered problem. What sets our work apart from Shapiro, 2021 is the Bayesian framework. While Shapiro, 2021 considers the risk measure directly conditioned on the realized data process, we compute Bayesian posterior distribution of the parameter based on realized data process and consider the risk functional with respect to the posterior. The Bayesian formulation is natural in the dynamic setting, and thus has been widely applied to MDPs (e.g., Poupard et al., 2006, Delage and Mannor, 2010, Imani et al., 2018, Petrik and Russel, 2019, and Derman et al., 2020). In particular, Derman et al., 2020 address the issue of learning in robust MDPs using a Bayesian approach, where they update the posterior of the transition after observing the state transition and construct the posterior uncertainty set dynamically; each time stage, one MDP model is sampled according to the current posterior distribution, and the agent acts greedily with respect to the so-called “robust Q-values” plus the posterior variance. While we also update the posterior distribution of transition probabilities, we utilize the entire posterior distribution and take a risk functional with respect to the posterior on the expected total costs.

To summarize, the contributions of this paper are three folds. First, we propose a Bayesian risk approach to MDPs with parameter uncertainty. Second, we propose a general algorithm to solve the finite-horizon problem under different risk functionals. Moreover, we develop an upper confidence bound (UCB) based adaptive sampling algorithm to efficiently solve the finite-horizon problem under risk-neutral measure, and rigorously show convergence of the algorithm. Third, we propose a
risk-adjusted Bellman operator for the infinite-horizon MDPs with parameter uncertainty and show the proposed operator is a contraction mapping. The remainder of the paper is organized as follows. Section 2 introduces the problem formulations in existing literature and presents our Bayesian risk MDP (BR-MDP) formulation. Section 3 considers the finite-horizon case and develops efficient algorithms to solve the corresponding dynamic programming equations. Section 4 considers the infinite-horizon case and shows theoretical properties of the corresponding Bellman operator. Section 5 demonstrates the empirical performance of the proposed algorithms with an inventory control problem and a path planning problem. Section 6 concludes the paper.

2 Problem formulations

Consider an MDP defined as \((S, A, P, C, \gamma)\), where \(S\) is the state space, \(A\) is the action space, \(P\) is the transition probability with \(P(s'|s, a)\) denoting the probability of transitioning to state \(s'\) from state \(s\) when action \(a\) is taken, \(C\) is the cost function with \(C(s, a)\) denoting the cost given the state \(s\) and the action \(a\), and \(\gamma \in (0, 1)\) is the discount factor. We assume \(S\) and \(A\) are finite sets. Assuming the time horizon is \(T\), a policy \(\pi = \{\pi_1, \ldots, \pi_T\}\) is a sequence of functions where each \(\pi_t\) is a function mapping from \(S\) to \(A\). Given an initial state \(s\), the goal is to find an optimal policy that minimizes the expected discounted total cost:

\[
\min_{\pi} \mathbb{E}_{\pi}^{s, P, C} \left[ \sum_{t=0}^{T-1} \gamma^t C(s_t, a_t) \right],
\]

where \(\mathbb{E}_{\pi}^{s, P, C}\) is the expectation with policy \(\pi\) under the condition that the transition probability is \(P\) and the cost is \(C\). In practice, \(P\) and \(C\) are often unknown, while we are only given a historical data set \(\phi_H\) of their parameters. To deal with this parameter uncertainty in MDPs, below we first review some existing formulations and discuss their limitations, and then present our proposed formulation.

2.1 Robust MDPs and distributionally robust MDPs

In robust MDPs (Nilim and Ghaoui [2004], Iyengar [2005], and Wiesemann et al. [2013]), it is assumed that the true unknown parameters \(P\) and \(C\) belong to a known uncertainty set \(U\) that is constructed from \(\phi_H\). Given an initial state \(s\), the decision maker aims to find a policy \(\pi\) that minimizes the expected discounted total cost under the most adversarial parameter scenarios, i.e.,

\[
\min_{\pi} \max_{(P, C) \in U} \mathbb{E}_{\pi}^{s, P, C} \left[ \sum_{t=0}^{T-1} \gamma^t C(s_t, a_t) \right].
\]

Solving this robust MDP can be reduced to solving a linear program and second-order cone program for the polytope and ellipsoid uncertainty sets respectively (Mannor and Xu [2019]). In distributionally robust MDPs (Xu and Mannor [2010]), the uncertain parameters \(P, C\) are modeled as random variables following an unknown distribution \(Q\), where \(Q\) is assumed to belong to an ambiguity set \(U\) constructed from historical data \(\phi_H\), i.e.,

\[
\min_{\pi} \max_{Q \in U} \mathbb{E}_{\pi, P, C \sim Q} \left[ \mathbb{E}_{\pi}^{s, P, C} \left[ \sum_{t=0}^{T-1} \gamma^t C(s_t, a_t) \right] \right].
\]

However, focusing on the worst-case scenario may result in overly conservative solutions, especially when the worst case has a very small probability to happen in reality. In view of this drawback of the distributionally robust approach, we intend to take the Bayesian risk approach (Zhou and Xie [2015], Wu et al. [2018]) to address the parameter uncertainty in MDPs. Following the above formulation of distributionally robust MDPs, it is tempting to replace the worst-case measure by a more general risk functional \(\rho\):

\[
\min_{\pi} \rho_{P, C \sim \mu} \left[ \mathbb{E}_{\pi}^{s, P, C} \left[ \sum_{t=0}^{T-1} \gamma^t C(s_t, a_t) \right] \right],
\]

where \(\rho\) is the risk functional applied to \(\mathbb{E}_{\pi}^{s, P, C} \left[ \sum_{t=0}^{T-1} \gamma^t C(s_t, a_t) \right]\) with respect to the uncertainty in \(P\) and \(C\), which is quantified by \(\mu\), the posterior distribution of the parameter based on
We denote an optimal policy to problem (2) as \( \text{VaR} \). Widely-used risk functionals include mean-variance, \( \text{VaR} \), and \( \text{CVaR} \). For a random variable \( X \), we only consider deterministic Markovian policy \( \pi \).

To address these issues, Shapiro [2021] proposes a nested distributionally robust formulation for multistage stochastic programming.

### 2.2 Bayesian risk MDPs

Motivated by the Bayesian risk approach of Wu et al. [2018] and the nested formulation of Shapiro [2021], we propose a new formulation named as Bayesian risk MDP (BR-MDP) to deal with the parameter uncertainty in MDPs. Specifically, we consider the case where the true underlying distribution of the randomness in the state transition (and cost) follows a parametric distribution. This assumption is widely applicable to real-world problems. For example, consider an inventory control problem, where the customer demand is unknown but assumed to follow a Poisson process (cf. Gallego and Van Ryzin [1994]). We use \( \xi \in \Xi \) to denote the randomness in the state transition (and cost), such that given the current state \( s \) and action \( a \), the next state \( s' \) satisfies \( s' = g(s, a, \xi) \), where \( g \) is a known function called the state equation. The state equation together with the distribution of \( \xi \) uniquely determines the transition probability of the MDP, i.e., \( P(s'|s, a) = P(s' = g(s, a, \xi)|s, a) \). We use the representation of state equation instead of transition probability in MDPs, for the purpose of decoupling the randomness and the policy, leading to a cleaner formulation in nested form. The cost is assumed to be a function of state \( s \) and action \( a \), and \( \xi \), i.e., \( C(s, a, \xi) \). Note that the transition function \( g \) and cost function \( C \) are both independent of time in the stationary setting. From our assumption, the randomness \( \xi \) follows distribution \( f(\cdot; \theta^0) \), which belongs to a (known) parametric family of distributions \( \{f(\cdot; \theta)|\theta \in \Theta\} \) but has unknown parameter value \( \theta^0 \). By taking a Bayesian perspective, we view the unknown parameter as a random variable \( \theta \) and estimate it using the Bayesian posterior distribution \( \mu_t \) at time \( t \), given a prior distribution \( \mu_{-1} \) (over the support \( \Theta \)) and historic data \( \phi_t \) before time 0. Finally, we only consider deterministic Markovian policy \( \pi = \{\pi_t|\pi_t: (S \times \mathcal{M}) \to A, t = 0, \ldots, T - 1\} \), where \( \mathcal{M} \) is the space of posterior distributions. The BR-MDP formulation is presented below.

\[
\min_{\pi} \rho_{0\theta_0} \mathbb{E}_{\xi_0|\theta_0} \left[ C(s_0, a_0, \xi_0) + \cdots + \gamma^{T-1} \rho_{\theta_{T-1}} \mathbb{E}_{\xi_{T-1}|\theta_{T-1}} \left[ C(s_{T-1}, a_{T-1}, \xi_{T-1}) + \cdots + \gamma^{T-2} \rho_{\theta_{T-2}} \mathbb{E}_{\xi_{T-2}|\theta_{T-2}} \left[ C(s_{T-2}, a_{T-2}, \xi_{T-2}) + \cdots + \gamma \rho_{\theta_1} \mathbb{E}_{\xi_0|\theta_0} \left[ C(s_0, a_0, \xi_0) \right] \right] \right] \right] 
\]

s.t. \( s_{t+1} = g(s_{t}, a_{t}, \xi_{t}) \), \( t = 0, \ldots, T - 2 \);
\[
\mu_{t+1}(\theta) = \frac{\mu_t(\theta) f(\xi_t|\theta)}{\int \mu_t(\theta) f(\xi_t|\theta) d\theta}, \quad t = 0, \ldots, T - 2; \quad \mu_0 = \frac{\mu_{-1}(\theta) f(\phi_0|\theta)}{\int \mu_{-1}(\theta) f(\phi_0|\theta) d\theta}.
\]

Note \( a_t = \pi_t(s_{t}, \mu_t) \), \( \mathbb{E}_{\xi_{t}|\theta_{t}} \) denotes the expectation with respect to \( \xi_t \sim f(\cdot; \theta_t) \) conditional on \( \theta_t \), and \( \rho_{\theta_0} \) denotes a risk functional take with respect to \( \theta_t \sim \mu_t \). Equation (3) is the update of the posterior \( \mu_t \), given a chosen prior \( \mu_{-1} \) and historic data \( \phi_t \) before time 0. The posterior \( \{\mu_t\} \) can be viewed as a self-evolving process that is independent of the action \( a_t \) (in other words, the policy \( \pi \)), and it is easy to see that \( \{\mu_t\} \) is a Markovian process. Hence, BR-MDP is an MDP with the augmented state \( (s_t, \mu_t) \). In fact, \( \mu_t \) is similar to the belief state in partially observable MDPs (POMDPs) (cf. Chapter 5 in Bertsekas et al. [2000]), which summarizes all the information of the data up to time and transforms the problem to an MDP.

