Model-based Policy Optimization under Approximate Bayesian Inference

Chaoqi Wang[1], Yuxin Chen[1], Kevin Murphy[2]

[1]University of Chicago; [2]Google DeepMind

Background

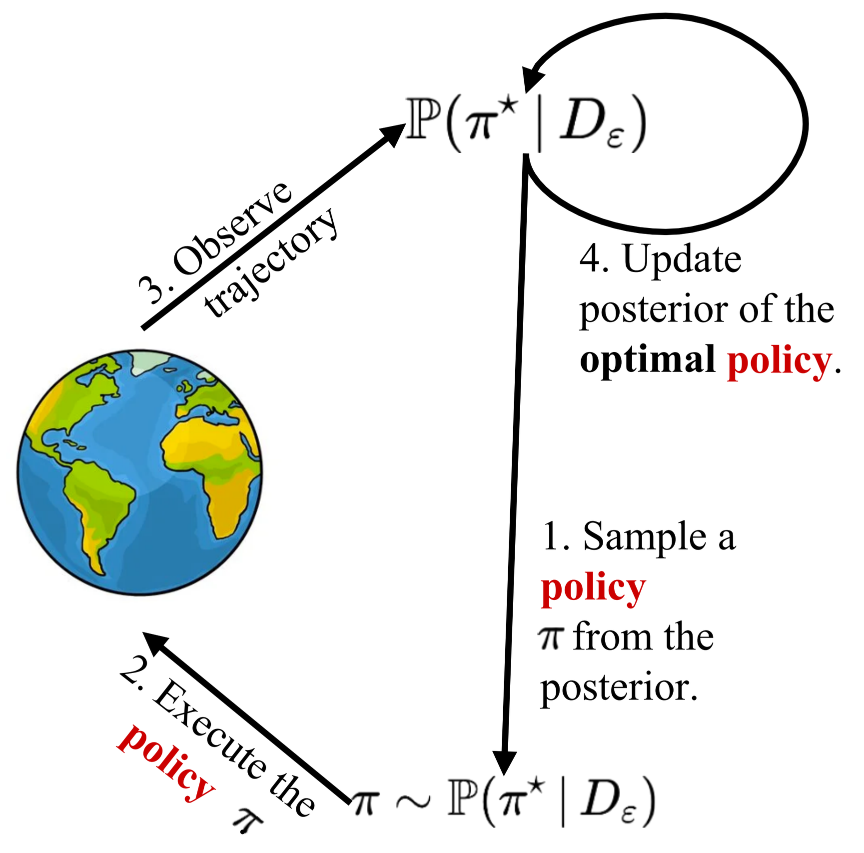

Figure: The pipeline of posterior sampling reinforcement learning (PSRL).

$$p(\pi|\mathcal{D}_\mathcal{E}) = \int \delta(\pi|\mathcal{M}) \underbrace{p(\mathcal{M}|\mathcal{D}_\mathcal{E})}_{\text{posterior over MDPs}} d\mathcal{M}, \quad (1)$$

where $\delta(\pi|\mathcal{M})$ is the Dirac delta distribution, and $\delta(\pi|\mathcal{M}) = 1$ if and only if the policy π optimally solves the MDP $\mathcal{M}$.

- PSRL attains a regret of $\sqrt{K}$ for K episodes.
- More *computationally efficient* than optimism-based methods and information-directed sampling.

But, the theoretical guarantee only holds under exact inference!

Research Question:

- What's the regret bound under approximate inference?
- What would be a good choice for approximating $p(\pi|\mathcal{D}_\mathcal{E})$?

Bayesian Regret under Approximate Inference

Theorem

For K episodes, the Bayesian regret of posterior sampling reinforcement learning algorithm $\mathcal{A}$ with any approximate posterior distribution q_k at episode k is upper bounded by

$$\sqrt{CK(HR_{\max})^2 \mathbb{H}(\pi^\star)} + 2HR_{\max}\sum_{k=1}^{K}\sqrt{\mathbb{E}\left[\mathbf{d}_{KL}\left(q_k(\pi)|\, p_k(\pi)\right)\right]},$$

where $\mathbb{H}(\pi^\star)$ is the entropy of the prior distribution of polices, i.e., $p(\pi) = \int \delta(\pi|\mathcal{M})p(\mathcal{M})d\mathcal{M}$, and C is some problem-dependent constant.

