# Conservative Dual Policy Optimization for Efficient Model-Based Reinforcement Learning

NeurIPS 2022

Shenao Zhang

Georgia Institute of Technology

# Provable Model-Based Reinforcement Learning

Different from greedy algorithms, provable MBRL often leverages the uncertainty:

- Optimism in the Face of Uncertainty (OFU)

$$\pi_t = \operatorname*{argmax}_{\pi} \max_{f_t \in \mathcal{F}_t} V_\pi^{f_t}. \tag{1}$$

- Posterior Sampling RL (or Thompson Sampling)

$$f_t \sim \phi(\cdot|\mathcal{D}_t), \; \pi_t = \operatorname*{argmax}_{\pi} V_\pi^{f_t}. \tag{2}$$

Sublinear regret $\tilde{O}(\sqrt{dT})$. Model complexity $d$ capture how effectively the observed samples can extrapolate to unobserved transitions.

**Theorem 1.** (Eluder Dimension of Nonlinear Models [Dong et al. 2021]) The eluder dimension of one-layer ReLU neural networks is at least $\Omega(\varepsilon^{-(d-1)})$, where $d$ is the state-action dimension.

# Limitations

- Assumption on the restricted model complexity is strong. Nonlinear model complexity is exponential in dimension.
- Over-exploration. Intuition: Explore regions with higher uncertainty and the optimistic/sampled model can be unrealistic.
- Policy is optimized for uncertainty elimination, not for value improvement. Each step only eliminates a small portion of uncertainty.

# Conservative Dual Policy Optimization

Sampling in PSRL is harmful. Can we abandon sampling while still provably exploring?

Selecting a reference model and optimizing a policy w.r.t. it resembles the sampling-then-optimization procedure in PSRL, while offering more stability when the reference is steady.

- Referential Update.

$$q_t = \operatorname*{argmax}_q V_q^{\hat{f}_t^{LS}}$$

- Constrained Conservative Update.

$$\pi_t = \operatorname*{argmax}_\pi \mathbb{E}\left[V_\pi^{f_t} \,\middle|\, \mathcal{H}_t\right], \text{ s.t. } \mathbb{E}_{s\sim\nu_{q_t}}\left[D_{\mathrm{TV}}\left(\pi_t(\cdot|s), q_t(\cdot|s)\right)\right] \le \eta$$

**Theorem 2.** [CDPO Matches PSRL in BayesRegret] Let $\pi^{\text{PSRL}}$ be the policy of any posterior sampling algorithm for reinforcement learning optimized by (2). If the $\mathrm{BayesRegret}$ bound of $\pi^{\text{PSRL}}$ satisfies that for any $T > 0$, $\mathrm{BayesRegret}(T, \pi^{\text{PSRL}}, \phi) \leq \mathcal{D}$, then for all $T > 0$, we have for the CDPO policy $\pi^{\text{CDPO}}$ that $\mathrm{BayesRegret}(T, \pi^{\text{CDPO}}, \phi) \leq 3\mathcal{D}$.

CDPO satisfies the following properties simultaneously:

- Global optimal with sublinear regret.
- Monotonic policy value improvement.

**Theorem 3.** [Policy Iterative Improvement] Suppose we have $\|\tilde{f}(\cdot,\cdot)\| \le C$ for $\tilde{f} \in \mathcal{F}$ where the model class $\mathcal{F}$ is finite. Define $\iota := \max_{s,a} |A_\pi^{f^*}(s,a)|$, where $A_\pi^{f^*}$ is the advantage function defined as $A_\pi^{f^*}(s,a) := Q_\pi^{f^*}(s,a) - V_\pi^{f^*}(s)$. With probability at least $1-\delta$, the policy improvement between successive iterations is bounded by

$$J(\pi_t) - J(\pi_{t-1}) \ge \Delta(t) - (1+\kappa) \cdot \frac{22\gamma C^2 \ln(|\mathcal{F}|/\delta)}{(1-\gamma)H} - \frac{2\eta\iota}{1-\gamma},$$

where $\Delta(t) := \mathbb{E}_{s\sim\zeta}\left[V_{q_t}^{\tilde{f}_t}(s) - V_{q_{t-1}}^{\tilde{f}_t}(s)\right] \ge 0$ due to the greediness of $q_t$.