We denote an optimal policy to problem (2) as \( \pi^* \). To distinguish \( \pi^* \) from the optimal policy in standard MDPs (without parameter uncertainty), we call \( \pi^* \) the risk-adjusted optimal policy. Some widely-used risk functionals include mean-variance, \( \text{VaR} \), and \( \text{CVaR} \). For a random variable \( X \), \( \text{VaR}^\alpha(X) \) is defined as the \( \alpha \)-quantile of \( X \) with \( \alpha \in (0, 1) \), i.e., \( \text{VaR}^\alpha(X) := \inf\{t : \mathbb{P}(X \leq t) \geq \alpha\} \), and \( \text{CVaR}^\alpha(X) \) is defined as the expected loss beyond \( \text{VaR} \), i.e., \( \text{CVaR}^\alpha(X) := \frac{1}{1-\alpha} \int_{\alpha}^1 \text{VaR}^\alpha(X) d\alpha \).

In particular, for the risk functional \( \rho = \text{VaR} \), when we choose \( \alpha = 1 \), it is similar to the worst-case scenario as studied in the distributionally robust optimization, except that the ambiguity set is the entire parameter space in our case. In sharp contrast to the formulations reviewed in Section 2.1, (2) imposes risk functionals in a nested form to incorporate the data process over time through the posterior updating.
3 Bayesian risk MDPs: the finite-horizon case

In this section, we consider solution approaches to problem (2) in the finite-horizon case. The exact dynamic programming is summarized in Algorithm 1.

Algorithm 1: Exact dynamic programming for finite-horizon BR-MDPs.

| input: | finite horizon $T$, historical data set of randomness $\phi_H$, initial state $s_0$, prior distribution $\mu_1$ |
| output: | risk-adjusted optimal value function $V_t^*(s, \mu_t)$ and corresponding policy $\pi^*$ |

Set $V_T^*(s, \mu) = 0, \forall s \in S, \mu \in M$; compute $\mu_0 = \frac{\mu_{-1}(\theta) f(\phi_H(\theta)) d\theta}{\int \mu_{-1}(\theta) f(\phi_H(\theta)) d\theta}$.

for $t \leftarrow T - 1$ to 0 do

\[ V_t^*(s_t, \mu_t) = \min_{a_t \in A} \rho_t \mathbb{E}_{\xi_t|\theta_t}[C(s_t, a_t, \xi_t) + \gamma V_{t+1}^*(s_{t+1}, \mu_{t+1})|s_t, \mu_t, a_t], \forall s_t, \mu_t, \]

where $s_{t+1} = g(s_t, a_t, \xi_t)$, $\mu_{t+1}(\theta) = \frac{\mu_t(\theta)f(\xi_t|\theta)}{\int \mu_t(\theta)f(\xi_t|\theta) d\theta}$;

Set $\pi_t^*(s_t, \mu_t) := a_t^*(s_t, \mu_t)$, where $a_t^*(s_t, \mu_t)$ attains $V_t^*(s_t, \mu_t)$;

end

There are two main challenges to carry out the exact dynamic programming. First, it is in general not possible to analytically compute the optimal value function. Compared with standard MDPs (without parameter uncertainty), there is a nested structure of risk functional $\rho$ composed with expectation in the computation of the value function, where the risk functional is taken with respect to the posterior distribution supported on a continuous set $\Theta$. Hence, computation of the value function usually requires Monte Carlo simulation from the posterior distribution or other approximation techniques. Second, the posterior distribution $\mu_t$ in general could be infinite dimensional due to Bayesian updating. We circumvent this difficulty by considering conjugate distributions (cf. Chapter 5 in Schlaifer and Raiffa [1961]), where the posterior distribution falls into the same parametric family as the prior distribution, and hence reduce the dimension of the posterior distributions to a finite (low) dimension of posterior parameters. For example, a uni-variate normal distribution is a conjugate prior and is characterized by a two-dimensional vector, namely mean and variance; hence, updating the posterior of normal distribution boils down to updating the mean and variance. However, the parameters of the posteriors are usually continuous valued, and therefore, the BR-MDP with the state $(s_t, \mu_t)$ is in fact a continuous-state MDP. Solving a continuous-state MDP usually requires discretizing the continuous state in some way, leading to a large (finite) state space.

3.1 Nested simulation optimization for general risk functions

In view of the first challenge mentioned above, we propose a nested simulation optimization (NSO) approach to approximating the value function, presented in Algorithm 2, which works for a general parameter space $\Theta$ and general risk functional $\rho$. Taking $\rho$ to be VaR as an example, in the risk functional approximation step, we sort the sample averages in ascending order $\bar{H}(\theta^{(1)}, a), \ldots, \bar{H}(\theta^{(N)}, a)$, and set $\bar{V}_t(s_t, \mu_t, a) := \bar{H}(\theta^{(N)}, a)$; for CVaR, we set $\bar{V}_t(s_t, \mu_t, a) := \frac{1}{1-\alpha} \sum_{i=0}^{N} \bar{H}(\theta^{(i)}, a)$. The convergence of Algorithm 2 is implied by Theorem 3.4 in Zhu et al. [2020]. In particular, when the inner-layer sampling budget $K$ and outer-layer sampling budget $N$ go to infinity simultaneously and satisfy certain mild conditions, the approximate risk-adjusted optimal value function converges to the true risk-adjusted optimal value function almost surely.

3.2 UCB-based adaptive sampling algorithm for the risk-neutral case

As mentioned in the second challenge above, the MDP we face potentially has a large state space due to the discretization of the posterior. However, we note that the finite action space is relatively small compared to the discretized state space. This gives us a unique opportunity to simulate every action and determine the best action in a probabilistic sense. Finding the optimal action in a single stage in this case is essentially a multi-armed bandit (MAB) problem. Motivated by this observation, we propose an upper confidence bound (UCB)-based adaptive sampling procedure for the risk-neutral case (i.e., the risk functional $\rho$ is an expectation). In particular, our procedure is an extension of the adaptive multistage sampling (AMS) algorithm proposed by Chang et al. [2005], which is a precursor of the Monte Carlo Tree Search (cf. Kocsis and Szepesvari [2006]) used in Alpha-Go (cf. Silver et al.)
It has been shown that the time complexity of AMS algorithm is independent of the state space, so the algorithm works well in the MDPs with large state space and small action space. To proceed, we first make the following assumptions.

Assumption 3.1.

- The risk functional $\rho = E$, i.e., we only consider risk-neutral case.
- The parameter space $\Theta$ is finite, i.e., $\Theta = \{\theta_1, \ldots, \theta_k\}$. Moreover, $\theta^c \in \Theta$.
- The prior distribution has a positive probability mass on $\theta^c$, i.e., $\mu_{-1}(\theta^c) > 0$.

Let $N_{a,t}^{\theta,\mu}$ denote the number of $\xi$ samples drawn (so far) from the distribution $f(\cdot; \theta)$ given a fixed $\theta$, for each action $a \in A$, augmented state $(s, \mu)$, and stage $t$. Let $\bar{n}$ denote the overall number of samples (so far), i.e., $\bar{n} = \sum_{a \in A} N_{a,t}^{\theta,\mu}$. We then present Algorithm 3 below.

Algorithm 2: An NSO approach to approximating value functions of BR-MDPs in stage $t$

**input**: inner-layer sampling budget $K$, outer-layer sampling budget $N$, state $s_t$, posterior distribution $\mu_t$, risk-adjusted optimal value function $V_{t+1}^*(s_t, \mu_t)$

**output**: approximate value function $\hat{V}_t(s_t, \mu_t)$ and corresponding policy $\pi_t^*(s_t, \mu_t)$

for $a \in A$ do
  sample $\theta_1, \ldots, \theta_N$ from posterior distribution $\mu_t$;
  for each sample $\theta_i$, $i = 1, \ldots, N$, sample $\xi_1^i, \ldots, \xi_K^i$ from distribution $f(\cdot; \theta_i)$;
  for each $\xi_i^j$, compute next state $s_{t+1}^i = g(s_t, a_t, \xi_i^j)$ and next posterior distribution $\mu_{t+1}(\theta) = \frac{\mu_t(\xi_i^j)}{\mu_t(f(\xi_i^j) \theta)}$, and evaluate $h_i(\xi_i^j, a) := C(s_t, a, \xi_i^j) + \gamma V_{t+1}^*(s_{t+1}^i, \mu_{t+1})$;
  compute sample average $H_i(\theta_i, a) := \frac{1}{K} \sum_{j=1}^K h_i(\xi_i^j, a)$ as approximation for $E_{\xi_0} \left[ C(s_t, a, \xi) + \gamma V_{t+1}^*(s_{t+1}^i, \mu_{t+1}) \right]_{s_t, \mu_t, a}$;
  compute $\hat{V}_t(s_t, \mu_t, a) := \rho_0 \left[ H_i(\theta_i, a) \right]$ (risk functional approximation);
end

return $\pi_t^*(s_t, \mu_t) := a_t^* = \arg \min_{a \in A} \hat{V}_t(s_t, \mu_t, a)$ as the policy;

return $\hat{V}_t(s_t, \mu_t) = \hat{V}_t(s_t, \mu_t, a_t^*)$ as the approximate risk-adjusted optimal value function.

