Bayesian Optimistic Optimization: Optimistic Exploration for Model-based Reinforcement Learning

Chenyang Wu[1], Tianci Li[1], Zongzhang Zhang[1], Yang Yu[1,2]

[1]National Key Lab for Novel Software Technology, Nanjing University, China [2]Pazhou Lab, Guangzhou, China

Learning And Mining from DatA
http://www.lamda.nju.edu.cn

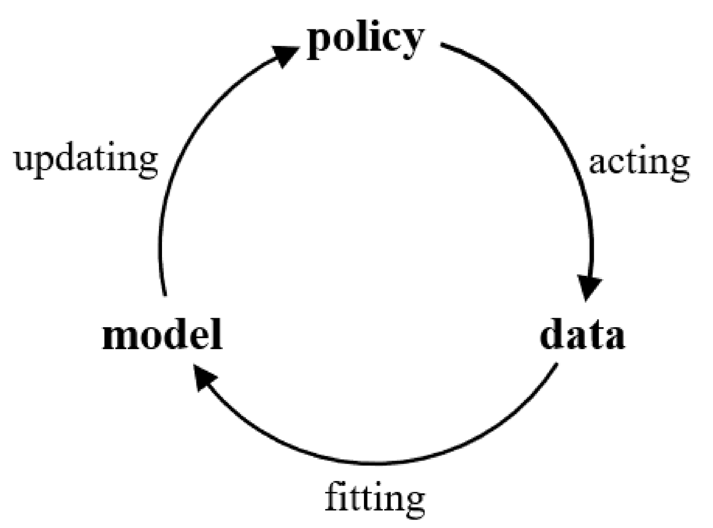
Backgrounds

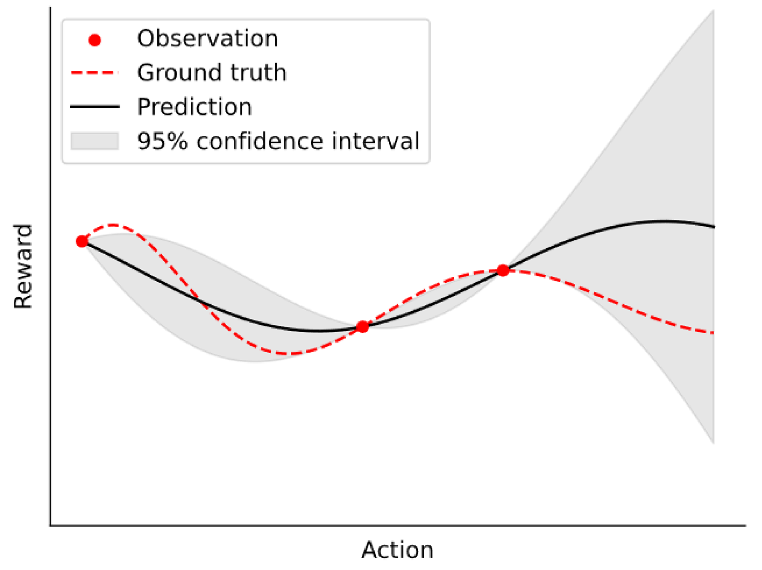
➢ Exploitation and exploration

Model-based RL collects data from the true environment and learns to get a model about the system dynamics, leading to the dilemma of exploitation and exploration:

- ➢ **Exploring uncertain states and actions** to improve future performance.
- ➢ **Exploiting existing knowledge** to improve short-run performance.

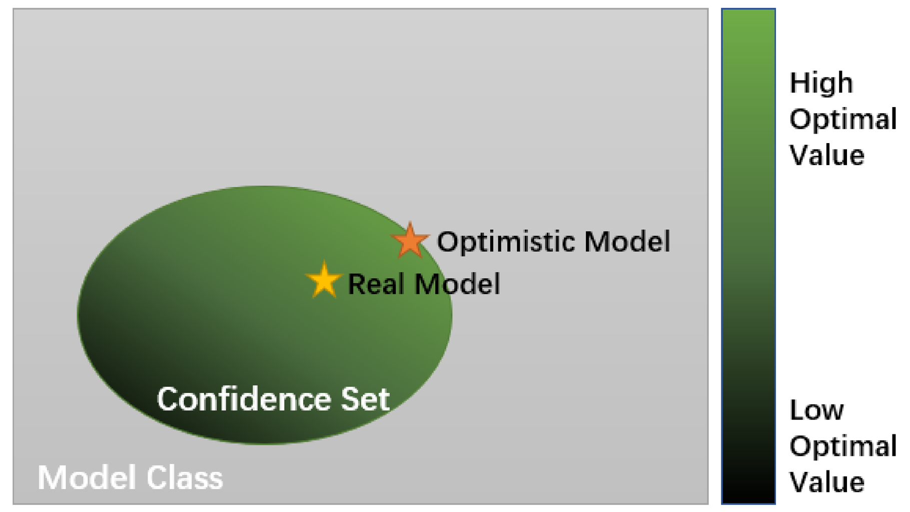
The performance of MBRL algorithm hinges on how we **face uncertainty** and how we **balance exploration and exploitation**.





