Supplementary Materials for Bayesian Risk Markov Decision Processes

A Nomenclature

- s: state in an MDP. $s \in \mathcal{S}$.
- a: action in an MDP. $a \in \mathcal{A}$.
- $\mathcal{P}$: transition probability in an MDP.
- $\mathcal{C}$: cost function in an MDP.
- T: time horizon in an MDP.
- ξ: the randomness in the system. $\xi \in \Xi \subseteq \mathbb{R}^k$. $\xi \sim f(\cdot; \theta^c)$.
- θ^c: the unknown parameter in the true distribution of ξ. $\theta^c \in \Theta \subseteq \mathbb{R}^d$.
- μ: the posterior distribution on θ. $\mu \in \mathcal{M}$.
- ρ_μ: the risk functional taken with respect to $\theta \sim \mu$.
- $g_t(s_t, a_t, \xi_t)$: state transition function at time stage t. $s_{t+1} = g_t(s_t, a_t, \xi_t)$.
- $V_t^*(s_t, \mu_t)$: the optimal value function at time stage t.
- $\pi_t^*(s_t, \mu_t)$: the optimal deterministic Markovian policy.
- α: risk level in CVaR risk functional.
- u: additional variable to optimize in CVaR minimization representation.
- $\alpha(s, \theta)$: α-function. $V_t^*(s_t, \mu_t) = \min_{\alpha_t \in \Gamma_t} \int_\Theta \alpha_t(s_t, \theta)\mu_t(\theta)d\theta$.
- $Q_t^*(s_t, \mu_t, a_t, u_t)$: optimal Q function. $V_t^*(s_t, \mu_t) = \min_{a_t \in \mathcal{A}, u_t \in \mathbb{R}} Q_t^*(s_t, \mu_t, a_t, u_t)$.
- $V_t(s_t, \mu_t) := \min_{a_t} Q_t^*(s_t, \mu_t, a_t, u_t)$: "optimal" value function for a given u_t.
- $\underline{V}_t(s_t, \mu_t)$: lower bound for $V_t(s_t, \mu_t)$.
- $\underline{\alpha}_t \in \underline{\Gamma}_t$: lower bound for α_t. $\underline{V}_t(s_t, \mu_t) := \min_{\underline{\alpha}_t \in \underline{\Gamma}_t} \int_\Theta \underline{\alpha}_t(s_t, \theta)\mu_t(\theta)d\theta$.
- $\bar{V}_t(s_t, \mu_t)$: upper bound for $V_t(s_t, \mu_t)$.
- $\bar{\alpha}_t \in \bar{\Gamma}_t$: upper bound for α_t. $\bar{V}_t(s_t, \mu_t) := \min_{\bar{\alpha}_t \in \bar{\Gamma}_t} \int_\Theta \bar{\alpha}_t(s_t, \theta)\mu_t(\theta)d\theta$.
- $\widetilde{V}_t(s_t, \mu_t)$: approximate for $V_t(s_t, \mu_t)$.
- $\widetilde{\alpha}_t \in \widetilde{\Gamma}_t$: approximate for α_t. $\widetilde{V}_t(s_t, \mu_t) := \min_{\widetilde{\alpha}_t \in \widetilde{\Gamma}_t} \int_\Theta \widetilde{\alpha}_t(s_t, \theta)\mu_t(\theta)d\theta$.

B Proof details

B.1 Proof of Proposition 4.1

Proof. We prove by induction. For $t = T$, we have $V_T^*(s_T, \mu_T) = \mathcal{C}_T(s_T)$. For $t = T-1$, let

$$Q_{T-1}^*(s_{T-1}, \mu_{T-1}, a_{T-1}, u_{T-1})$$
$$= \int_\Theta \underbrace{\left\{ u_{T-1} + \frac{1}{1-\alpha} \left(\int_\Xi f(\xi; \theta) \left(\mathcal{C}_{T-1}(s_{T-1}, a_{T-1}, \xi) + \mathcal{C}_T(s_T) \right) d\xi - u_{T-1} \right)^+ \right\}}_{\alpha_{T-1}(s_{T-1}, \theta | a_{T-1}, u_{T-1})} \mu_{T-1}(\theta) d\theta.$$