Algorithm 3: A UCB-based adaptive sampling approach to approximate value function in stage $t$ (when $\rho$ is expectation)

**input**: state $(s, \mu) \in (S \times \mathcal{M})$, total sampling budget $N_t$

**output**: approximate value function $\hat{V}_t(s, \mu)$

**initialization**: take each action $a \in A$ sequentially once at state $(s, \mu)$. For each $\theta \in \Theta$, sample one $\xi$ from the distribution $f(\cdot; \theta)$. Set $N_{a,t}^{\theta,\mu} = 1$ for all actions, $\bar{n} = |A|$. Set

$\hat{Q}_t(s, \mu, a) := \rho_0 \left( \frac{1}{N_{a,t}^{\theta,\mu}} \sum_{i=1}^{N_{a,t}^{\theta,\mu}} \left( C(s, a, \xi_i^j) + \gamma \hat{V}_{t+1}^*(s_i^j, \mu_i^j) \right) \right)$, where

$s_i^j = g(s_a, a, \xi_i^j), \mu_i^j(\theta) = \frac{\mu_t(\xi_i^j)}{\mu_t(f(\xi_i^j) \theta)}$.

for $\bar{n} < N_t$ do
  take an action $a^* = \arg \min_{a \in A} \left( \hat{Q}_t(s, \mu, a) - \sqrt{\frac{2 \ln t}{N_{a,t}^{\theta,\mu}}} \right)$;
  update $\hat{Q}_t(s, \mu, a^*)$ with additional $\xi$ sample;
  set $N_{a^*,t}^{\theta,\mu} = N_{a^*,t}^{\theta,\mu} + 1$ and $\bar{n} := \bar{n} + 1$.
end

return $\hat{V}_t(s, \mu) := \sum_{a \in A} \frac{N_{a,t}^{\theta,\mu}}{N_{a^*,t}^{\theta,\mu}} \hat{Q}_t(s, \mu, a)$.

Compared with the AMS algorithm in [Chang et al., 2005], Algorithm 3 considers the parameter uncertainty in MDPs and incorporates the risk associated with uncertain parameters into the value function. The convergence analysis of Algorithm 3 is shown in the next section.
3.3 Convergence Analysis of Algorithm

The convergence analysis of Algorithm 3 is composed of two parts. First, we consider the risk-adjusted MAB problem, where there are a finite number of scenarios within each bandit machine, and the risk is associated with the scenarios. This is similar to the concept we develop in MDPs with parameter uncertainty in previous sections. We then carry out finite time analysis of the risk-adjusted MAB, and derive the regret bound for the problem. Second, we use the regret bound to show the convergence of the one-stage algorithm, and extend the result to the multi-stage algorithm, i.e., Algorithm 3.

3.3.1 Introduction to risk-adjusted MAB

We consider MAB with nested structure. In particular, there are L bandit machines. Within each machine, there are k scenarios, denoted by θ1, · · · , θk. At each time t = 1, 2, · · · , we play machine At, where At ∈ {1, · · · , L} denotes the machine played at time t, and receive a random cost vector X̄ι(t) := (X̄ι,1,t, · · · , X̄ι,k,t), where X̄ι,1,t, · · · , X̄ι,k,t are independent, each with unknown distribution that has unknown mean m̄ι,1, · · · , m̄ι,k. For each i = 1, · · · , K, suppose the number of times machine i has been played is ni, define the average cost (over time) as X̄(ni) := (X̄1(ni), · · · , X̄K(ni)) = 1 ni n ∑ i=1 (X̄ι,1,t, · · · , X̄ι,k,t). The risk functional with respect to the scenario is then applied to the average cost, and we have the risk-adjusted average cost ρθ(X̄1(ni), · · · , X̄K(ni)), where we use θ to denote the scenario. For example, when ρ is expectation, the averaged cost is further averaged over the k scenarios, i.e., Eθ( ¯X(1) = 1 k ∑ j=1 X̄j(ni)). When ρ is VaR, the averaged cost is rearranged in ascending order such that X̄(1)(ni) < · · · < X̄(k)(ni), and VaRθ(X̄1(ni), · · · , X̄K(ni)) = X̄∗ (n). MAB under risk criteria has received increasing attention recently. Sani et al. [2012] consider the mean-variance risk criterion and present MV-UCB algorithm. Vakili and Zhao [2016] further complete the regret analysis of MV-UCB algorithm. Galichet et al. [2013] consider the CVaR risk criterion and present MaRaB algorithm. Maillard [2013] considers entropic risk measure and presents RA-UCB algorithm. Cassel et al. [2018] provide a more systematic approach to analyzing general risk criteria within the stochastic MAB formulation. However, in the aforementioned work, the risk measure is applied to the costs collected over time. While in this paper, the risk functional is applied to an inner expectation, and the risk is taken with respect to the scenarios within each machine. We present the finite time analysis of the risk-adjusted MAB in the next section, by extending the analysis in Auer et al. [2002].

3.3.2 Finite time analysis of the risk-adjusted MAB

Let vi = ρθ(m̄ι,1, · · · , m̄ι,k), v∗ = min i=1,...,L vi. Define the optimality gap ∆i = vi − v∗. The goal of risk-adjusted MAB is to find the machine with the lowest risk-adjusted average cost, or to minimize the regret, defined as

\[ R(n) = \sum_{j=1}^{L} \rho_{\theta}(E(\sum_{t=1}^{n} X̄j_1(A_t = j))) - v^* n. \]  (5)

We show in the next lemma that the risk-adjusted regret can be decomposed as the sum of optimality gap ∆i times the expected times a machine has been played, where the sum is over all L machines. So we can bound the regret by simply bounding the expected times a machine has been played.

Definition 3.2 (Positive homogeneity). Let h be any function on θ. A risk functional is said to have positive homogeneity if for every a > 0, ρθ(ah(θ1), · · · , ah(θk)) = apθ(h(θ1), · · · , h(θk)).

Lemma 3.3 (Regret decomposition). Suppose the risk functional ρ has positive homogeneity. Denote by Ti(n) the number of times machine i has been played after a total of n plays. The risk-adjusted regret specified in (5) can be decomposed as

\[ R(n) = \sum_{i=1}^{L} \Delta_i E[T_i(n)]. \]  (6)

Proof. Recall that R(n) = ∑Ij=1 ρθ(E(∑t=1^n X̄j_1(A_t = j))) − v^* n. Knowing A_t, we have

E[X̄j_1(A_t = j)|A_t] = 1{A_t = j}m_j,
which implies that

$$\mathbb{E}[X^j_t 1\{A_t = j\}] = \mathbb{E}[1\{A_t = j\}]m^j,$$

and

$$\mathbb{E}\left[\sum_{t=1}^n 1\{A_t = j\}\right] = \mathbb{E}[T_j(n)].$$

Now we have

$$\sum_{j=1}^L \rho_\theta(\mathbb{E}\left[\sum_{t=1}^n X^j_t 1\{A_t = j\}\right]) = \sum_{j=1}^L \rho_\theta(\mathbb{E}[T_j(n)]m^j)$$

$$= \sum_{j=1}^L \rho_\theta(\mathbb{E}[X^j])\mathbb{E}[T_j(n)],$$

where we use positive homogeneity in the last equality. Also, note that $\sum_{j=1}^L T_j(n) = n$, so $\sum_{j=1}^L \mathbb{E}[T_j(n)] = n$. The regret can then be decomposed as $R(n) = \sum_{i=1}^L \Delta_i \mathbb{E}[T_i(n)]$, where $\Delta_i = \bar{v}_i - v^*$. \qed

We then present the UCB algorithm that determines which machine to play next based on the past actions and obtained costs. Due to technical difficulty, we only consider the risk-neutral case where $\rho$ is expectation, i.e., the average cost over $k$ scenarios.

**Algorithm 4: UCB algorithm for risk-adjusted MAB**

**Initialization:** Play each machine once. In each play, we receive a random cost vector of dimension $k$;

**Loop:** Play machine $j$ that minimizes $\mathbb{E}_\theta(\bar{X}_i^j(n_j), \cdots, \bar{X}_k^j(n_j)) - \sqrt{\frac{2\ln n}{n_j}}$, where $\bar{X}_i^j(n_j)$ is the average cost obtained from machine $j$ and scenario $i, j = 1, \cdots, L, i = 1, \cdots, k$; $n_j$ is the number of times machine $j$ has been played so far; $n$ is the overall number of plays done so far.

---

Before we present the regret bound theorem, we first present one important lemma that shows the concentration bound for the estimator $\mathbb{E}_\theta(\bar{X}_i^j(n_j), \cdots, \bar{X}_k^j(n_j))$ that has been shown in Algorithm 4.

**Lemma 3.4 (Expectation convergence).** Let $\{\bar{X}_{1,i}\}, \cdots, \{\bar{X}_{k,i}\}$ be $k$ sequences. Each sequence $\{\bar{X}_{i,t}\}_{t}$ is the sample average with $\bar{X}_{i,t} = \frac{1}{t} \sum_{s=1}^t X_{i,s}$, where the cost at each time stage $s$ is $X_{i,s} \sim \mathcal{D} \zeta_i$ with mean $m_i$, $i = 1, \cdots, k$. The cost $X_{i,s}$ is assumed to fall into range $[0,1]$, $\forall i \in \{1, \cdots, L\}, \forall s \in \{1, \cdots, t\}$. At each time $t$, we take the average of $\bar{X}_{i,t}, i = 1, \cdots, k$, and define a new sequence $Y_t := \mathbb{E}(\bar{X}_{1,t}, \cdots, \bar{X}_{k,t}) = \frac{1}{k} \sum_{i=1}^k \bar{X}_{i,t}$. Let $\bar{m} = \frac{1}{k} \sum_{i=1}^k m_i$. Then

$$\lim_{t \to \infty} Y_t = \bar{m} \text{ a.s..}$$

Moreover, for $\alpha \geq 0$, we have the following concentration bound for $Y_t$:

$$\mathbb{P}(Y_t \geq \bar{m} + \alpha) \leq \exp(-2\alpha^2),$$

$$\mathbb{P}(Y_t \leq \bar{m} - \alpha) \leq \exp(-2\alpha^2).$$

**Proof.**

$$\lim_{t \to \infty} Y_t = \lim_{t \to \infty} \frac{1}{k} \sum_{i=1}^k \bar{X}_{i,t}$$

$$= \lim_{t \to \infty} \frac{1}{k} \sum_{i=1}^k \frac{1}{t} \sum_{s=1}^t X_{i,s}$$

$$= \lim_{t \to \infty} \frac{1}{t} \sum_{s=1}^t \left( \frac{1}{k} \sum_{i=1}^k X_{i,s} \right).$$
By Strong Law of Large Number (SLLN) and the exchange of the limit and finite sum, we have
\[ \lim_{t \to \infty} \frac{1}{t} \sum_{s=1}^{t} \{ (\frac{1}{k} \sum_{i=1}^{k} X_{i,s}) = \frac{1}{k} \sum_{s=1}^{t} X_{i,s} \} = \frac{1}{k} \sum_{i=1}^{k} m_i \text{ a.s.} \]
Let \( X_s = \frac{1}{k} \sum_{i=1}^{k} X_{i,s} \sim \zeta \) with mean \( \bar{m} \). For \( a \geq 0 \),
\[
P(Y_t \geq \bar{m} + a) = \mathbb{P}(\frac{1}{l} \sum_{s=1}^{t} \{ \frac{1}{k} \sum_{i=1}^{k} X_{i,s} \} \geq \bar{m} + a)
\]
\[
= \mathbb{P}(\frac{1}{l} \sum_{s=1}^{t} X_s \geq \bar{m} + a)
\]
\[
\leq \exp(-2tl^2),
\]
where the last inequality follows from Hoeffding’s inequality as we assume the cost \( X_{i,s} \in [0, 1] \). \( \square \)

**Remark 3.5.** If we consider other risk-averse functionals such as VaR, CVaR, mean-variance, we can easily show the convergence of the estimator \( \rho_0(\tilde{X}_1(n), \cdots, \tilde{X}_L(n)) \) by continuous mapping theorem. However, the concentration bound will no longer be valid.