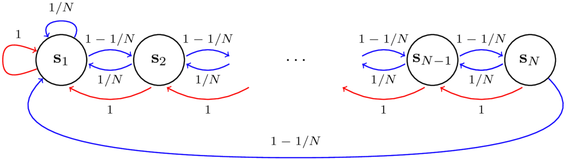
**Theorem 4.**[Expected Regret of CDPO] Let $N(\mathcal{F}, \alpha, \|\cdot\|_2)$ be the $\alpha$-covering number of $\mathcal{F}$. Denote $d_E := \dim_E(\mathcal{F}, T^{-1})$ for the eluder dimension of $\mathcal{F}$ at precision $1/T$. Under Lipschitz assumptions, the cumulative expected regret of CDPO in $T$ iterations is bounded by

$$\mathrm{BayesRegret}(T, \pi, \phi) \le \frac{\gamma T(3T-5)L}{(T-1)(T-2)} \cdot \left(1 + \frac{1}{1-\gamma} C d_E + 4\sqrt{T d_E \beta}\right) + 4\gamma C,$$

where $L := \mathbb{E}[L_t]$ and

$$\beta := 8\sigma^2 \log\left(2N(\mathcal{F}, 1/(T^2), \|\cdot\|_2)\, T\right) + 2(8C + \sqrt{8\sigma^2 \log(8T^3)})/T.$$

Tabular $N$-Chain MDP:




*Right* actions are optimal, *left* actions are suboptimal, at each of the $N$ states.

Different Exploration Mechanisms in the tabular $N$-Chain MDPs: CDPO gives more accurate and certain estimates *only* for the optimal *right* actions, while PSRL explores *both* directions.

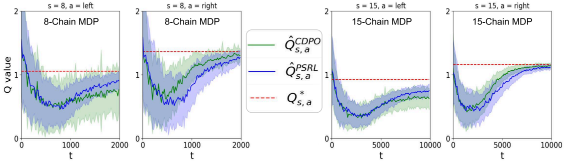



Figure 1: CDPO and PSRL posterior on an 8-Chain MDP and a 15-Chain MDP, where the *right* actions are optimal.

Over-exploration issue in PSRL: as long as the uncertainty contains unrealistically large values, it can perform uninformative exploration according to an inaccurate *sampled* model.

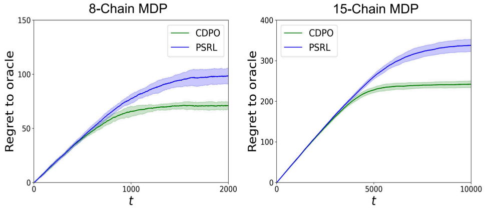
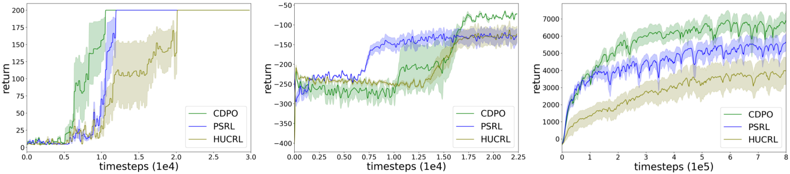



Figure 2: Regret curve of CDPO and PSRL when $N = 8$ and $N = 15$.

Although CDPO has much larger uncertainty for the suboptimal *left* actions, its regret is lower.

# Mujoco Experiments

Exploration Efficiency with Nonlinear Model Class:

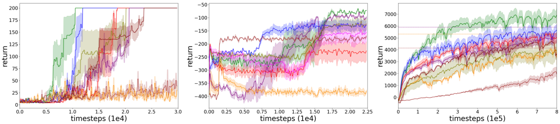



(a) Inverted Pendulum. (b) 7-DOF Pusher. (c) Half-Cheetah.

Figure 3: Performance of CDPO, PSRL, and HUCRL equipped with nonlinear models in several MuJoCo tasks: inverted pendulum swing-up, pusher goal-reaching, and half-cheetah locomotion.

In higher dimensional tasks such as half-cheetah, CDPO achieves a higher asymptotic value with faster convergence.

## Full Results:

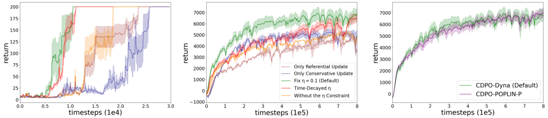



(a) Inverted Pendulum. (b) 7-DOF Pusher. (c) Half-Cheetah.

CDPO PSRL HUCRL SLBO ME-TRPO
MBPO MPO SAC PPO Convergence

Figure 4: Comparison between CDPO and model-free, model-based RL baseline algorithms.

## Ablation Study:




(a) Inverted Pendulum. (b) Half-Cheetah. (c) Half-Cheetah.

Figure 5: Ablation studies on the effect of the dual update steps and the trust-region constraint. The robustness and generalizability of the CDPO framework are demonstrated by the results of different choices of the constraint threshold and different solvers.

*Thanks!*