Exponential Family Model-Based Reinforcement Learning via Score Matching

Gene Li [1] Junbo Li [2] Anmol Kabra [1] Nathan Srebro [1] Zhaoran Wang [3] Zhuoran Yang [4]

[1]TTI Chicago [2]UC Santa Cruz [3]Northwestern University [4]Yale University

Problem setting

Consider the setting of online learning in finite horizon episodic Markov Decision Process: $\mathsf{MDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ is the state space, $\mathcal{A}$ is any arbitrary action set, $H \in \mathbb{N}$ is the horizon.

Reward function. $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is deterministic and **known**.

Transition probability. $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ follows an exponential family model introduced by Chowdhury et al. (2021):

$$\mathbb{P}_{W_0}(s' \mid s, a) = q(s') \cdot \exp\left(\langle \psi(s'), W_0\phi(s, a)\rangle - Z_{sa}(W_0)\right), \quad (1)$$

where *feature mappings* $\psi : \mathcal{S} \to \mathbb{R}^{d_\psi}$ and $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_\phi}$, and *base measure* $q : \mathcal{S} \to \mathbb{R}$ are **known**, but matrix $W_0 \in \mathbb{R}^{d_\psi \times d_\phi}$ is **unknown**.

Interacting with the MDP

In every round $k \in [K]$:

- Observe initial state s_1^k.
- Select policy $\pi^k : \mathcal{S} \to \mathcal{A}$
- Run policy on MDP and observe trajectory $\{(s_h, a_h, r_h)\}_{h \in [H]}$, where

$$a_h = \pi^k(s_h), r_h = r(s_h, a_h), \text{ and } s_{h+1} \sim \mathbb{P}(\cdot | s_h, a_h), \quad \text{for all } h \in [H].$$

Objective

Value functions. For any policy π, denote $V_h^\pi : \mathcal{S} \to \mathbb{R}$ as the expected value of future cumulative rewards when the learner plays π starting from a state in step h:

$$V_h^\pi(s) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \,\middle|\, s_h = s, a_{h:H} \sim \pi\right].$$

We also let $V_h^\pi(\cdot; W)$ denotes value function under transition param. by W.

Optimal policy. Denote $\pi^\star$ to be a policy such that $V_h^{\pi^\star}(s)$ is maximized at every state s and step h.

Measure performance as regret against the optimal policy:

$$\mathsf{Regret}(K) := \sum_{k=1}^{K}\left(V_1^{\pi^\star}(s_1^k) - V_1^{\pi^k}(s_1^k)\right).$$

Discussion on model assumption

The exponential family transition in Equation (1) captures previously studied models in RL.

Special case: (non)linear dynamical systems

Linear dynamical systems are an important theoretical model; they govern the dynamics for the linear quadratic regulator (LQR).

$$s' = As + Ba + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, \Sigma).$$

Mania et al. (2020); Kakade et al. (2020) study nonlinear extensions:

$$s' = W_0\phi(s, a) + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, \Sigma).$$

Exponential family transitions model a richer class of densities beyond (non)linear dynamical systems due to the added flexibility in q and ψ!
E.g., nonadditive and nongaussian noise.

Motivation: how do we do model estimation?

Chowdhury et al. (2021) propose an optimistic model-based RL algorithm called Exp-UCRL that uses MLE for model estimation.

- Estimating model parameter W_0 with MLE requires computing the log-partition function $Z_{sa}(\cdot)$.
- For nonlinear dynamical systems, this is efficient.
- In general, one can estimate $Z_{sa}(\cdot)$ via Markov Chain Monte Carlo methods, but this can be *slow* and *induce approximation errors*.

Exp-UCRL is *statistically* efficient, but not *computationally* efficient in general, so we need an alternative model estimation procedure.

Our approach: score matching

Score matching (Hyvärinen, 2005)

For any (s, a), the population loss function is

$$J(W) := \frac{1}{2}\int_{\mathcal{S}} \mathbb{P}_{W_0}(s' \mid s, a)\left\|\nabla_{s'} \log \frac{\mathbb{P}_{W_0}(s' \mid s, a)}{\mathbb{P}_W(s' \mid s, a)}\right\|^2 ds'.$$

$$= \frac{1}{2}\int_{\mathcal{S}} \mathbb{P}_{W_0}(s' \mid s, a)\sum_{i=1}^{d_s}\left((\partial_i \log \mathbb{P}_W(s' \mid s, a))^2 + 2\partial_i^2 \log \mathbb{P}_W(s' \mid s, a)\right) ds' + C,$$

where second line uses integration by parts trick under some regularity conditions (see paper for more details).

- Score matching is an unnormalized density estimation procedure, which does not require computing of $Z_{sa}(\cdot)$.
- Empirical loss function $\hat{J}(W)$ can be minimized via $d_\phi \cdot d_\psi$-dimensional ridge regression problem $\Rightarrow$ **computationally efficient!**

Algorithm: score matching for reinforcement learning (SMRL)

We use score matching as a subroutine for parameter estimation for an optimistic planning algorithm, SMRL.

Main result: SMRL algorithm and regret guarantee

In every round $k \in [K]$:

- Estimate $\hat{W} = \min_W \hat{J}(W) + \frac{\lambda}{2}\|W\|_F^2$ using transition samples from previous $k - 1$ episodes.
- Construct confidence set $\mathcal{W}_k$ centered at $\hat{W}$.
- Choose the optimistic policy $\pi^k = \arg\max_\pi \sup_{W \in \mathcal{W}_k} V_1^\pi(s_1^k; W)$.