We now present the regret bound theorem for the risk-neutral case.

**Theorem 3.6 (Regret bound).** If the Algorithm 4 is run on \( L \) machines and \( k \) scenarios within each machine, where the cost falls into range \( [0, 1] \), then its risk-adjusted expected regret after any number \( n \) of plays is at most
\[
\left[ 8 \sum_{i: v_i > n^*} \left( \frac{\ln n}{\Delta_i} \right) \right] + \left( 1 + \frac{\pi^2}{3} \right) \left( \sum_{j=1}^{L} \Delta_j \right).
\]

**Proof.** Let \( l \) be any positive integer that is less than \( n \). Let \( c_{t,s} = \sqrt{\frac{2\ln t}{s}} \), \( \tilde{X}^i(t) = \frac{1}{t} \sum_{s=1}^{t} (X_{i,s}^1, \cdots, X_{i,s}^k) \), \( \tilde{X}^i(t) = \frac{1}{t} \sum_{s=1}^{t} (X_{i,s}^1, \cdots, X_{i,s}^k) \), where \( i_s = \arg \min_{i=1, \ldots, L} v_i \).

\[
T_i(n) \leq l + \sum_{t=L+1}^{n} 1\{A_t = i, T_i(t-1) \geq l\}
\]
\[
\leq l + \sum_{t=L+1}^{n} 1\{\rho_0(\tilde{X}^i(T^*(t-1))) - c_{t-1,s} \geq \rho_0(\tilde{X}^i(T_i(t-1))) - c_{t-1,s}, T_i(t-1) \geq l\}
\]
\[
\leq l + \sum_{t=L+1}^{n} \max_{0<s<t} \rho_0(\tilde{X}^i(s)) - c_{t-1,s} \geq \min_{1 \leq s < t} \rho_0(\tilde{X}^i(s)) - c_{t-1,s}, T_i(t-1) \geq l\}
\]
\[
\leq l + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} 1\{\rho_0(\tilde{X}^i(s)) - c_{t,s} \geq \rho_0(\tilde{X}^i(s_i)) - c_{t,s}\}
\]

Observing that \( \rho_0(\tilde{X}^i(s)) - c_{t,s} \geq \rho_0(\tilde{X}^i(s_i)) - c_{t,s} \) implies one of the following three inequalities must hold:

- \( \rho_0(\tilde{X}^i(s)) \geq \rho_0(\tilde{E}(\tilde{X}^i)) + c_{t,s} \);
- \( \rho_0(\tilde{X}^i(s_i)) \leq \rho_0(\tilde{E}(\tilde{X}^i)) - c_{t,s} \);
- \( \rho_0(\tilde{E}(\tilde{X}^i)) > \rho_0(\tilde{E}(\tilde{X}^i)) - 2c_{t,s} \).

Since we only consider \( \rho = \tilde{E} \), the probability of first inequality can be bounded using the concentration bound given in Lemma [3.3].
\[
P(\rho_0(\tilde{X}^i(s)) \geq \rho_0(\tilde{E}(\tilde{X}^i)) + c_{t,s}) \leq \exp(-2s \frac{2\ln t}{s}) = t^{-4}.
\]
Similarly, the probability of second inequality can be bounded as
\[ P(\rho_\theta(\hat{X}^i(s_i)) \leq \rho_\theta(\mathbb{E}(X_i)) - c_{t,s_i}) \leq \exp(-2s_i \frac{2 \ln t}{s_i}) = t^{-4}. \]

Finally, notice that when the third inequality does not hold, i.e.,
\[ \rho_\theta(\mathbb{E}(X^*)) - \rho_\theta(\mathbb{E}(X_i)) + 2c_{t,s_i} \leq 0, \]
it implies that \( s_i \geq \frac{8 \ln t}{\Delta^2} + \forall t \leq n \). Note that the summation over \( s_i \) is from \( l \) to \( t - 1 \), so when \( l = \lceil \frac{8 \ln n}{\Delta^2} \rceil \), the third inequality does not hold, such that the first inequality or the second inequality must hold. Therefore, we have
\[
\mathbb{E}[T_i(n)] \leq \left[ \frac{8 \ln n}{\Delta^2} \right] + \sum_{t=1}^{\infty} \sum_{s=1}^{\lceil \frac{8 \ln n}{\Delta^2} \rceil} \sum_{s_i=1}^{t-1} (P(\rho_\theta(\hat{X}^i(s)) \leq \rho_\theta(\mathbb{E}(X^*)) - c_{t,s}) + P(\rho_\theta(\hat{X}^i(s_i)) \geq \rho_\theta(\mathbb{E}(X_i)) + c_{t,s_i}))
\leq \left[ \frac{8 \ln n}{\Delta^2} \right] + \sum_{t=1}^{\infty} \sum_{s=1}^{t} \sum_{s_i=1}^{t} 2t^{-4}
\leq \frac{8 \ln n}{\Delta^2} + 1 + \frac{\pi^2}{3}.
\]

Finally, with the regret decomposition lemma, we obtain the required result. \( \square \)

### 3.3.3 Convergence analysis for one-stage and multi-stage algorithm

To show the convergence (under BR-MDP formulation) of Algorithm 3, we first need the following lemma for the one-stage version of Algorithm 3 which would simply replace the estimated next-stage value function \( \hat{V}_{i+1}(s', \mu') \) by a stochastic value function \( U(s', \mu') \). Here \( U(s', \mu') \) is a non-negative random variable with unknown distribution and bounded above by a constant \( U_{\text{max}} \) for all augmented states. Without loss of generality, we assume \( U_{\text{max}} \leq 1 \) (in Lemma 3.7) and \( C_{\text{max}} \leq \frac{1}{\Delta} \) (in Theorem 3.8) for the convenience of theoretical analysis, since we can always normalize the cost function to satisfy these assumptions. Note that these assumptions can be further relaxed. For example, \( U(s', \mu') \) can be set to the interval \([a, b]\), where \( 0 < a < b < \infty \). Then applying the Hoeffding’s inequality would give a slight different regret bound in Theorem 3.6

**Lemma 3.7 (One-stage convergence).** Given a stochastic value function \( U \) defined over \( S \times M \) with \( U_{\text{max}} = \max_{s, \mu, a} \rho_\theta \mathbb{E}_{\xi}[|C(s, a, \xi) + \gamma U(s', \mu')|] \leq 1 \). Let \( n \) be the total sampling budget. Define
\[
V(s, \mu) = \min_a \rho_\theta \mathbb{E}_{\xi}[|C(s, a, \xi) + \gamma U(s', \mu')|] s, \mu, a,\]
where \( s' = g(s, a, \xi), \mu'(\theta) = \frac{\mu(\xi(\theta)|\xi(\theta))}{\int \mu(\xi(\theta))d\xi(\theta)}, \) then we have
\[
\lim_{n \to \infty} \mathbb{E}(\hat{V}(s, \mu)) = V(s, \mu).
\]

**Proof.** Consider risk-adjusted MAB with \(|A|\) machines and \( k \) scenarios within each machine. Successive plays of machine \( a \) yield costs vectors that are independent and identically distributed (for each scenario \( \theta \) with unknown risk-adjusted expectation:
\[
Q(s, \mu, a) = \rho_\theta \mathbb{E}_{\xi|\theta}[|C(s, a, \xi) + \gamma U(s', \mu')|] s, \mu, a\]
The corresponding risk-adjusted expected regret after \( n \) plays is given by
\[
R(n) = \sum_{a=1}^{|A|} Q(s, \mu, a) \mathbb{E} \left( T^{(s, \mu)}(n) \right) - V(s, \mu)n.
\]
We apply Theorem 3.6 and obtain the following upper bound on the regret:
\[
R(n) \leq \sum_{a:Q(s, \mu, a) > V(s, \mu)} \left[ \frac{8 \ln n}{Q(s, \mu, a) - V(s, \mu)} + \left( 1 + \frac{\pi^2}{3} \right) (Q(s, \mu, a) - V(s, \mu)) \right].
\]

10
Now it remains to show the convergence of $\mathbb{E}\left(\hat{V}(s, \mu)\right)$. The idea is to use decomposition. Define

$$
\hat{V}(s, \mu) = \sum_{a=1}^{|A|} T_a^{(s, \mu)}(n) Q(s, \mu, a).
$$

Note that

$$
\mathbb{E}\left(\hat{V}(s, \mu)\right) - V(s, \mu) = \mathbb{E}\left(\hat{V}(s, \mu)\right) - V(s, \mu)
$$

$$
+ \mathbb{E}\left(\sum_{a \in A} \frac{T_a^{(s, \mu)}(n)}{n} (\hat{Q}(s, \mu, a) - Q(s, \mu, a))\right)
$$

