Exponential Family Model-Based Reinforcement Learning via Score Matching

Gene Li
Toyota Technological Institute at Chicago
gene@ttic.edu

Junbo Li
UC Santa Cruz
jli753@ucsc.edu

Anmol Kabra
Toyota Technological Institute at Chicago
anmol@ttic.edu

Nathan Srebro
Toyota Technological Institute at Chicago
nati@ttic.edu

Zhaoran Wang
Northwestern University
zhaoranwang@gmail.com

Zhuoran Yang
Yale University
zhuoran.yang@yale.edu

Abstract

We propose an optimistic model-based algorithm, dubbed SMRL, for finite-horizon episodic reinforcement learning (RL) when the transition model is specified by exponential family distributions with d parameters and the reward is bounded and known. SMRL uses score matching, an unnormalized density estimation technique that enables efficient estimation of the model parameter by ridge regression. Under standard regularity assumptions, SMRL achieves $\tilde{O}(d\sqrt{H^3T})$ online regret, where H is the length of each episode and T is the total number of interactions (ignoring polynomial dependence on structural scale parameters).

1 Introduction

This paper studies the regret minimization problem for finite horizon, episodic reinforcement learning (RL) with infinitely large state and action spaces. Empirically, RL has achieved success in diverse domains, even when the problem size (measured in the number of states and actions) explodes [35, 41, 28]. The key to developing sample-efficient algorithms is to leverage *function approximation*, enabling us to generalize across different state-action pairs. Much theoretical progress has been made towards understanding function approximation in RL. Existing theory typically requires strong linearity assumptions on transition dynamics [e.g., 51, 26, 8, 36] or action-value functions [e.g., 30, 52] of the Markov Decision Process (MDP). However, most real world problems are *nonlinear*. Our theoretical understanding of these settings remains limited. Thus, we ask the question:

Can we design provably efficient RL algorithms in nonlinear environments?

Recently, Chowdhury et al. [11] introduced a nonlinear setting where the state-transition measures are finitely parameterized exponential family models, and they proposed to estimate model parameters via maximum likelihood estimation (MLE). The exponential family is a well-studied and powerful statistical framework, so it is a natural model class to consider beyond linear models. Chowdhury et al. study exponential family transitions of the form:

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

where $\psi \in \mathbb{R}^{d_\psi}$ and $\phi \in \mathbb{R}^{d_\phi}$ are known feature mappings, q is a known base measure, W_0 is the unknown parameter to be learned, and Z_{sa} is the log partition function that ensures the density integrates to 1. This transition model covers both linear dynamical systems as well as the nonlinear dynamical system (nonLDS), introduced by Mania et al. [33]. Linear dynamical systems with quadratic rewards, i.e., the linear quadratic regulator (LQR), have received much attention recently as

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

an important testbench for RL in unknown, complex environments [19, 42, 27]. Thus, the work of Chowdhury et al. is a crucial step in bridging the gap between RL and continuous control.

However, MLE has several shortcomings. In order to estimate the parameter W_0 in (1), MLE requires estimating the log partition function Z_{sa}, which is computationally intensive. Practical implementations for MLE which estimate the log partition function via Markov Chain Monte Carlo (MCMC) methods can be slow and induce approximation errors [10]. These approximation errors can propagate in undesirable ways to the algorithm's planning procedure. Since the MLE $\hat{W}$ cannot be computed in closed form, Chowdhury et al. leave their estimator implicitly defined as solutions of the likelihood equations. As is typical for upper confidence RL (UCRL) algorithms, one constructs high probability confidence sets around the estimator. Due to the challenging modeling assumption, Chowdhury et al. employ confidence sets which are sums of KL divergences taken over the dataset.

In this work, we bypass these difficulties by instead proposing to learn the model parameters with *score matching*, an unnormalized density estimation technique introduced by Hyvärinen [22]. Score matching provides an explicit, easily computable closed form estimator for the model parameters by solving a certain ridge regression problem (Theorem 1). Moreover, we can employ high probability confidence sets which are ellipsoids centered at the estimator, a standard component in prior theoretical work on linear bandits and linear MDPs [e.g., 2, 26].