➢ Optimism in the face of uncertainty (OFU)

- OFU models the environment as optimistically as statistically plausible and involves constructing a confidence set of possible MDPs and solving for the most optimistic one within the confidence set : $\max_{\pi_k, M_k \in \mathcal{M}_k} V_1^k(s_1)$



Notions

- $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively.
- $\mathcal{H}_k$ is the history prior to k-th episode, and s_1 is the initial state.
- M is the MDP model, and V_h is value function at the h-th period of an episode. Here, $h \in [H]$ and H is the length of episode.
- $V_1^{\pi,M}(s_1)$ represents the value of the initial state s_1 under the policy π and the model M at the initial period $h=1$.

Bayesian Optimistic Optimization

- Following the idea of OFU, we select $\mathcal{M}_k$ to be **the highest density region** (HDR) defined as follows:

$$\Pr(\mathcal{M}_k|\mathcal{H}_k) \geq 1-\alpha_k, \qquad \mathcal{M}_k = \{M_k | \Pr(M_k|\mathcal{H}_k) \geq \epsilon_k\},$$

- We transform this problem into an **unconstrained optimization** problem via Lagrangian Relaxation:

$$\max_{\pi, M} (V_1^{\pi,M}(s_1) + \lambda_k(\log \Pr(M|\mathcal{H}_k) - \log \epsilon_k)),$$

- Once the Lagrange multiplier is determined, the optimization is equivalent to:

$$\max_{\pi_k, M_k} (V_1^{\pi_k, M_k}(s_1) + \lambda_k \ln \Pr(M_k|\mathcal{H}_k)),$$

- We show that any properly set λ_k assures convergence to the optimal policy. The optimal $\lambda_k^* = ck^{-v_1^*}(\log k)^{-v_2^*}$ is determined by the **exploration complexity and the size of the model class** (see our paper for details).
- The λ_k^* can be interpreted as $\lambda_k^* = \xi_k^V / \xi_k^M$, where ξ_k^V represents the **value uncertainty** and ξ_k^M stands for **the variation of the log-posterior density** for the model.

BOO via Posterior Sampling (BPS)

- Get d random samples from posterior and solve :

$$\max_\pi \max_{i=1}^d (V_1^{\pi, M_k^i}(s_1) + \lambda_k \ln \Pr(M_k^i|\mathcal{H}_k)),$$

where $\lambda_k^* = \xi_k^V$ since ξ_k^M **is a constant** for posterior samples.

BOO via Gradient Ascent(FiniteBOO)

➢ Value model gradient

Theorem 5.1 (Value model gradient). *Suppose that the transition function P^{M_θ} and reward function R^{M_θ} of model M_θ, the gradient of the value $V_1^{\pi,M_\theta}(s_1)$ w.r.t. the model is*

$$\nabla_\theta V_1^{\pi,M_\theta}(s_1) = \mathbb{E}_{\tau\sim\pi,M_\theta}\left[\sum_{h=1}^{H} \nabla_\theta \bar{R}^{M_\theta}(s_h,a_h) + \sum_{h=1}^{H-1} V_{h+1}^{\pi,M_\theta}(s_{h+1}) \nabla_\theta \log P^{M_\theta}(s_{h+1}|s_h,a_h)\right], \tag{4}$$

where $\tau = (s_1, a_1, \ldots, s_H, a_H)$ is a trajectory, $\tau \sim \pi, M_\theta$ means that the trajectory is formed by the interaction of the policy π and the model M_θ, and $P^{M_\theta}(s_{h+1}|s_h,a_h)$ is the probability of s_{h+1} under distribution $P^{M_\theta}(s_h,a_h)$.