Issues with Existing Solutions

Q: What would be a good choice for approximating $p(\pi|\mathcal{D}_\mathcal{E})$?

$$q^\delta(\pi|\mathcal{D}_\mathcal{E}) = \int \delta(\pi|\mathcal{M})q(\mathcal{M}|\mathcal{D}_\mathcal{E})d\mathcal{M}.$$

$q(\mathcal{M}|\mathcal{D}_\mathcal{E})$ is usually implemented with deep ensemble or Bayesian neural networks. However, $q^\delta(\mathcal{M}|\mathcal{D}_\mathcal{E})$ can perform arbitrarily poorly in terms of the KL divergence!

EXAMPLE 1. SUBOPTIMALITY OF $q^\delta(\pi|\mathcal{D}_\mathcal{E})$.

Consider a toy setting, where the support set of MDPs is $\{\mathcal{M}_1, \mathcal{M}_2\}$, and the support set of policies is $\{\pi_1, \pi_2\}$. Suppose that the true posterior distribution of MDPs is $p(\mathcal{M}_1|\mathcal{D}_\mathcal{E}) = 1/3$, $p(\mathcal{M}_2|\mathcal{D}_\mathcal{E}) = 2/3$, and the optimal policy per MDP is $\delta(\pi_1|\mathcal{M}_1) = 1$ and $\delta(\pi_2|\mathcal{M}_2) = 1$. This we get the following exact distribution over policies: $p(\pi|\mathcal{D}_\mathcal{E})$ is

$$p(\pi|\mathcal{D}_\mathcal{E}) = \underbrace{\begin{bmatrix} \delta(\pi_1|\mathcal{M}_1){=}1,\ \delta(\pi_1|\mathcal{M}_2){=}0 \\ \delta(\pi_2|\mathcal{M}_1){=}0,\ \delta(\pi_2|\mathcal{M}_2){=}1 \end{bmatrix}}_{\delta(\pi|\mathcal{M})} \underbrace{\begin{bmatrix} p(\mathcal{M}_1|\mathcal{D}_\mathcal{E}){=}\frac{2}{3} \\ p(\mathcal{M}_2|\mathcal{D}_\mathcal{E}){=}\frac{1}{3} \end{bmatrix}}_{p(\mathcal{M}|\mathcal{D}_\mathcal{E})} = \begin{bmatrix} p(\pi_1|\mathcal{D}_\mathcal{E}){=}\frac{2}{3} \\ p(\pi_2|\mathcal{D}_\mathcal{E}){=}\frac{1}{3} \end{bmatrix}$$

Now suppose we use the approximate posterior distribution over models, $q(\mathcal{M}_1|\mathcal{D}_\mathcal{E}) = 0$ and $q(\mathcal{M}_2|\mathcal{D}_\mathcal{E}) = 1$. We can optimize $q(\pi|\mathcal{M})$ by minimizing $\mathbf{d}_{\mathrm{KL}}\left(q(\pi|\mathcal{D}_\mathcal{E})|\, p(\pi|\mathcal{D}_\mathcal{E})\right)$. One solution could be

$$q(\pi|\mathcal{D}_\mathcal{E}) = \underbrace{\begin{bmatrix} q(\pi_1|\mathcal{M}_1){=}\frac{1}{2},\ q(\pi_1|\mathcal{M}_2){=}\frac{2}{3} \\ q(\pi_2|\mathcal{M}_1){=}\frac{1}{2},\ q(\pi_2|\mathcal{M}_2){=}\frac{1}{3} \end{bmatrix}}_{q(\pi|\mathcal{M})} \underbrace{\begin{bmatrix} q(\mathcal{M}_1|\mathcal{D}_\mathcal{E}){=}0 \\ q(\mathcal{M}_2|\mathcal{D}_\mathcal{E}){=}1 \end{bmatrix}}_{q(\mathcal{M}|\mathcal{D}_\mathcal{E})} = \begin{bmatrix} q(\pi_1|\mathcal{D}_\mathcal{E}){=}\frac{2}{3} \\ q(\pi_2|\mathcal{D}_\mathcal{E}){=}\frac{1}{3} \end{bmatrix}$$