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Then $V^*_{T-1}(s_{T-1},\mu_{T-1}) = \min_{a_{T-1}\in\mathcal{A},u_{T-1}\in\mathbb{R}} Q^*_{T-1}(s_{T-1},\mu_{T-1},a_{T-1},u_{T-1})$ takes the desired form. For $t \le T-2$, assuming $V^*_{t+1}(s_{t+1},\mu_{t+1})$ takes the desired form, then by induction we have

$$
\begin{aligned}
&Q^*_t(s_t,\mu_t,a_t,u_t)\\
&= \int_\Theta \left\{ u_t + \frac{1}{1-\alpha}\left(\int_\Xi f(\xi;\theta)\left(\mathcal{C}(s_t,a_t,\xi) + V^*_{t+1}(s_{t+1},\mu_t)\right)d\xi - u_t\right)^+ \right\}\mu_t(\theta)d\theta\\
&= \underbrace{\int_\Theta \left\{ u_t + \frac{1}{1-\alpha}\left(\int_\Xi f(\xi;\theta)\left(\mathcal{C}_t(s_t,a_t,\xi) + \min_{\alpha_{t+1}}\int_\Theta \alpha_{t+1}(s_{t+1},\theta)\frac{\mu_t(\theta)f(\xi;\theta)}{\int_\Theta \mu_t(\theta)f(\xi;\theta)}d\theta\right)d\xi - u_t\right)^+ \right\}}_{\alpha_t(s_t,\theta|a_t,u_t)}\mu_t(\theta)d\theta.
\end{aligned}
$$

Then $V^*_t(s_t,\mu_t) = \min_{a_t\in\mathcal{A},u_t\in\mathbb{R}} Q^*_t(s_t,\mu_t,a_t,u_t)$ takes the desired form. □

B.2 Proof of Proposition 4.2

Proof. For the lower bound, we have for $t \le T-1$,

$$
\begin{aligned}
&Q^*_t(s_t,\mu_t,a_t,u_t)\\
&= u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi f(\xi;\theta)\left(\mathcal{C}_t(s_t,a_t,\xi) + \min_{\alpha_{t+1}}\int_\Theta \alpha_{t+1}(s_{t+1},\theta)\frac{\mu_t(\theta)f(\xi;\theta)}{\int_\Theta \mu_t(\theta)f(\xi;\theta)}d\theta\right)d\xi - u_t\right)^+ \mu_t(\theta)d\theta\\
&\ge u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi f(\xi;\theta)\left(\mathcal{C}(s_t,a_t,\xi) + \min_{\alpha_{t+1}}\sum_\Theta \alpha_{t+1}(s_{t+1},\theta)\frac{\mu_t(\theta)f(\xi;\theta)}{\int_\Theta \mu_t(\theta)f(\xi;\theta)}d\theta\right)d\xi - u_t\right)\mu_t(\theta)d\theta\\
&= u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi f(\xi;\theta)\left(\mathcal{C}(s_t,a_t,\xi) - u_t\right)d\xi\right)\mu_t(\theta)d\theta + \frac{1}{1-\alpha}\int_\Xi \left(\min_{\alpha_{t+1}}\int_\Theta \alpha_{t+1}(s_{t+1},\theta)f(\xi;\theta)\mu_t(\theta)d\theta\right)d\xi\\
&\ge u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi f(\xi;\theta)\left(\mathcal{C}_t(s_t,a_t,\xi) - u_t\right)d\xi\right)\mu_t(\theta)d\theta + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi \min_{\alpha_{t+1}} \alpha_{t+1}(s_{t+1},\theta)f(\xi;\theta)d\xi\right)\mu_t(\theta)d\theta\\
&:= \underline{Q}_t(s_t,\mu_t,a_t,u_t)
\end{aligned}
$$