Define the set of sub-optimal actions for the augmented state $(s, \mu)$:

$$
\phi(s, \mu) = \{ a \in A : Q(s, \mu, a) > V(s, \mu) \}.
$$

Define

$$
\alpha(s, \mu) = \min_{a \in \phi(s, \mu)} (Q(s, \mu, a) - V(s, \mu)).
$$

Note that $0 < \alpha(s, \mu) \leq U_{\text{max}}$ and $\max_a (Q(s, \mu, a) - V(s, \mu)) \leq U_{\text{max}}$. We then have

$$
\mathbb{E}\left(\hat{V}(s, \mu)\right) - V(s, \mu) = \frac{R(n)}{n}
$$

$$
\leq \frac{8(|A| - 1) \ln n}{n \alpha(s, \mu)} + \frac{\pi^2}{3} \left(\frac{|A| - 1}{n} U_{\text{max}}\right)
$$

$$
\leq \frac{C_1 \ln n}{n} + \frac{C_2}{n},
$$

for some constants $C_1, C_2 > 0$. Also, since the regret is bounded by zero from below, we have

$$
\mathbb{E}\left(\hat{V}(s, \mu)\right) - V(s, \mu) \geq 0.
$$

It only remains to bound the second term in $[7]$. Let $\mu_a(s, \mu) = \mathbb{E}_{\xi \mid \theta}[C(s, a, \xi) + \gamma U(s', \mu')|s, \mu, a]$. Note that

$$
\mathbb{E}\left(\sum_{a \in A} \frac{T_a^{(s, \mu)}(n)}{n} (Q(s, \mu, a) - Q(s, \mu, a))\right)
$$

$$
= \mathbb{E}\left(\sum_{a \in A} \frac{T_a^{(s, \mu)}(n)}{n} \rho \left(\frac{1}{T_a^{(s, \mu)}(n)} \sum_{i=1}^{T_a^{(s, \mu)}(n)} (C(s, a, \xi) + \gamma U(s_i', \mu'))\right) - \rho \left(\mu_a(s, \mu)\right)\right)
$$

$$
= \frac{1}{n} \left(\sum_{a \in A} \mathbb{E} \left(\rho \left(\frac{1}{T_a^{(s, \mu)}(n)} \sum_{i=1}^{T_a^{(s, \mu)}(n)} (C(s, a, \xi) + \gamma U(s_i', \mu'))\right)\right) - \sum_{a \in A} \mathbb{E} \left(T_a^{(s, \mu)}(n) \rho \left(\mu_a(s, \mu)\right)\right)\right)
$$

$$
= \frac{1}{n} \left(\sum_{a \in A} \rho \left(\sum_{i=1}^{T_a^{(s, \mu)}(n)} (C(s, a, \xi) + \gamma U(s_i', \mu'))\right)\right) - \sum_{a \in A} \mathbb{E} \left(T_a^{(s, \mu)}(n) \rho \left(\mu_a(s, \mu)\right)\right)
$$

$$
= 0.
$$

In the above equations, $\xi^{i, t, \omega} \sim f(\cdot; \theta)$ conditional on $\theta$, $s_i' = g(s, a, \xi), \mu'_i(\theta) = \frac{\mu(\theta) f(\xi_i(\theta))}{\int \mu(\theta) f(\xi_i(\theta)) d\theta}$. Note that in the second last equality, we use two facts. First, $\xi$ is independent of $T_a^{(s, \mu)}(n)$. So we can apply Wald’s identity. Second, $\rho$ and expectation can be interchanged. Note that if the risk functional is VaR or CVaR, this condition is not satisfied, and we will not have the last equality. Now [7] is only bounded by the first term, and we get

$$
\mathbb{E}\left(\hat{V}(s, \mu)\right) - V(s, \mu) \leq O\left(\frac{\ln n}{n}\right).
$$

As $n \to \infty$, we have $\mathbb{E}\left(\hat{V}(s, \mu)\right) \to V(s, \mu)$. \qed
We do it backwards inductively, and finally have the result

\[ \Phi \]

Let \( C_{\max} = \max_{t,a,T} C(s,a,\xi) \) and assume that \( C_{\max} \leq \frac{1}{T} \). Suppose Algorithm 3 is run with input \( N_t \) for stage \( t = 0, \cdots, T - 1 \). Then

\[
\lim_{N_0 \to \infty} \lim_{N_1 \to \infty} \cdots \lim_{N_{T-1} \to \infty} E \left[ \hat{V}_0(s, \mu) \right] = V_0^*(s, \mu).
\]

**Proof.** We approximate the value function backwards.

\[
\hat{V}_{T-1}(s, \mu) = \sum_{a \in A} \frac{N_a^{(s,\mu)}}{N_{T-1}} \left( \rho_\theta \left( \frac{1}{N_a^{(s,\mu)}} \sum_{i=1}^{N_a^{(s,\mu)}} \left( C(s, a, \xi_i) + \gamma \hat{V}_T(s_i', \mu_i') \right) \right) \right) 
\]

\[
\leq \sum_{a \in A} \frac{N_a^{(s,\mu)}}{N_{T-1}} (\rho_\theta(C_{\max} + 0)) = C_{\max}.
\]

Similarly,

\[
\hat{V}_{T-2}(s, \mu) = \sum_{a \in A} \frac{N_a^{(s,\mu)}}{N_{T-2}} \left( \rho_\theta \left( \frac{1}{N_a^{(s,\mu)}} \sum_{i=1}^{N_a^{(s,\mu)}} \left( C(s, a, \xi_i) + \gamma \hat{V}_{T-1}(s_i', \mu_i') \right) \right) \right) 
\]

\[
\leq \sum_{a \in A} \frac{N_a^{(s,\mu)}}{N_{T-2}} (\rho_\theta(C_{\max} + \gamma C_{\max})) = (1 + \gamma)C_{\max}.
\]

Therefore we have \( \hat{V}_i(s, \mu) \leq C_{\max}(T - i), i = T - 1, \cdots, 0 \). Note that \( U_{\max} \) in Lemma 3.7 satisfies \( U_{\max} = C_{\max}(T - i) \leq 1 \). So we can apply Lemma 3.7 and we can get, as \( N_{T-1} \to \infty \),

\[
E \left[ \hat{V}_{T-1}(s, \mu) \right] \to V_{T-1}^*(s, \mu).
\]

We do it backwards inductively, and finally have the result

\[
\lim_{N_0 \to \infty} \lim_{N_1 \to \infty} \cdots \lim_{N_{T-1} \to \infty} E \left[ \hat{V}_0(s, \mu) \right] = V_0^*(s, \mu).
\]

Finally, we show the error bound for the value function approximation in the next proposition.

**Proposition 3.9.** Suppose Algorithm 3 is run with input \( N_t \) for stage \( t = 0, \cdots, T - 1 \). Then

\[
E \left[ \hat{V}_0(s, \mu) \right] - V_0^*(s, \mu) \leq O(\sum_{t=0}^{T-1} \frac{\ln N_t}{N_t}).
\]

**Proof.** Consider an operator defined in (12). Define \( \Phi_t(s, \mu) = E \left[ \hat{V}_t(s, \mu) \right], t = 1, \cdots, T - 1 \), and \( \Phi_T(s, \mu) = V_T^*(s, \mu) = 0 \). We replace the stochastic value function \( U \) in Lemma 3.7 by \( \Phi_{t+1} \). By applying the operator \( T \) to \( \Phi_{t+1} \) and combining with inequality (8), we get

\[
\Phi_t(s, \mu) - (T \Phi_{t+1})(s, \mu) \leq O\left( \frac{\ln N_t}{N_t} \right), t = 0, \cdots, T - 1.
\]

This implies

\[
\Phi_0(s, \mu) \leq (T \Phi_1)(s, \mu) + O\left( \frac{\ln N_0}{N_0} \right), \quad (9)
\]

\[
\Phi_0(s, \mu) \leq (T \Phi_1)(s, \mu) + O\left( \frac{\ln N_0}{N_0} \right).
\]
\[ \Phi_1(s, \mu) \leq (T \Phi_2)(s, \mu) + O\left(\frac{\ln N_1}{N_1}\right). \] (10)

By applying the monotonicity property of the operator \( T \) that will be shown in Lemma 4.4 to (10), we have

\[ (T \Phi_1)(s, \mu) \leq (T^2 \Phi_2)(s, \mu) + O\left(\ln N_1\right). \] (11)

Combining (9) and (11), we get

\[ \Phi_0(s, \mu) \leq (T \Phi_2)(s, \mu) + O\left(\ln N_1\right). \] (12)

Repeating this process, we obtain

\[ \Phi_0(s, \mu) \leq (T^T \Phi_T)(s, \mu) + O\left(\sum_{t=0}^{T-1} \ln N_t\right). \]

Therefore, we get

\[ E\left[\hat{V}_0(s, \mu)\right] - V^*_0(s, \mu) \leq O\left(\sum_{t=0}^{T-1} \ln N_t\right). \]

\[ \square \]

4 Bayesian risk MDPs: the infinite-horizon case

In this section, we show how to compute the risk-adjusted optimal value function to the BR-MDP in infinite horizon. We generalize the formulation in (2) to infinite horizon with \( T \to \infty \) and discounted factor \( \gamma \in (0, 1) \). We consider a general risk functional \( \rho \). We first make the following assumptions on the cost function \( C \) and the risk functional \( \rho \).

**Assumption 4.1.**

- The cost at each stage is uniformly bounded by a positive constant \( Z \), i.e., \( |C(s, a, \xi)| \leq Z \), \( \forall s \in S, a \in A, \xi \in \Xi \).
- The risk functional \( \rho \) satisfies \( \rho(X + c) = c + \rho(X) \), where \( c \) is any constant, and \( X \) is a random variable.

We consider the following operator obtained by applying the dynamic programming mapping to the risk-adjusted value function \( V \).

**Definition 4.2.** Let \( B(S, M) \) be the space of real-valued bounded measurable functions on \( S \times M \). For any bounded value function \( V \in B(S, M) \), define an operator \( T : B(s, \mu) \to B(s, \mu) \) as:

\[ (TV)(s, \mu) = \min_{a \in A} \rho_a E_{\xi|\theta}[C(s, a, \xi) + \gamma V(s', \mu')|s, \mu, a], \] (12)

where state transition is implied by \( s' = g(s, a, \xi) \), \( \mu'(\theta) = \frac{\mu(\theta) f(\xi|\theta) d\theta}{\int \mu(\theta) f(\xi|\theta) d\theta} \).

The next proposition shows that for any initial value function \( V \), applying the mapping \( T \) in (12) to \( V \) for an infinite number of times will give us the risk-adjusted optimal value function.