We see that the optimal $q(\pi|\mathcal{M})$ requires modeling uncertainty in the policy even conditional on the model. By contrast, if we adopt $q^\delta(\pi|\mathcal{D}_\mathcal{E})$ as our approximation, we will have

$$\mathbf{d}_{\mathrm{KL}}\left(q^\delta(\pi|\mathcal{D}_\mathcal{E})|\, p(\pi|\mathcal{D}_\mathcal{E})\right) = \log 3 = \max_{q \in \Delta^1} \mathbf{d}_{\mathrm{KL}}\left(q(\pi|\mathcal{D}_\mathcal{E})|\, p(\pi|\mathcal{D}_\mathcal{E})\right).$$

Observation: Approximation error of $q(\mathcal{M}|\mathcal{D}_\mathcal{E})$ ruins $q^\delta(\pi|\mathcal{D}_\mathcal{E})$.

A Better Choice of $q(\pi|\mathcal{D}_\mathcal{E})$

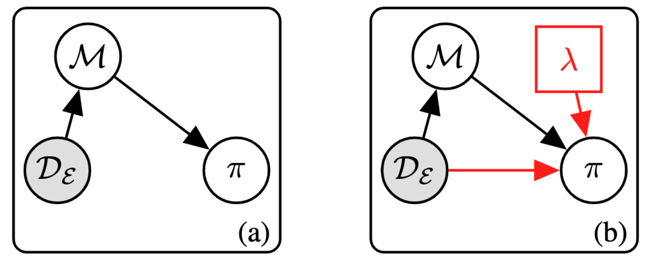

Figure 1: Graphical models for (a) the standard and (b) our posterior over policies π. Differences are shown in red.

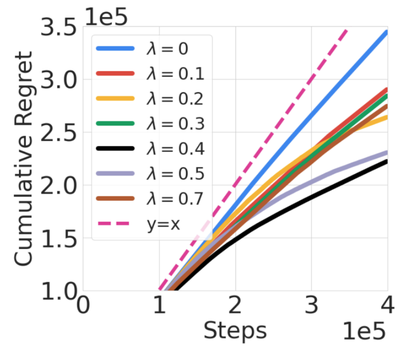

Figure 2: A comparison of cumulative regret for different λ.

A more flexible posterior decomposition to handle the error:

$$q(\pi|\mathcal{D}_\mathcal{E}, \lambda) = \int q(\pi|\mathcal{M}, \mathcal{D}_\mathcal{E}, \lambda)q(\mathcal{M}|\mathcal{D}_\mathcal{E})d\mathcal{M},$$

where $\lambda \in [0, 1]$. In particular, we define

$$q(\pi|\mathcal{M}, \mathcal{D}_\mathcal{E}, \lambda = 0) = q(\pi|\mathcal{M}) = \delta(\pi|\mathcal{M})$$
$$q(\pi|\mathcal{M}, \mathcal{D}_\mathcal{E}, \lambda = 1) = q(\pi|\mathcal{D}_\mathcal{E})$$

Sampling Policies

Ensemble Sampling (PS). Given the posterior distributions, it remains to specify the sampling approach for policies. The simplest sampling strategy is uniform sampling,

$$\pi \sim \mathcal{U}(\{\pi_{1,1}, \ldots, \pi_{N,M}\}).$$

Optimistic Ensemble Sampling (OPS). PS may overly explore unpromising regions, hence we propose OPS, which gradually discards unpromising ensemble members.

$$p_k(\pi = \pi_i) := \frac{\exp\left(\sum_{l=1}^{k} R_\mathcal{E}(\pi_i, l)/\tau\right)}{\sum_{j=1}^{N\cdot M}\exp\left(\sum_{l=1}^{k} R_\mathcal{E}(\pi_j, l)/\tau\right)},$$

where τ controls the level of optimism, and $R_\mathcal{E}(\pi_i, l)$ is the empirical cumulative reward of π_i at the l_{th} episode.