where the last inequality is justified by Jensen's inequality as we exchange $\min$ and summation over θ. Therefore, $\underline{V}_t(s_t,\mu_t) := \min_{a_t\in\mathcal{A}} \underline{Q}_t(s_t,\mu_t,a_t,u_t) \le V_t(s_t,\mu_t)$. For the upper bound, we have for $t \le T-1$,

$$
\begin{aligned}
&Q_t(s_t,\mu_t,a_t,u_t)\\
&= u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi f(\xi;\theta)\left(\mathcal{C}_t(s_t,a_t,\xi) + \min_{\alpha_{t+1}}\int_\Theta \alpha_{t+1}(s_{t+1},\theta)\frac{\mu_t(\theta)f(\xi;\theta)}{\int_\Theta \mu_t(\theta)f(\xi;\theta)}d\theta\right)d\xi - u_t\right)^+ \mu_t(\theta)d\theta\\
&\le u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi f(\xi;\theta)\left(\mathcal{C}_t(s_t,a_t,\xi) + \alpha^*_{t+1}(s_{t+1},\theta) - u_t\right)d\xi\right)^+ \mu_t(\theta)d\theta\\
&\le u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi f(\xi;\theta)\left(\mathcal{C}_t(s_t,a_t,\xi) - u_t\right)d\xi\right)^+ \mu_t(\theta)d\theta + \frac{1}{1-\alpha}\int_\Xi \left(\min_{\alpha_{t+1}}\int_\Theta \alpha_{t+1}(s_{t+1},\theta)f(\xi;\theta)\mu_t(\theta)d\theta\right)d\xi\\
&\le u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi f(\xi;\theta)\left(\mathcal{C}_t(s_t,a_t,\xi) - u_t\right)d\xi\right)^+ \mu_t(\theta)d\theta + \frac{1}{1-\alpha}\min_{\alpha_{t+1}}\int_\Theta \left(\int_\Xi \alpha_{t+1}(s_{t+1},\theta)f(\xi;\theta)d\xi\right)\mu_t(\theta)d\theta\\
&:= \bar{Q}_t(s_t,\mu_t,a_t,u_t)
\end{aligned}
$$

where $\alpha^*_{t+1}(s_{t+1},\theta)$ attains the minimum of $\int_\Theta \alpha_{t+1}(s_{t+1},\theta)f(\xi;\theta)\mu_t(\theta)d\theta$. The last inequality is justified by Jensen's inequality as we exchange $\min$ and integral over ξ. Therefore, $\bar{V}_t(s_t,\mu_t) := \min_{a_t\in\mathcal{A}} \bar{Q}_t(s_t,\mu_t,a_t,u_t) \ge V_t(s_t,\mu_t)$. In the following, we derive another approximate value function $\widetilde{V}_t$. We start from the lower bound. By applying Jensen's inequality and exchanging $\min$

and integral over ξ, we have

$$
\begin{aligned}
&\underline{Q}_t(s_t, \mu_t, a_t, u_t) \\
&= u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi \left(\mathcal{C}_t(s_t, a_t, \xi) - u_t + \min_{\alpha_{t+1}} \alpha_{t+1}(s_{t+1}, \theta)\right) f(\xi;\theta)d\xi\right)\mu_t(\theta)d\theta \\
&\le u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi \mathcal{C}_t(s_t, a_t, \xi) f(\xi;\theta)d\xi - u_t + \min_{\alpha_{t+1}} \int_\Xi \alpha_{t+1}(s_{t+1}, \theta) f(\xi|\theta)d\xi\right)\mu_t(\theta)d\theta \\
&\le u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi \mathcal{C}_t(s_t, a_t, \xi) f(\xi;\theta)d\xi - u_t + \min_{\alpha_{t+1}} \int_\Xi \alpha_{t+1}(s_{t+1}, \theta) f(\xi;\theta)d\xi\right)^+\mu_t(\theta)d\theta \\
&:= \widetilde{Q}_t(s_t, \mu_t, a_t, u_t) \\
&\le u_t + \frac{1}{1-\alpha}\int_\Theta \left(\int_\Xi \mathcal{C}_t(s_t, a_t, \xi) f(\xi;\theta)d\xi - u_t\right)^+\mu_t(\theta)d\theta + \frac{1}{1-\alpha}\int_{\theta\in\Theta}\left(\min_{\alpha_{t+1}} \int_\Xi \alpha_{t+1}(s_{t+1}, \theta) f(\xi;\theta)d\xi\right)\mu_t(\theta)d\theta \\
&= \bar{Q}_t(s_t, \mu_t, a_t, u_t)
\end{aligned}
$$