**Proposition 4.3.** Under Assumption 4.1, the risk-adjusted optimal value function \( V^*(s, \mu) \) to the infinite horizon problem is

\[ V^*(s, \mu) = \lim_{N \to \infty} (T^N V)(s, \mu), \]

\( \forall s \in S, \mu \in M \), for any initial value function \( V \).
We show two key properties of the mapping $T$ we have obtained the result.

\[ V^\pi(s_0, \mu_0) = \rho_{\theta_0} \mathbb{E}_{\xi_0|\theta_0} [C(s_0, a_0, \xi_0) + \gamma \rho_{\theta_1} \mathbb{E}_{\xi_1|\theta_1} [C(s_1, a_1, \xi_1) + \cdots + \gamma \rho_{\theta_K} \mathbb{E}_{\xi_K|\theta_K} [C(s_K, a_{K-1}, \xi_{K-1}) + \gamma V^\pi(s_K, \mu_K)]]], \]

where $a_k = \pi(s_k, \mu_k)$, $s_{k+1} = g(s_k, a_k, \xi_k)$, $\mu_{k+1} = f^{\theta}(\xi_k|\theta)$, $\forall k = 0, \cdots, K - 1$.

\[ V^\pi(s_K, \mu_K) = \lim_{N \to \infty} \rho_{\theta_N} \mathbb{E}_{\xi_N|\theta_N} [C(s_N, a_N, \xi_N) + \cdots + \gamma \rho_{\theta_N} \mathbb{E}_{\xi_N|\theta_N} [C(s_N, a_N, \xi_N)]] \]

Since we assume the cost at each stage is bounded, i.e., $|C(s_k, a_k, \xi_k)| \leq Z$ for all $k$, we then have $|V^\pi(s_k, \mu_K)| \leq Z \sum_{k=1}^{\infty} \gamma^k = \frac{\gamma K}{1 - \gamma} Z$. Further, since for any constant $c$, $\rho_{\theta} \mathbb{E}_{\xi|\theta} [C(s, a, \xi) + c] = c + \rho_{\theta} \mathbb{E}_{\xi|\theta} [C(s, a, \xi)]$, we have

\[ V^\pi(s_0, \mu_0) - \frac{\gamma K}{1 - \gamma} Z - \gamma^K \max_{s_0 \in S, \mu_0 \in M} |V(s_0, \mu_0)| \leq \rho_{\theta_0} \mathbb{E}_{\xi_0|\theta_0} [C(s_0, a_0, \xi_0) + \gamma \rho_{\theta_1} \mathbb{E}_{\xi_1|\theta_1} [C(s_1, a_1, \xi_1) + \cdots + \gamma \rho_{\theta_K} \mathbb{E}_{\xi_K|\theta_K} [C(s_K, a_{K-1}, \xi_{K-1}) + \gamma V^\pi(s_K, \mu_K)]]] \leq V^\pi(s_0, \mu_0) + \frac{\gamma K}{1 - \gamma} Z + \gamma^K \max_{s_0 \in S, \mu_0 \in M} |V(s_0, \mu_0)|. \]

Taking minimum over policy $\pi$, we have, for all $s_0 \in S, \mu_0 \in M, K$,

\[ V^*(s_0, \mu_0) - \frac{\gamma K}{1 - \gamma} Z - \gamma^K \max_{s_0 \in S, \mu_0 \in M} |V(s_0, \mu_0)| \leq (T^K V)(s_0, \mu_0) \leq V^*(s_0, \mu_0) + \frac{\gamma K}{1 - \gamma} Z + \gamma^K \max_{s_0 \in S, \mu_0 \in M} |V(s_0, \mu_0)|. \]

Finally, as $K \to \infty$, $V^*(s_0, \mu_0) = \lim_{K \to \infty} (T^K V)(s_0, \mu_0)$. Since it holds for all $s_0 \in S, \mu_0 \in M$, we obtain the result.

We show two key properties of the mapping $T$ in Eq. 12 in the next two lemmas, which lead to the risk-adjusted Bellman equation for the infinite horizon problem in the next theorem.

**Lemma 4.4 (Monotonicity).** If $V(s, \mu) \leq V'(s, \mu)$, then $(T^k V)(s, \mu) \leq (T^k V')(s, \mu)$, $\forall s \in S, \mu \in M, k \in \mathbb{N}$.

**Proof.** We expand $(T^k V)$ up to stage $k$:

\[ (T^k V)(s, \mu) = \min_{a_0} \rho_{\theta_0} \mathbb{E}_{\xi_0|\theta_0} [C(s_0, a_0, \xi_0) + \gamma \min_{a_1} \rho_{\theta_1} \mathbb{E}_{\xi_1|\theta_1} [C(s_1, a_1, \xi_1) + \cdots + \gamma \min_{a_{k-1}} \rho_{\theta_{k-1}} \mathbb{E}_{\xi_{k-1}|\theta_{k-1}} [C(s_{k-1}, a_{k-1}, \xi_{k-1}) + \gamma V(s_k, \mu_k)]]]. \]

Note that as the terminal value function increases uniformly for all $s \in S, \mu \in M$, so will the $k$-stage costs. Therefore, if $V(s, \mu) \leq V'(s, \mu)$, we have $(T^k V)(s, \mu) \leq (T^k V')(s, \mu)$.

**Lemma 4.5 (Weight-shift).** $(T^k (V + rE))(s, \mu) = (T^k V)(s, \mu) + \gamma^k r$, $\forall s \in S, \mu \in M, k \in \mathbb{N}$, $r \in \mathbb{R}$, and $E$ is matrix of ones.

**Proof.** Applying the operator $T$ to $V + rE$:

\[ (T(V + rE))(s, \mu) = \min_{a} \rho_{\theta} \mathbb{E}_{\xi|\theta} [C(s, a, \xi) + \gamma (V(s', \mu') + r)] = \min_{a} \rho_{\theta} \mathbb{E}_{\xi|\theta} [C(s, a, \xi) + \gamma V(s', \mu')] + \gamma r = (TV)(s, \mu) + \gamma r, \forall s \in S, \mu \in M. \]

Then by induction on $k$, we obtain the result.
Theorem 4.6. The optimal value function $V^*$ satisfies the following risk-adjusted Bellman equation:

$$V^*(s, \mu) = \min_{a \in A} \mathbb{E}_{\xi}[\mathcal{C}(s, a, \xi) + \gamma V^*(s', \mu') | s, \mu, a],$$

and $V^*$ is the unique solution within the class of bounded functions.

Proof. From Proposition 4.3, we have $\forall s \in S, \mu \in M, k \in \mathbb{N}$ and any bounded value function $V$,

$$V^*(s, \mu) \leq \frac{\gamma^k}{1 - \gamma} Z \leq (T^k V)(s, \mu) \leq V^*(s, \mu) + \frac{\gamma^k}{1 - \gamma} Z.$$  

Applying the mapping $T$ to this relation and using Lemma 4.4 and Lemma 4.5, we have

$$(TV^*)(s, \mu) - \frac{\gamma^k}{1 - \gamma} Z \leq (T^{k+1} V)(s, \mu) \leq (TV^*)(s, \mu) + \frac{\gamma^k}{1 - \gamma} Z.$$  

As we take the limit as $k \to \infty$, we have $\lim_{k \to \infty} (T^{k+1} V)(s, \mu) = V^*(s, \mu)$ and $V^*(s, \mu) = (TV^*)(s, \mu)$ and we obtain the result. For the uniqueness, observe that if $V$ is bounded and satisfies $V = TV$, then $V = \lim_{k \to \infty} T^k V$, then by Proposition 4.3, we have $V = V^*$.

The next theorem shows the operator $T$ is actually a $\gamma$ contraction for $\| \cdot \|_\infty$ norm.

Theorem 4.7 (Contraction). The operator $T : \mathbb{R}^{|S| \times |M|} \to \mathbb{R}^{|S| \times |M|}$ is a $\gamma$ contraction for $\| \cdot \|_\infty$ norm, i.e., for any two bounded functions $V : S \times M \to \mathbb{R}$, $V' : S \times M \to \mathbb{R}$, $k \in \mathbb{N}$, we have

$$\max_{s, \mu \in M} |(T^k V)(s, \mu) - (T^k V')(s, \mu)| \leq \gamma^k \max_{s, \mu \in M} |V(s, \mu) - V'(s, \mu)|.$$  

Proof. Let $c = \max_{s, \mu \in M} |V(s, \mu) - V'(s, \mu)|$, we have

$$V(s, \mu) - c \leq V'(s, \mu) \leq V(s, \mu) + c.$$  

Applying $T^k$ in this relation and using Lemma 4.4 and Lemma 4.5, we have

$$(T^k V)(s, \mu) - \gamma^k c \leq (T^k V')(s, \mu) \leq (T^k V)(s, \mu) + \gamma^k c.$$  

It follows that

$$|(T^k V)(s, \mu) - (T^k V')(s, \mu)| \leq \gamma^k c.$$  

Corollary 4.8. For any initial bounded value function $V$, the convergence rate is shown to be

$$\max_{s, \mu \in M} |(T^k V)(s, \mu) - V^*(s, \mu)| \leq \gamma^k \max_{s, \mu \in M} |V(s, \mu) - V^*(s, \mu)|.$$  

Proof.

$$\max_{s, \mu \in M} |(T^k V)(s, \mu) - V^*(s, \mu)| = \max_{s, \mu \in M} |(T^k V)(s, \mu) - (TV^*)(s, \mu)|$$

$$= \max_{s, \mu \in M} |(T^k V)(s, \mu) - (T^k V^*)(s, \mu)|$$

$$\leq \gamma^k \max_{s, \mu \in M} |V(s, \mu) - V^*(s, \mu)|.$$  

Remark 4.9. Corollary 4.8 implies that applying the operator $T$ on any initial value function repeatedly, the resulting value function will converge to the optimal value function exponentially fast. Given any initial bounded value function $V$, applying mapping $T$ is equivalent to solving a nested simulation optimization problem, which can be solved approximately by Algorithm 2.
5 Numerical experiments

We illustrate the performance of our proposed algorithms in the finite-horizon case with two numerical examples, namely the inventory control problem (cf. Bertsekas et al. [2000]) and the path planning problem (cf. Xu and Mannor [2010]). We will show that our BR-MDP formulation can handle the parameter uncertainty properly, which leads to more robust policies compared to the empirical approach. All the experiments are run on a single Intel Xeon Gold 6134 CPU.