➢ Entropy regularization

- **Use entropy regularization to smooth the BOO objective:**

$$\max_{\pi,M}\left(V_1^{\pi,M}(s1) + \xi_k^V \mathbb{E}_{\tau\sim\pi,M}\left[\sum_{h=1}^{H} \mathrm{KL}\left(\pi_h(s_h)||\hat{\pi}_h(s_h)\right)\right] + \lambda_k \log \Pr(M|\mathcal{H}_k)\right),$$

where ξ_k^V downscales the entropy term in proportion to the value uncertainty, $\hat{\pi}_h$ is a prior policy ensuring π_h is absolutely continuous w.r.t. $\hat{\pi}_h$, and ζ controls the amount of entropy.

➢ Mean reward bonus

- The model becomes optimistic on state-action pairs it visits frequently, which in turn makes these state-action pairs more appealing. This mutual strengthening phenomenon makes optimization easily stuck at local optima and can be solved by **adding a bonus term to the BOO objective:**

$$\xi_k^V H \mathbb{E}_{(s,a)\sim U_{\mathcal{S}\times\mathcal{A}}}[R(s,a)]$$

where $U_{\mathcal{S}\times\mathcal{A}}$ is the uniform distribution over the state-action space, and the coefficient ξ_k^V ensures that the bonus decays with the value uncertainty.

Empirical Study

➢ Empirical study on several standard benchmarks:

Environment	$(\lvert\mathcal{S}\rvert, \lvert\mathcal{A}\rvert, H)$	Uncertainty
River Swim	(5,2,5)	High **transition** uncertainty
Chain	(40,2,40)	High **reward** uncertainty
Random MDPS	(5,5,5)	High **reward and transition** uncertainty

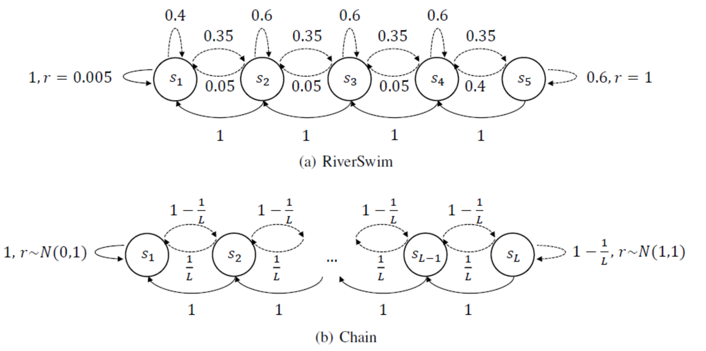
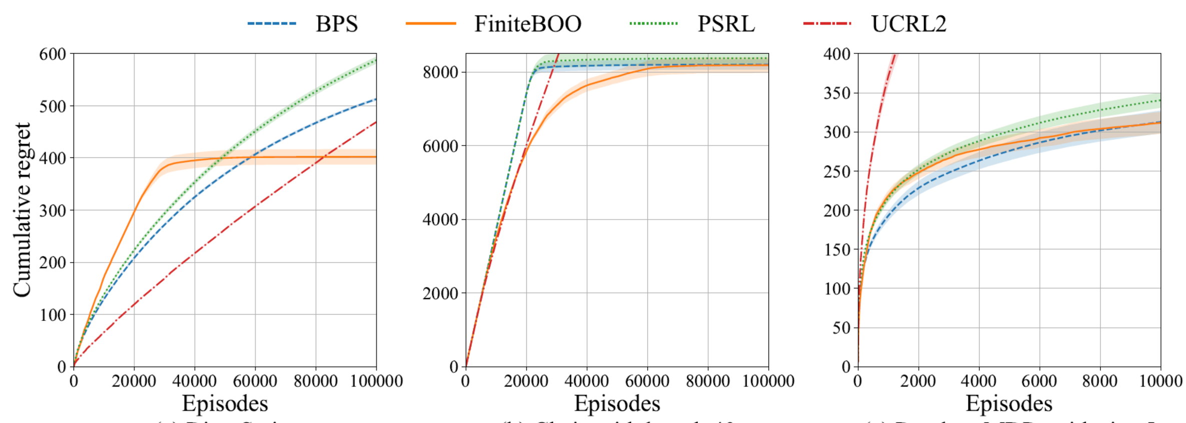


Figure 1: Illustrative diagrams for RiverSwim and Chain. Solid and dotted arrows represent actions, "left" and "right", respectively, with the transition probability tagged. The action left never fails, but the action right often fails. Except for the rewards shown in the figure, the rewards of other state-action pairs are all zero.



(a) RiverSwim (b) Chain with length 40 (c) Random MDPs with size 5