Therefore, $\underline{V}_t(s_t, \mu_t) \le \widetilde{V}_t(s_t, \mu_t) := \min_{a_t\in\mathcal{A}} \widetilde{Q}_t(s_t, \mu_t, a_t, u_t) \le \bar{V}_t(s_t, \mu_t)$. □

B.3 Proof of Theorem 4.3

Proof. We prove by induction. For $t = T-1$, we have

$$\widetilde{V}_{T-1}(s_{T-1}, \mu_{T-1}, u_{T-1}) = \min_{a_{T-1}}\left\{u_{T-1} + \frac{1}{1-\alpha}\int_\Theta\left(\int_\Xi \left(\mathcal{C}_{T-1}(s_{T-1}, a_{T-1}, \xi) + \mathcal{C}_T(s_T)\right) f(\xi;\theta)d\xi - u_{T-1}\right)^+\mu_{T-1}(\theta)d\theta\right\}.$$

Since $\mathcal{C}_{T-1}(s_{T-1}, a_{T-1}, \xi)$ is jointly convex in s_{T-1} and a_{T-1}, and state transition $g(s_{T-1}, a_{T-1}, \xi)$ is jointly convex in s_{T-1} and a_{T-1}, we have $\mathcal{C}_{T-1}(s_{T-1}, a_{T-1}, \xi) + \mathcal{C}_T(s_T)$ is convex in a_{T-1}, and it follows that $\widetilde{\alpha}_{T-1}(s_{T-1}, a_{T-1}, \theta)$ is jointly convex in a_{T-1} and u_{T-1}. Thus $\widetilde{V}_{T-1}(s_{T-1}, \mu_{T-1}, u_{T-1})$ is convex in u_{T-1}. Suppose now it holds for some $t \le T-2$, i.e., $\alpha_{t+1}(s_{t+1}, a_{t+1}, \theta)$ is jointly convex in $(u_{t+1}, \cdots, u_{T-1})$ and a_{t+1}. Note that

$$\widetilde{V}_t(s_t, \mu_t, u_{t:}) = \min_{a_t\in\mathcal{A}}\left\{u_t + \frac{1}{1-\alpha}\int_\Theta\left(\min_{a_{t+1}}\int_\Xi \left(\mathcal{C}_t(s_t, a_t, \xi) + \widetilde{\alpha}_{t+1}(s_{t+1}, a_{t+1}, \theta)\right) f(\xi;\theta)d\xi - u_t\right)^+\mu_t(\theta)d\theta\right\}.$$

By induction, $\int_\Xi \left(\mathcal{C}_t(s_t, a_t, \xi) + \widetilde{\alpha}_{t+1}(s_{t+1}, a_{t+1}, \theta)\right) f(\xi;\theta)d\xi$ is jointly convex in $(u_{t+1}, \cdots, u_{T-1})$ and a_{t+1}. Also from the convex assumption on the state transition, we have the joint convexity in a_t and $(u_{t+1}, \cdots, u_{T-1})$ of the term inside $(\cdot)^+$ operator. Therefore, the convexity of $\widetilde{V}_t(s_t, \mu_t, u_{t:})$ w.r.t. $u_{t:}$ holds. □

B.4 Proof of Theorem 4.4

Proof. For $t = T-1$, clearly we have

$$\min_{u_{T-1}} \widetilde{V}_{T-1}(s_{T-1}, \mu_{T-1}, u_{T-1}) = V^*_{T-1}(s_{T-1}, \mu_{T-1}).$$