5.1 Inventory control problem

Let \( s_t \) denote the inventory level at the start of stage \( t \), where \( s_t \in \{0, 1, \cdots, S\} \) and \( S \) is the inventory storage capacity. Let \( a_t \) denote the order amount in stage \( t \), where \( a_t \in \{0, 1, \cdots, S-s_t\} \). Let \( \xi_t \) denote the demand in stage \( t \), and the demand is realized at the end of the stage. We assume the demands \( \xi_0, \ldots, \xi_{T-1} \) are independent and identically distributed according to a Poisson distribution with mean \( \theta^* \), which is unknown. Given \( s_t, a_t, \) and \( \xi_t \), the state (i.e., inventory level) at the start of the next stage is given by \( s_{t+1} = \max(s_t + a_t - \xi_t, 0) \). The cost \( C \) incurred at stage \( t \) is:
\[
C(s_t, a_t, \xi_t) = h \cdot \max(s_t + a_t - \xi_t, 0) + p \cdot \max(\xi_t - s_t - a_t, 0) + c \cdot a_t,
\]
where \( h \) is the holding cost per unit, \( p \) is the lost sale penalty cost per unit, \( c \) is the ordering cost per unit. To formulate the BR-MDP, we view the unknown parameter \( \theta^* \) as a random variable \( \theta \in \Theta \) and estimate it using the Bayesian posterior distribution \( \mu_t \) at stage \( t \), given a prior distribution \( \mu_{t-1} \). The posterior update is given by (4).

We consider BR-MDP with three risk functions including expectation, VaR and CVaR, and we run Algorithms 1 (for expectation) and 2 (for VaR and CVaR) to find the corresponding risk-adjusted optimal policy. As a comparison, we also consider the empirical approach, i.e., solve the MDP with the unknown parameter replaced by its maximum likelihood estimator (MLE) computed from the given data. For each formulation, we run \( J \) replications in total. For each replication \( j \), we generate an independent dataset of size \( H \), i.e., \( \phi^j = \{d^j_1, \ldots, d^j_H\} \sim f(\cdot; \theta^*) \). Let \( \hat{\pi}^j_t \) denote the respective optimal policy of each formulation, and \( \hat{V}_t^j \) denote the true performance of \( \hat{\pi}^j_t \), which is equivalent to the expected discounted total cost in (1) evaluated with the policy \( \hat{\pi}^j_t \) and true parameter \( \theta^* \). To assess the performance of these policies, we define a performance measure \( D \) by the average square-deviation in the evaluation of each policy from the optimal value function \( V^* \) computed by (1), i.e.,
\[
D = \frac{1}{J} \sum_{j=1}^{J} \left( \frac{V^*_j - V^*}{V^*} \right)^2.
\]
Therefore, a larger \( D \) value implies a more significant deviation from the true optimal value function in average and thus more risk of the corresponding formulation due to parameter uncertainty.

In implementation, the parameters are set as follows: \( T = 7 \), \( S = 3 \), \( s_0 = 1 \), \( \Theta = \{1,2,1.6,2,2.4,2.8\} \), \( \theta^* = 2 \), \( h = 4 \), \( p = 4 \), \( c = 1 \), \( J = 100 \), \( \alpha = 0.8 \), \( \mu_{-1} \) is the uniform distribution on the support in \( \Theta \). Note that in this case, \( V^* = 30.05 \). Figure[1] shows the histogram of the true performance of the optimal policies under different formulations (BR-MDP mean, BR-MDP VaR, BR-MDP CVaR, and empirical approach), where the data size is \( H = 20 \). Table[1] shows the average, standard deviation (std) of the true performance of the optimal policies, and the \( D \) value under different formulations, where the data size \( H \) varies from 10 to 1000. Figure[1] shows that a fair portion of empirical optimal policies (Figure[1d]) perform badly on the true system with values (\( > 30.6 \)) far from the optimal value function \( V^* = 30.05 \), while the optimal policies of BR-MDP formulations (Figure[1a][1c][1e]) have true performance concentrated around \( V^* = 30.05 \). This is more obvious in the VaR and CVaR formulations, which is also reflected by the smaller \( D \) value and smaller standard deviation shown in Table[1]. This is because VaR and CVaR formulations aim to avoid the extreme large costs by choosing the optimal policy under adversarial parameter scenarios, and hence are more robust. Table[1] further shows that as the data size becomes larger, we have better estimation of the parameter value for both the Bayesian risk approach and the empirical approach, and therefore the true performance of optimal policies under all four formulations tend to be smaller and approaches 0.

5.2 Path planning problem

We consider the following path planning problem: an agent wants to exit a \( 3 \times 9 \) maze using the least possible time. Starting from the upper-left corner, the agent can move up, down, left and right, but it
cannot go outside of the maze, and it can only exit the grid at the lower-right corner. Let $s_t$ denote the location of the agent at stage $t$, $a_t$ denote the direction of the agent move at stage $t$, $c_t$ denote the passing time of the agent at stage $t$. As shown in Figure 2, a white cell stands for a normal place where the agent needs one time unit to pass through, and a grey cell represents a “shaky” place where the agent has an uncertain time or probability to pass.

Table 1: Average, standard deviation of the true performance of the optimal policies, and the $D$ value under different formulations. Inventory control problem.

| Data size $H$ | 10   | 20   | 100  | 1000 |
|---------------|------|------|------|------|
| BR-MDP mean   |      |      |      |      |
| average       | 30.37| 30.34| 30.18| 30.05|
| std           | 0.23 | 0.21 | 0.19 | 0.00 |
| $D$ value     | $1.65 \times 10^{-5}$ | $1.45 \times 10^{-5}$ | $0.78 \times 10^{-5}$ | $0.00 \times 10^{-5}$ |
| BR-MDP VaR    |      |      |      |      |
| average       | 30.22| 30.21| 30.10| 30.05|
| std           | 0.13 | 0.13 | 0.12 | 0.00 |
| $D$ value     | $0.52 \times 10^{-5}$ | $0.47 \times 10^{-5}$ | $0.20 \times 10^{-5}$ | $0.00 \times 10^{-5}$ |
| BR-MDP CVaR   |      |      |      |      |
| average       | 30.22| 30.20| 30.10| 30.05|
| std           | 0.13 | 0.12 | 0.12 | 0.00 |
| $D$ value     | $0.50 \times 10^{-5}$ | $0.44 \times 10^{-5}$ | $0.20 \times 10^{-5}$ | $0.00 \times 10^{-5}$ |
| Empirical     |      |      |      |      |
| average       | 30.34| 30.26| 30.13| 30.05|
| std           | 0.25 | 0.25 | 0.19 | 0.00 |
| $D$ value     | $1.68 \times 10^{-5}$ | $1.22 \times 10^{-5}$ | $0.49 \times 10^{-5}$ | $0.00 \times 10^{-5}$ |

Figure 2: The maze for the path planning problem.
We consider both the cases of uncertain cost and uncertain transition. In the uncertain cost case, the agent needs $\xi$ time unit to pass through the "shaky" place, and $\xi$ is a random variable that is no less than 1 with parameter $\theta_c$ that is unknown to the agent. The cost at each stage is the time for the agent to pass through the current cell. If $s_t$ is white cell, the cost at stage $t$ is 1; if $s_t$ is "shaky" cell, the cost at stage $t$ is $\xi_t$. In the uncertain transition case, if an agent reaches a “shaky” place, then the transition becomes unpredictable: in the next step, it will stay at the current cell regardless of the action with probability $1 - \theta_c$, and it will transition to the next cell (according to the action) using one time unit with probability $\theta_c$. Note that in this case, we can view it as an uncertain cost case, where $\xi$ follows a geometric distribution with mean $\frac{1}{\theta_c}$. The goal is to find the optimal policy that minimizes the total time to exit the maze. Same with the problem setting in inventory control problem, we take a Bayesian perspective and view the unknown parameter as a random variable $\theta \in \Theta$ and estimate it using the Bayesian posterior distribution $\mu_t$ at stage $t$, given a prior distribution $\mu_{t-1}$. The posterior update depends on the state: if $s_t$ is white cell, the posterior distribution will not be updated; if $s_t$ is "shaky" cell, the posterior distribution will be updated with $\mu_{t+1}(\theta) = \frac{\mu_t(\theta) f(\xi_t | \theta)}{\int \mu_t(\theta) f(\xi_t | \theta) d\theta}$. The implementation and evaluation of different formulations is similar to the inventory control problem.

5.2.1 Finite parameter space

We first show the numerical simulation on the case of uncertain transition with a finite parameter space. Note that in this case the randomness $\xi$ follows a geometric distribution with mean $\frac{1}{\theta_c}$. The parameters are set as follows: $s_0 = (0, 0)$, exit of the maze locates at $(2, 8)$, $\Theta = \{0.245, 0.25, 0.255\}$, $\theta_c = 1.5$, $J = 100$, $\alpha = 0.6$, $\mu_{-1}$ is the uniform distribution on the support in $\Theta$. Note that in this case, $V^* = 18$, corresponding to the strategy that avoids passing the "shaky" cells. This policy corresponds to the green route in Figure 2.

Figure 3: Histogram of the true performance of optimal policies under different formulations. Input dataset size $H = 20$. Maze problem with a finite parameter space.

Figure 3 shows the histogram of the true performance of the optimal policies under different formulations (BR-MDP mean, BR-MDP VaR, BR-MDP CVaR, and empirical approach), where the data size is $H = 20$. Table 2 shows the average, standard deviation (std) of the true performance of the optimal policies, and the $D$ value under different formulations, where the data size $H$ varies from 10 to 1000. Note that with the BR-MDP approach one can learn from passing the "shaky"
cells and avoid such behavior (passing the "shaky" cells) afterwards. This corresponds to the blue route in Figure 2. Therefore, the BR-MDP approach can have true performance between 18 and 19. In contrast, the empirical approach does not learn from passing the "shaky" cells and ends up with the red route in Figure 2. The conclusion we can draw from Figure 3 and Table 2 is similar to the inventory control problem: with BR-MDP VaR and BR-MDP CVaR formulations, the $D$ value and the standard deviation are smaller, which reflects the robustness of the BR-MDP formulation; the true performance of optimal policies under all four formulations tend to be smaller and approaches 0 as the data size becomes larger.