Regret guarantee. With high probability, SMRL achieves regret:

$$\mathsf{Regret}(K) \le \tilde{O}\left(d_\psi d_\phi \sqrt{H^3 T}\right),$$

where $\tilde{O}(\cdot)$ hides log factors and poly factors of problem constants.

Remark. Optimistic planning can be NP-hard, but this step can be approximated by model predictive control algorithms.

Proof ingredients.

1. Show that whp, for all episodes $k \in [K]$, that $W_0 \in \mathcal{W}_k$.
2. By optimism, regret is bounded by (learners est. of value of π^k) − (true value of π^k).
3. Bound the difference in value function under distributions $\tilde{W}_k$ and W_0, where $\tilde{W}_k$ is the model attaining supremum in the optimistic planning step.

Experiments

We demonstrate the benefit of using SMRL with an expressive transition model vs the conventional approach of fitting an LDS (Kakade et al., 2020).

Experimental problem

Consider a synthetic MDP with the following multimodal transition function and reward structure.

Multimodal characteristic of MDP

- Next state density $\mathbb{P}$ for $a = +1$ and $a = -1$ have disjoint modes.
- Crests for $\mathbb{P}(s' \mid s, a = +1)$ are located at troughs for $\mathbb{P}(s' \mid s, a = -1)$, and vice versa.
- Rewards peak at crests of $\mathbb{P}(s' \mid s, a = +1) \Rightarrow a = +1$ is always the optimal action.

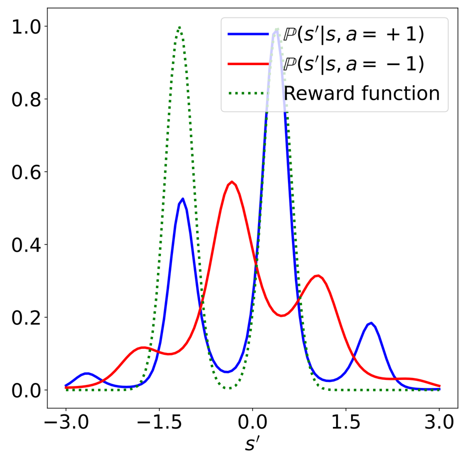


Figure 1. Synthetic MDP

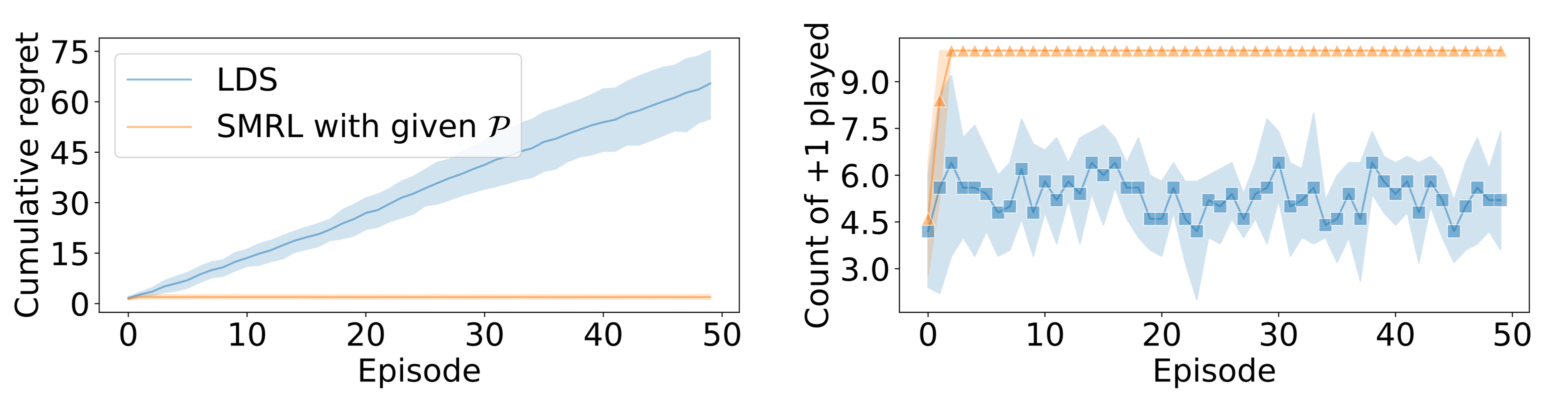
Experimental setup

We fix the simple random sampling shooting planner that at every step, (i) simulates *lookaheads* of playing $[+1, \ldots, +1]$ and $[-1, \ldots, -1]$, and (ii) chooses action depending on which yields higher reward. We compare these model estimation methods:

1. Using SMRL with given transition probability class $\mathcal{P}$.
2. Fitting an LDS using MLE to get $\hat{W}_k$.

Results

- Score matching estimates transition density well, and the planner quickly learns to play the optimal action (Figure 2b).
- LDS is not expressive enough to distinguish action choices.



(a) Cumulative regret (b) Planner's a choices

Figure 2. SMRL with expressive density vs LDS. Regret is w.r.t. the planner with ground truth model ($\hat{W}_k = W_0$), a surrogate for the optimal policy.

References

Sayak Ray Chowdhury, Aditya Gopalan, and Odalric-Ambrym Maillard. Reinforcement learning in parametric mdps with exponential families. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1855–1863, 2021.

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *Advances in Neural Information Processing Systems*, pages 15312–15325, 2020.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

NeurIPS 2022 (Selected as Oral)

gene@ttic.edu, anmol@ttic.edu