For $t = T-2$, we have

$$
\begin{aligned}
&V^*_{T-2}(s_{T-2}, \mu_{T-2}) \\
&= \min_{a_{T-2}, u_{T-2}} u_{T-2} + \frac{1}{1-\alpha}\int_\Theta\left(\int_\Xi\left(\mathcal{C}_{T-2}(s_{T-2}, a_{T-2}, \xi) + \min_{u_{T-1}}\widetilde{V}_{T-1}(s_{T-1}, \mu_{T-1}, u_{T-1})\right) f(\xi;\theta)d\xi - u_{T-2}\right)^+\mu_{T-2}(\theta)d\theta \\
&\le \min_{u_{T-2}, u_{T-1}}\min_{a_{T-2}} u_{T-2} + \frac{1}{1-\alpha}\int_\Theta\left(\int_\Xi\left(\mathcal{C}_{T-2}(s_{T-2}, a_{T-2}, \xi) + \widetilde{V}_{T-1}(s_{T-1}, \mu_{T-1}, u_{T-1})\right) f(\xi;\theta)d\xi - u_{T-2}\right)^+\mu_{T-2}(\theta)d\theta \\
&= \min_{u_{T-2}, u_{T-1}}\min_{a_{T-2}} u_{T-2} + \frac{1}{1-\alpha}\int_\Theta\Big(\int_\Xi(\mathcal{C}_{T-2}(s_{T-2}, a_{T-2}, \xi) + \min_{a_{T-1}} u_{T-1} \\
&\quad + \frac{1}{1-\alpha}\int_\Theta\Big(\int_\Xi(\mathcal{C}_{T-1}(s_{T-1}, a_{T-1}, \xi) + \mathcal{C}_T(s_T))f(\xi;\theta)d\xi - u_{T-1})^+\mu_{T-1}(\theta)d\theta\Big)f(\xi;\theta)d\xi - u_{T-2})^+\mu_{T-2}(\theta)d\theta \\
&\le \min_{u_{T-2}, u_{T-1}}\min_{a_{T-2}} u_{T-2} + \frac{1}{1-\alpha}\int_\Theta\Big(\min_{a_{T-1}}\int_\Xi(\mathcal{C}_{T-2}(s_{T-2}, a_{T-2}, \xi) + u_{T-1} \\
&\quad + \frac{1}{1-\alpha}\int_\Theta\Big(\int_\Xi(\mathcal{C}_{T-1}(s_{T-1}, a_{T-1}, \xi) + \mathcal{C}_T(s_T))f(\xi;\theta)d\xi - u_{T-1})^+\mu_{T-1}(\theta)d\theta\Big)f(\xi;\theta)d\xi - u_{T-2})^+\mu_{T-2}(\theta)d\theta \\
&\le \widetilde{V}_{T-2}(s_{T-2}, \mu_{T-2}, u_{T-2}, u_{T-1}).
\end{aligned}
$$

Repeating the above process for $t = T-3, \cdots, 0$, we have $V_t^*(s_t, \mu_t) \leq \min_{u_t, \cdots, u_{T-1}} \widetilde{V}_t(s_t, \mu_t)$. □

C Implementing details

Code in Python for the numerical experiments is included in the supplementary. Computational time is reported for a 1.4 GHz Intel Core i5 processor with 8 GB memory.