### 5.2.2 Continuous parameter space

We then consider the case of uncertain cost with a continuous parameter space. We assume the randomness $\xi$ follows a normal distribution that is left truncated at one with unknown mean $\theta$ and

| Data size $H$ | 10  | 20  | 100 | 1000 |
|---------------|-----|-----|-----|------|
| BR-MDP mean   |     |     |     |      |
| average       | 18.49 | 18.37 | 18.17 | 18.00 |
| std           | 0.48 | 0.47 | 0.37 | 0.00 |
| $D$ value     | $1.43 \times 10^{-3}$ | $1.10 \times 10^{-3}$ | $0.50 \times 10^{-3}$ | $0.00 \times 10^{-3}$ |
| BR-MDP VaR    |     |     |     |      |
| average       | 18.33 | 18.26 | 18.11 | 18.00 |
| std           | 0.45 | 0.42 | 0.30 | 0.00 |
| $D$ value     | $0.95 \times 10^{-3}$ | $0.76 \times 10^{-3}$ | $0.32 \times 10^{-3}$ | $0.00 \times 10^{-3}$ |
| BR-MDP CVaR   |     |     |     |      |
| average       | 18.33 | 18.26 | 18.11 | 18.00 |
| std           | 0.45 | 0.42 | 0.30 | 0.00 |
| $D$ value     | $0.95 \times 10^{-3}$ | $0.76 \times 10^{-3}$ | $0.32 \times 10^{-3}$ | $0.00 \times 10^{-3}$ |
| Empirical     |     |     |     |      |
| average       | 18.44 | 18.29 | 18.12 | 18.00 |
| std           | 0.50 | 0.45 | 0.32 | 0.00 |
| $D$ value     | $1.36 \times 10^{-3}$ | $0.90 \times 10^{-3}$ | $0.37 \times 10^{-3}$ | $0.00 \times 10^{-3}$ |

Table 2: Average, standard deviation of the true performance of the optimal policies, and the $D$ value under different formulations. Maze problem with a finite parameter space.

Figure 4: Histogram of the true performance of optimal policies under different formulations. Input dataset size $H = 10$. Maze problem with a continuous parameter space.

We then consider the case of uncertain cost with a continuous parameter space. We assume the randomness $\xi$ follows a normal distribution that is left truncated at one with unknown mean $\theta$ and
known variance \( (\sigma^c)^2 \). We assume the conjugate prior \( \mu_{-1} \) is a normal distribution with mean \( m_0 \) and variance \( \sigma_0^2 \). The other parameters are set as follows: \( \theta^c = 5.5, \Theta = \{ \theta \in \mathbb{R} | \theta \geq 1 \}, \sigma^c = 2, J = 100, \alpha = 0.6, m_0 = 0, \sigma_0^2 = 10^6 \). Note that in this case \( V^* = 18 \).

| BR-MDP mean | BR-MDP VaR | BR-MDP CVaR | Empirical |
|------------|------------|-------------|-----------|
| average    | 18.36      | 18.07       | 18.07     | 18.20     |
| std        | 0.46       | 0.25        | 0.25      | 0.39      |
| \( D \) value | \( 1.43 \times 10^{-3} \) | \( 0.21 \times 10^{-3} \) | \( 0.21 \times 10^{-3} \) | \( 0.60 \times 10^{-3} \) |

Table 3: Average, standard deviation of the true performance of the optimal policies, and the \( D \) value under different formulations. Input dataset size \( H = 10 \). Maze problem with a continuous parameter space.

Similar to the case of finite parameter space, Figure 4 shows the histogram of the true performance of the optimal policies under different formulations (BR-MDP mean, BR-MDP VaR, BR-MDP CVaR, and empirical approach), where the data size is \( H = 10 \). Table 3 shows the average, standard deviation (std) of the true performance of the optimal policies, and the \( D \) value under different formulations, where the data size \( H = 10 \). We can draw the same conclusion as the case of finite parameter space.

6 Conclusion

We propose a Bayesian risk approach to MDPs (BR-MDPs) with parameter uncertainty. For the finite-horizon case, we solve the proposed formulation with dynamic programming, and develop a nested simulation optimization algorithm for general risk functions and a more efficient UCB-based adaptive sampling algorithm for the risk-neutral case. For the infinite-horizon case, we obtain the Bellman equation for BR-MDPs and theoretically show its convergence properties. As demonstrated in our numerical results, the proposed BR-MDPs are able to find more robust policies, however, at a sacrifice of more expensive computation than solving a standard MDP without parameter uncertainty. A future direction would be to develop more computationally efficient algorithms, possibly by exploiting problem structures and leveraging on better value function approximations.

References

Arnab Nilim and Laurent Ghaoui. Robustness in markov decision problems with uncertain transition matrices. In S. Thrun, L. Saul, and B. Schölkopf, editors, Advances in Neural Information Processing Systems, volume 16. MIT Press, 2004.

Garud N Iyengar. Robust dynamic programming. Mathematics of Operations Research, 30(2):257–280, 2005.

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

Shie Mannor and Huan Xu. Data-driven methods for markov decision problems with parameter uncertainty. In Operations Research & Management Science in the Age of Analytics, pages 101–129. INFORMS, 2019.

Huan Xu and Shie Mannor. Distributionally robust markov decision processes. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, Advances in Neural Information Processing Systems, volume 23. Curran Associates, Inc., 2010.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations research, 58(3):595–612, 2010.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. Operations Research, 62(6):1358–1376, 2014.

Di Wu, Helin Zhu, and Enlu Zhou. A bayesian risk approach to data-driven stochastic optimization: Formulations and asymptotics. SIAM Journal on Optimization, 28(2):1588–1612, 2018.

Enlu Zhou and Wei Xie. Simulation optimization when facing input uncertainty. In L. Yilmaz, W. K V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, editors, Proceedings of the 2015 Winter Simulation Conference, pages 3714–3724, Piscataway, New Jersey, 2015. Institute of Electrical and Electronics Engineers, Inc.
Malcolm J. A. Strens. A bayesian framework for reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning, ICML ’00*, page 943–950, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Raphaël Fonteneau, Nathan Korda, and Rémi Munos. An optimistic posterior sampling strategy for bayesian reinforcement learning. In *NIPS 2013 Workshop on Bayesian Optimization (BayesOpt2013)*, 2013.

Aditya Gopalan and Shie Mannor. Thompson Sampling for Learning Parameterized Markov Decision Processes. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 861–898, Paris, France, 2015. PMLR.

Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI’15*, page 2–11, Arlington, Virginia, USA, 2015. AUAI Press.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2701–2710. PMLR, 2017.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Philippe Jorion. *Value at risk: the new benchmark for controlling market risk*. Irwin Professional Pub., 1997.

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

Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

Marek Petrik and Dharmashankar Subramanian. An approximate solution method for large risk-averse markov decision processes. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI’12*, page 805–814, Arlington, Virginia, USA, 2012. AUAI Press. ISBN 9780974903989.

Stefano Carpin, Yin-Lam Chow, and Marco Pavone. Risk aversion in finite markov decision processes using total cost criteria and average value at risk. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 335–342, 2016.

Daniel R Jiang and Warren B Powell. Risk-averse approximate dynamic programming with quantile-based risk measures. *Mathematics of Operations Research*, 43(2):554–579, 2018.

Mohamadreza Ahmadi, Ugo Rosolia, Michel D Ingham, Richard M Murray, and Aaron D Ames. Constrained risk-averse markov decision processes. *arXiv preprint arXiv:2012.02423*, 2020.

Marc Rigter, Bruno Lacerda, and Nick Hawes. Risk-averse bayes-adaptive reinforcement learning. *arXiv preprint arXiv:2102.05762*, 2021.

Andy Philpott, Vitor de Matos, and Erlon Finardi. On solving multistage stochastic programs with coherent risk measures. *Operations Research*, 61(4):957–970, 2013.

Alexander Shapiro. Rectangular sets of probability measures. *Operations Research*, 64(2):528–541, 2016.

Oscar Dowson, David P Morton, and Bernardo K Pagnoncelli. Multistage stochastic programs with the entropic risk measure, 2020.

Alexander Shapiro. Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming. *European Journal of Operational Research*, 288(1):1–13, 2021.

Alois Pichler and Alexander Shapiro. Mathematical foundations of distributionally robust multistage optimization. *arXiv preprint arXiv:2101.02498*, 2021.
Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete bayesian reinforcement learning. In Proceedings of the 23rd International Conference on Machine Learning, ICML ’06, page 697–704, New York, NY, USA, 2006. Association for Computing Machinery.

Erick Delage and Shie Mannor. Percentile optimization for markov decision processes with parameter uncertainty. Operations research, 58(1):203–213, 2010.

Mahdi Imani, Seyed Fatemeh Ghoreishi, and Ulisses M. Braga-Neto. Bayesian control of large mdps with unknown dynamics in data-poor environments. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018.

Marek Petrik and Reazul Hasan Russel. Beyond confidence regions: Tight bayesian ambiguity sets for robust mdps. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

Esther Derman, Daniel Mankowitz, Timothy Mann, and Shie Mannor. A bayesian approach to robust reinforcement learning. In Uncertainty in Artificial Intelligence, pages 648–658. PMLR, 2020.

Guillermo Gallego and Garrett Van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. Management science, 40(8):999–1020, 1994.

Dimitri P Bertsekas et al. Dynamic programming and optimal control: Vol. 1. Athena scientific Belmont, 2000.

Robert Schlaifer and Howard Raiffa. Applied statistical decision theory. Division of Research, Harvard Business School, Boston, Massachusetts, 1961.

Helin Zhu, Tianyi Liu, and Enlu Zhou. Risk quantification in stochastic simulation under input uncertainty. ACM Transactions on Modeling and Computer Simulation (TOMACS), 30(1):1–24, 2020.

Hyeong Soo Chang, Michael C. Fu, Jiqiaio Hu, and Steven I. Marcus. An adaptive sampling algorithm for solving markov decision processes. Operations Research, 53(1):126–139, 2005.

Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. IEEE Journal of Selected Topics in Signal Processing, 10(6):1093–1111, 2016.

Nicolas Galichet, Michèle Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In Cheng Soon Ong and Tu Bao Ho, editors, Proceedings of the 5th Asian Conference on Machine Learning, volume 29 of Proceedings of Machine Learning Research, pages 245–260. PMLR, 2013.

Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In Sanjay Jain, Rémi Munos, Frank Stephan, and Thomas Zeugmann, editors, Algorithmic Learning Theory, pages 218–233, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, Proceedings of the 31st Conference On Learning Theory, volume 75 of Proceedings of Machine Learning Research, pages 1295–1306. PMLR, 2018.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2):235–256, 2002.