Practical Algorithm: (O)PS-MBPO

Require: Initialize an ensemble of dynamics models $\Theta = \{\hat{\boldsymbol{\theta}}_n\}_{n=1}^{N}$ i.i.d. $\sim q(\boldsymbol{\theta})$.
Require: Initialize an ensemble of policy networks $\Phi = \{\hat{\phi}_{n,m}\}_{n,m=1}^{N,M}$ i.i.d. $\sim q(\boldsymbol{\phi})$.
Require: Initialize empty datasets $\mathcal{D}_\mathcal{E}$ and $\{\mathcal{D}_\mathcal{M}^{n,m}\}_{n,m=1}^{N,M}$. Real data vs. synthetic data ratio λ.

1: **for** K episodes **do**
▷ /* Dynamics training. (Line 2) */
2: Train the ensemble models Θ on $\mathcal{D}_\mathcal{E}$ using MLE.
▷ /* Policy sampling. (Line 3) */
3: Sample a policy π from Φ **uniformly at random** or based on **the optimistic distribution** .
4: Sample state s_1 from the initial state distribution $\rho(\mathrm{s})$
5: **for** $h = 2 : H$ steps **do**
▷ /* Data collection. (Lines 6-11) */
6: s_h = rollout(world dynamics $\mathcal{E}$, π, s_{h-1}, #steps 1)
7: Add s_h to $\mathcal{D}_\mathcal{E}$
8: Sample state $\mathrm{s} \sim \mathcal{D}_\mathcal{E}$
9: **for** each model n, policy m **do**
10: $\mathcal{D}_\mathcal{M}^{n,m}$ = rollout(dynamics $\hat{\boldsymbol{\theta}}_n$, policy $\hat{\phi}_{n,m}$, s, R)
11: Created mixed dataset $D = \lambda\mathcal{D}_\mathcal{E} + (1-\lambda)\mathcal{D}_\mathcal{M}^{n,m}$
▷ /* Policy optimization (Line 12) */
12: $\hat{\phi}_{n,m}$ = update-policy($\hat{\phi}_{n,m}$, D, #steps G)
13: **end for**
14: **end for**
15: **Update the optimistic policy distribution.**
16: **end for**

Experimental Results

1. Results on Continuous Control Benchmarks

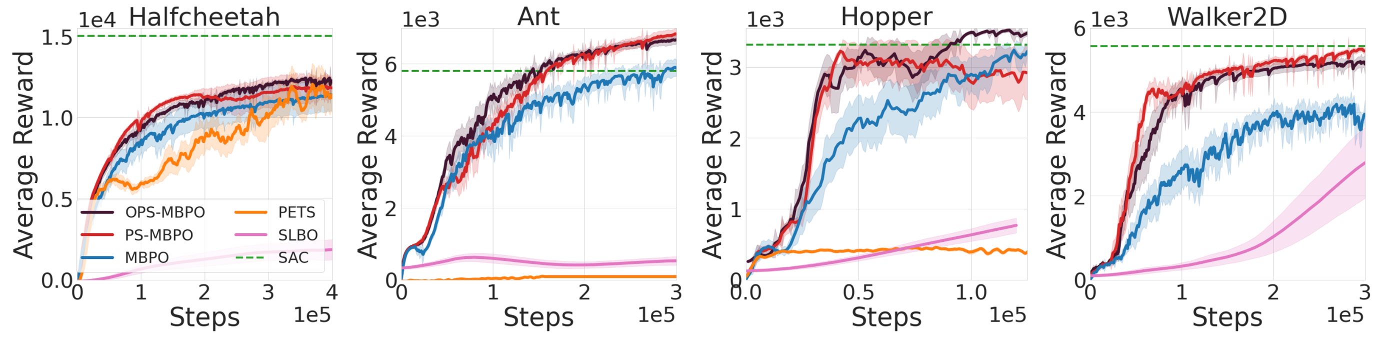

Figure: Comparisons on four tasks with dense rewards.