C.1 Parameter setup

In the gambler's betting problem, the initial wealth $s_0 = 60$, and the parameter space is set to $\Theta = \{0.1, 0.3, 0.45, 0.55, 0.7, 0.9\}$. In CVaR BR-MDP with exact dynamic programming, to obtain the "exact" (more precisely, should be close-to-exact) optimal value function, we discretize the continuous state, i.e., the posterior distribution, with small step size 0.1, which results in very large state space, and then we conduct dynamic programming on the discretized problem to obtain the optimal value function. This is a brute-force way to compute the "exact" value function, and that's why the computational time for the exact BR-MDP formulation is extremely large compared to the approximate formulation. In CVaR BR-MDP with approximate dynamic programming, the initial u-vector $u^0 = (60, 50, 40, 30, 20, 10)$. The gradient descent is run for $K = 100$ iterations. The learning rate is set to $\eta_k = \frac{100}{1+k}$. In the inventory control problem, the initial inventory level $s_0 = 5$, the parameter space is set to $\Theta = \{4, 6, 8, 10, 12, 14, 16\}$. The storage capacity is set to $M = 15$. Maximal customer demand is set to $M_C = 20$. The holding cost is set to $h_t = 4$, and the penalty cost is set to $p_t = 6$. In CVaR BR-MDP with exact dynamic programming, the posterior distribution space $\mathcal{M}$ is a probability simplex with support over Θ and is discretized with gap 0.1. In CVaR BR-MDP with approximate dynamic programming, the initial u-vector $u^0 = (10, 10, 10, 10, 10, 10)$. The gradient descent is run for $K = 100$ iterations. The learning rate is set to $\eta_k = \frac{10}{1+k}$. In both problems, the prior is set to uniform distribution with support over Θ. Given the historical data, the posterior is then updated by Bayes' rule. The resulted posterior then serves as the prior input for Algorithm 1, Algorithm 2, nominal approach and DR-MDP approach.

C.2 DR-MDP details

In the DR-MDP approach, as we have argued before, the construction of the ambiguity set requires aprior knowledge of the probabilistic information, which is not readily available from a given data set. However, we note that BRO has a distributionally robust optimization (DRO) interpretation. In particular, for a static stochastic optimization problem, it is shown in [1] that the BRO formulation with the risk functional taken as VaR with confidence level $\alpha = 100\%$ is equivalent to a DRO formulation with the ambiguity set constructed for θ. Therefore, we adapt DR-MDP to our considered problem as follows: we draw samples of θ from the posterior distribution computed for a given data set, and obtain the optimal policy that minimizes the total expected cost under the most adversarial θ.

C.3 Gradient descent details

In Algorithm 2, to accelerate the gradient computation and convergence of the algorithm, we can instead use stochastic gradient descent. Let $\hat{\xi}_0, \hat{\xi}_1, \cdots, \hat{\xi}_{t-1}$ be a trajectory up to time $t-1$. Let the subsequent states and actions along this trajectory be $\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_t$ and $\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_t$ respectively. Note that

$$
\begin{aligned}
&\frac{\partial \widetilde{\alpha}_0(\hat{s}_0, \hat{a}_0, \theta)}{\partial u_t}(\hat{\xi}_0, \hat{\xi}_1, \cdots, \hat{\xi}_{t-1}) \\
&= \frac{1}{1-\alpha} \mathbb{1}\left\{ \int_\Xi \mathcal{C}_0(\hat{s}_0, \hat{a}_0, \xi_0) f(\xi_0; \theta) d\xi_0 - u_0 + \min_{\hat{a}_1} \int_\Xi \widetilde{\alpha}_1(\hat{s}_1, \hat{a}_1, \theta) f(\xi_0; \theta) d\xi_0 \geq 0 \right\} \\
&\cdot \frac{1}{1-\alpha} \mathbb{1}\left\{ \int_\Xi \mathcal{C}_1(\hat{s}_1, \hat{a}_1, \xi_1) f(\xi_1; \theta) d\xi_1 - u_1 + \min_{\hat{a}_2} \int_\Xi \widetilde{\alpha}_2(\hat{s}_2, \hat{a}_2, \theta) f(\xi_1; \theta) d\xi_1 \geq 0 \right\} \\
&\cdots
\end{aligned}
$$

