# Exponential Family Model-Based Reinforcement Learning via Score Matching

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

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

## Problem Setting

We consider the setting of online learning in a finite horizon episodic Markov Decision Process: MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$.

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

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

Objective: minimize

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

**Question.** How can we leverage function approximation to design statistically and computationally efficient algorithms?

## Exponential Family Transitions [1]

**Assumption.**

Suppose $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ and $\mathcal{A}$ is any arbitrary action set.

- *Transition Probabilities*: Let *feature mappings* $\psi : \mathcal{S} \mapsto \mathbb{R}^{d_\psi}$ and $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_\phi}$, as well as *base measure* $q : \mathcal{S} \to \mathbb{R}$ be known to the learner.
  The state transition measures are conditional exp. family models, parameterized by an unknown matrix $W_0 \in \mathbb{R}^{d_\psi \times d_\phi}$:
  $$\mathbb{P}_{W_0}(s'|s,a) = q(s') \exp\left(\langle \psi(s'), W_0 \phi(s,a) \rangle - Z_{sa}(W_0)\right).$$
- *Rewards*: We assume that the rewards $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ are bounded a.s. in $[0,1]$ and known to the learner.

## Motivation

- Prior work imposes strong model-based or model-free *linearity* assumptions.
- Nonlinear settings are not well understood theoretically.

### Special Case: (non)Linear Dynamical Systems

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

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

Recent work [2, 3] considers nonlinear extensions, where the transition dynamics are:

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

## Model Estimation via Score Matching

### Issues with MLE

- Estimating model parameters with MLE requires computing the log partition function $Z_{sa}(\cdot)$.
- Practically, estimating the log partition function can be done via Markov Chain Monte Carlo methods, but these are *slow* and *induce approximation errors*.
- Approximation errors in model estimation can propagate to planning in an undesirable way.

### Score Matching

Instead, we propose to use *score matching*, an unnormalized density estimation procedure [4].

For any $(s,a)$ pair, we can define the *Fischer divergence* between two conditional distributions on $s'$:

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

**Key observation:** under some regularity conditions, $J(\mathbb{P}_{W_0} \| \mathbb{P}_W)$ can be estimated with samples as:

$$\hat{J}(W) := \frac{1}{2} \sum_{t=1}^{n} \sum_{i=1}^{d_s} \left( (\partial_i \log \mathbb{P}_W(s_t'|s_t,a_t))^2 + 2\partial_i^2 \log \mathbb{P}_W(s_t'|s_t,a_t) \right)$$

This loss function can be minimized by solving a $d_\phi d_\psi$-dimensional ridge regression problem.

We use score matching as a subroutine for parameter estimation for an optimistic planning algorithm and prove the following regret guarantee:

## Main Result: Regret Guarantee for SMRL

With high probability, SMRL achieves the regret guarantee of:

$$\text{Regret}(K) \leq \tilde{O}\left( d_\psi d_\phi \sqrt{H^3 T} \right).$$

(This matches the guarantee provided by [1], who use MLE instead of score matching.)

### Proof Ingredients

1. Show that with high probability, for all episodes $k \in [K]$: the ground truth lies in a shrinking confidence set, i.e. $W_0 \in \mathcal{W}_k$.
2. By optimism, the regret is bounded by the learners estimate of the value of $\pi^k$ minus the true value of $\pi^k$.
3. Apply information-theoretic machinery to bound the difference in value function under distributions $\tilde{W}_k$ and $W_0$.

## Algorithm: Score Matching for RL

**Algorithm 1** Score Matching for RL (SMRL)

1: **Input:** Regularizer $\lambda$ and constants (omitted for clarity)
2: **Initialize:** starting confidence set $\mathcal{W}_1 = \mathbb{R}^{d_\psi \times d_\phi}$, confidence widths $\{\beta_k\}_{k \geq 1}$, dataset $\mathcal{D} = \emptyset$.
3: **for** episode $k = 1, 2, 3, \cdots, K$ **do**
4: Observe initial state $s_1^k$
5: Choose the optimistic policy:
$$\pi^k = \arg\max_{\pi} \max_{W \in \mathcal{W}_k} V^{\pi}_{\mathbb{P}_W,1}(s_1^k)$$
6: Execute $\pi^k$ to get $\tau = \{(s_h^k, a_h^k, r_h^k)\}_{h \in [H]}$ and add it to $\mathcal{D}$.
7: Solve for score matching estimator:
$$\hat{W}_k = \arg\min_{W} \hat{J}(W) + \frac{\lambda}{2} \|W\|_F^2$$
8: Compute confidence set $\mathcal{W}_{k+1}$

## Discussion

### When should we use score matching?

- Score matching requires more regularity conditions than MLE does. In particular, it requires $\mathcal{S}$ to be a Euclidean space and $\psi$ to be twice-differentiable.
- Score matching provides a computationally tractable estimator and simpler analysis.

### Future Directions

- Optimistic planning is NP-hard, but we can implement a variant of SMRL with Thompson Sampling and approximate planning algorithms.
- Arbitrary state spaces?
- Analyzing SMRL for kernelized exponential family settings?
- Handling unbounded costs?

## References

[1] 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.

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

[3] 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.

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



gene@ttic.edu, ljb121002@gmail.com