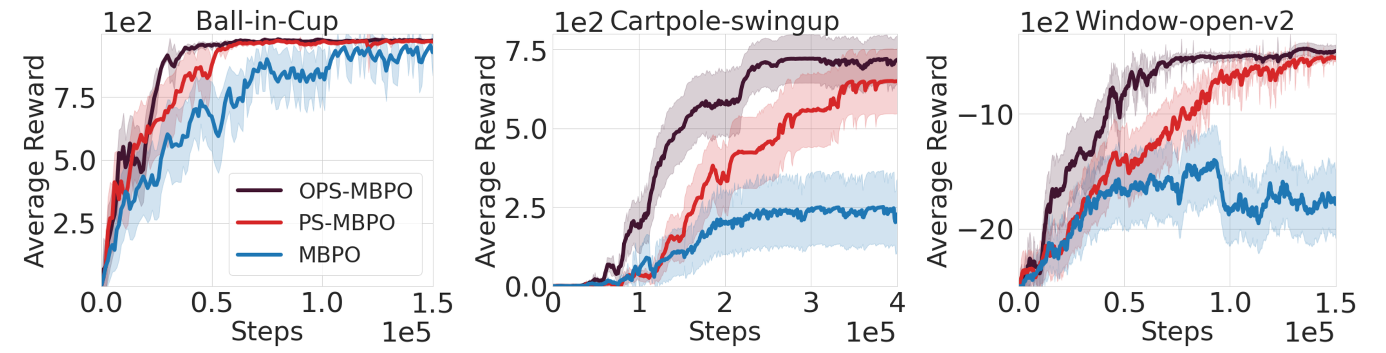

Figure: Comparisons on four tasks with sparse rewards.

2. Ablation Studies

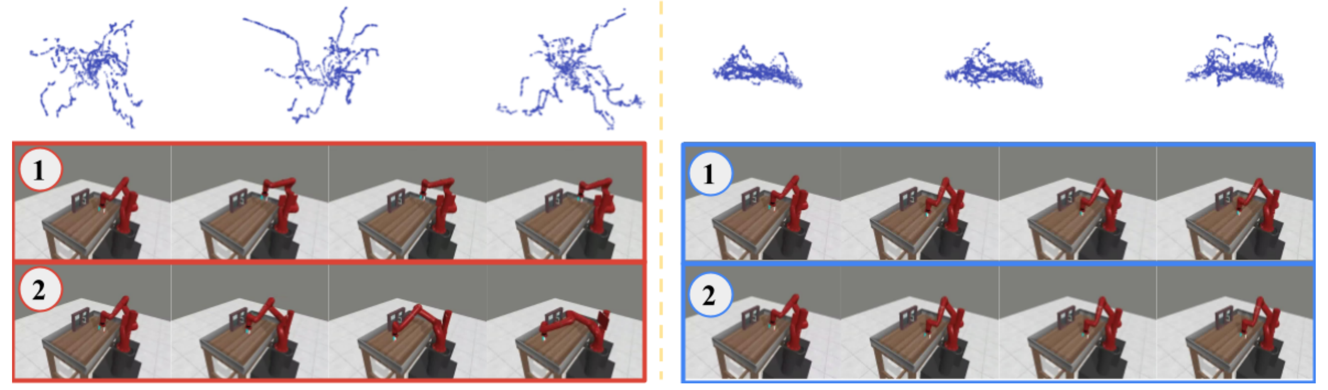

Figure: Visualization of the visited state space of PS-MBPO (top left) and MBPO (top right) on `Window-open-v2`.