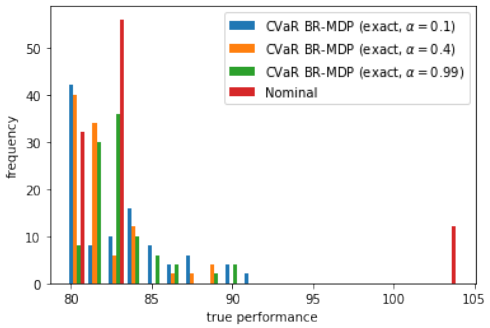
$$\cdot \frac{1}{1-\alpha}\mathbb{1}\left\{\int_{\Xi}\mathcal{C}_{t-1}(\hat{s}_{t-1},\hat{a}_{t-1},\xi_{t-1})f(\xi_{t-1};\theta)d\xi_{t-1}-u_{t-1}+\min_{\hat{a}_t}\int_{\Xi}\widetilde{\alpha}_t(\hat{s}_t,\hat{a}_t,\theta)f(\xi_{t-1};\theta)d\xi_{t-1}\geq 0\right\}$$
$$\cdot\left(1-\frac{1}{1-\alpha}\mathbb{1}\left\{\int_{\Xi}\mathcal{C}_t(\hat{s}_t,\hat{a}_t,\xi_t)f(\xi_t;\theta)d\xi_t-u_t+\min_{\hat{a}_{t+1}}\int_{\Xi}\widetilde{\alpha}_{t+1}(\hat{s}_{t+1},\hat{a}_{t+1},\theta)f(\xi_t;\theta)d\xi_t\geq 0\right\}\right).$$

Since $\frac{\partial\widetilde{\alpha}_0(s_0,a_0,\theta)}{\partial u_t}=\mathbb{E}\left[\frac{\partial\widetilde{\alpha}_0(\hat{s}_0,\hat{a}_0,\theta)}{\partial u_t}(\hat{\xi}_0,\hat{\xi}_1,\cdots,\hat{\xi}_{t-1})\right]$, $(\frac{\partial\widetilde{\alpha}_0(s_0,a_0,\theta)}{\partial u_0},\cdots,\frac{\partial\widetilde{\alpha}_0(s_0,a_0,\theta)}{\partial u_{T-1}})$ can be substituted by an unbiased gradient estimator $(\frac{\partial\widetilde{\alpha}_0(s_0,a_0,\theta)}{\partial u_0},\cdots,\frac{\partial\widetilde{\alpha}_0(\hat{s}_0,\hat{a}_0,\theta)}{\partial u_{T-1}}(\hat{\xi}_0,\hat{\xi}_1,\cdots,\hat{\xi}_{T-2}))$.

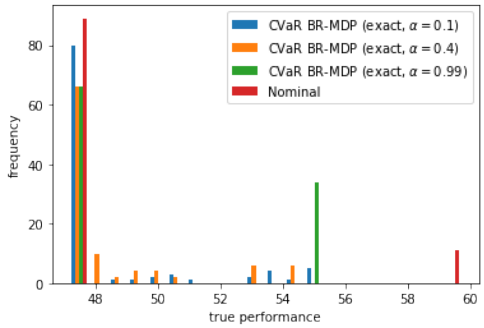
C.4 Relative gaps between $\widetilde{V}_0^*(s_0,\mu_0)$ and $V_0^*(s_0,\mu_0)$

Table 1: Relative gaps between $\widetilde{V}_0^*(s_0,\mu_0)$ and $V_0^*(s_0,\mu_0)$. Betting problem. $N=10$. $\theta^c=0.45$.

prior distribution	$\mu_0(1)$	$\mu_0(2)$	$\mu_0(3)$	$\mu_0(4)$	$\mu_0(5)$	$\mu_0(6)$	$\mu_0(7)$	$\mu_0(8)$	$\mu_0(9)$	$\mu_0(10)$	$\mu_0(11)$
relative gap (%)	32.98%	29.09%	25.29%	18.78%	16.79%	12.34%	9.37%	3.38%	0.10%	0.08%	0.00%



(a) Histogram of actual performance over 100 replications for CVaR BR-MDP (exact) with different α. $\theta^c=12$.



(b) Histogram of actual performance over 100 replications for CVaR BR-MDP (exact) with different α.$\theta^c=4$.

Figure 1: Inventory control problem.

C.5 Inventory control details

We report additional results for the inventory control problem in Figure 1. Figure 1 shows the histogram of actual performance over 100 replications for the nominal approach and CVaR BR-MDP (exact) with different confidence levels $\alpha=0.1,0.5,0.99$ under distributional parameter $\theta^c=4$ and $\theta^c=12$ respectively. The same conclusions can be drawn as the gambler's betting problem.

References

[1] 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.