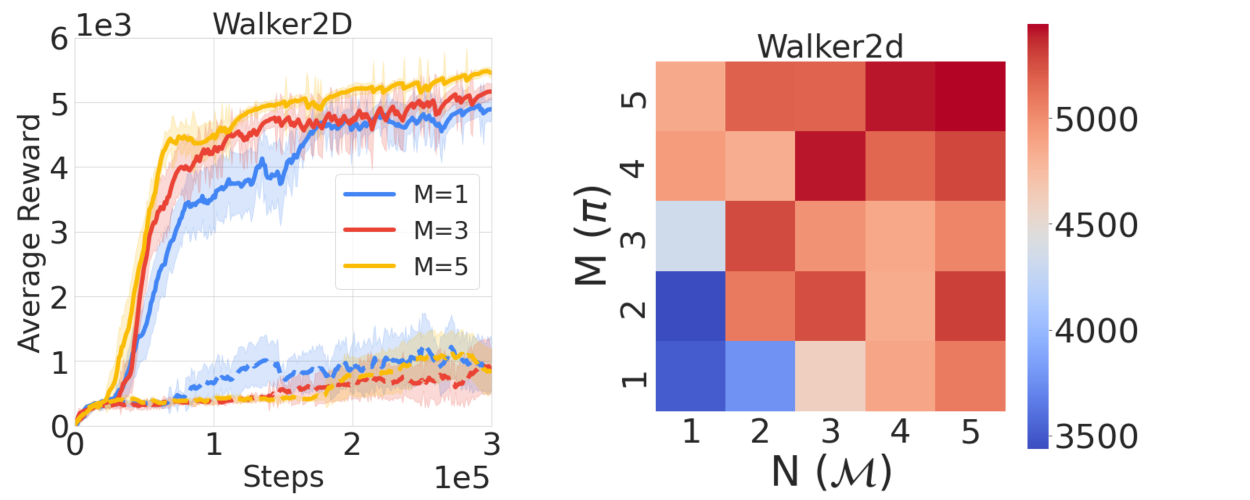

Figure: Left: Ablation study on the performance of with (solid curves) and without (dashed curves) the sampling step. Right: Average reward for varying number of dynamics model (N) and policies (M).

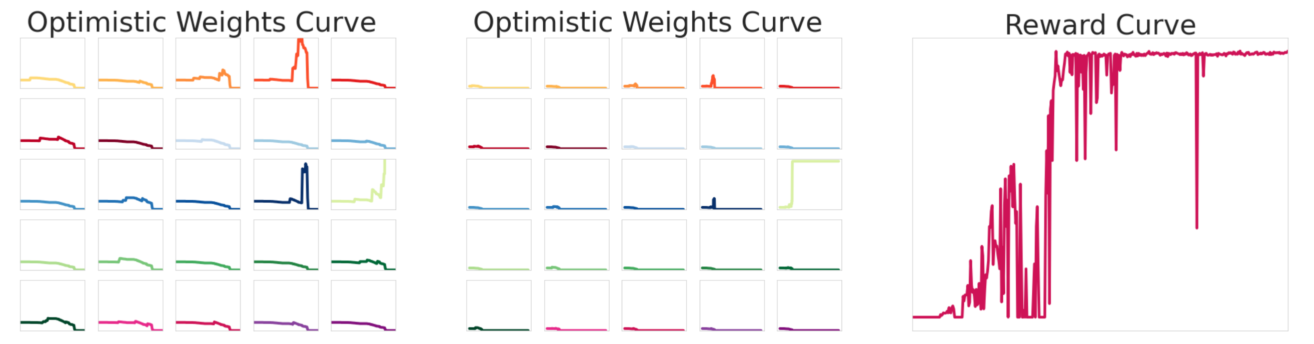

Figure: The optimistic weights of the first 100K iterations (left) and during the entire training process (middle), and the reward curve (right) on `Cartpole-Swingup`.