Our main results are as follows:

- We extend prior work on the score matching estimator in the i.i.d. setting by proving nonasymptotic concentration guarantees for non-i.i.d. data (Theorem 2).
- We consider regret minimization for the setting of exponential family transitions and bounded and known rewards. We design a model-based algorithm, dubbed SMRL, which achieves regret of $\tilde{O}(d\sqrt{H^3T})$, with polynomial dependence on structural scale parameters (Theorem 3). Here, $d = d_\psi \times d_\phi$ is the total number of parameters of W_0, H is the episode length, and T is the total number of interactions. In each episode, SMRL uses score matching as a computationally efficient subroutine to estimate W_0 from data, then it constructs elliptic confidence regions around the estimator which contain W_0 w.h.p. and chooses policies optimistically based on such confidence regions. (This work assumes computational oracle access to an optimistic planner.)

Our regret guarantee matches that of Exp-UCRL, the model-based algorithm proposed by Chowdhury et al. When specialized to the nonLDS with bounded costs and features, score matching and MLE are equivalent estimators (Proposition 10). Here, the work of Kakade et al. [27] gives a tighter guarantee of $\tilde{O}(\sqrt{d_\phi(d_\phi + d_\psi + H)H^2T})$; however we stress that our analysis applies to a broader class of models. Broadly speaking, we view score matching and MLE as complementary estimation techniques; while MLE relies on less assumptions, score matching enjoys computational efficiency and allows us to simplify both the algorithm and proofs. A detailed comparison is deferred to Section 5. In this work, we mainly compare against the papers [11, 27], but a broader summary of related work can be found in Appendix A.

Notation. For a vector $x \in \mathbb{R}^d$, we let $\|x\| := \|x\|_2$ denote the ℓ_2 norm. For a matrix $M \in \mathbb{R}^{n\times d}$, we denote $\mathsf{vec}\,(M) \in \mathbb{R}^{nd}$ to be the vectorized version of M. For a matrix M, we also denote $\|M\|_2$ to be the operator norm and $\|M\|_F$ to be the Frobenius norm, i.e., $\|M\|_F := \|\mathsf{vec}\,(M)\|$. We also let $e_i \in \mathbb{R}^d$ and $E_{ij} \in \mathbb{R}^{n\times d}$ denote the canonical basis vectors and matrices respectively. For positive semidefinite matrices A, B, we let $A \preceq B$ to be $B - A \succeq 0$. For positive semidefinite matrix A and vector x we define $\|x\|_A := \sqrt{x^\top Ax}$. For any $n \in \mathbb{N}$, we let $[n] := \{1, 2, \ldots, n\}$. For a twice differentiable function $f : \mathbb{R}^m \mapsto \mathbb{R}^n$ and any $i \in [m]$, we let $\partial_i f(x) := \left(\frac{\partial}{\partial x_i} f_1(x), \ldots, \frac{\partial}{\partial x_i} f_n(x)\right)^\top \in \mathbb{R}^n$ and $\partial_i^2 f(x) := \left(\frac{\partial^2}{\partial x_i^2} f_1(x), \ldots, \frac{\partial^2}{\partial x_i^2} f_n(x)\right)^\top \in \mathbb{R}^n$. We use the word "algorithm" liberally, since methods discussed in this paper as well as other papers require solving optimization procedures which can be computationally intractable.

2 Problem statement

We consider the setting of an episodic Markov Decision Process, denoted by $\mathrm{MDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $H \in \mathbb{N}$ is the horizon length of each episode, $\mathbb{P}$ is state transition probability measure, and $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function.

The agent interacts with the episodic MDP as follows. At the beginning of each episode, a state s_1 is chosen by an adversary and revealed to the agent. The agent picks a **policy function**, which is a collection of (possibly random) functions $\pi := \{\pi_h : \mathcal{S} \mapsto \Delta(\mathcal{A})\}_{h\in[H]}$ that determines the agent's strategy for interacting with the world. For each step $h \in [H]$, the agent observes the state s_h and plays action $a_h \sim \pi_h(s_h)$. Afterwards, they observe reward $r_h(s_h, a_h)$, and the MDP evolves to a new state $s_{h+1} \sim \mathbb{P}(\cdot \mid s_h, a_h)$. The episode terminates at state s_{H+1} after which the world resets.

The goal of the agent is to maximize their cumulative rewards through interactions with the MDP. Concretely, in our model-based setting the agent knows the reward function r and that the transition model $\mathbb{P}$ lies in some model class $\mathcal{P}$, and they want to pick policies every episode to minimize **regret**, which we formally define later on.

Now we define the value function and action-value function. For every policy π, we can define a **value function** $V^\pi_{\mathbb{P},h} : \mathcal{S} \mapsto \mathbb{R}$, which is the expected value of the cumulative future rewards when the agent plays policy π starting from state s in step h, and the world transitions according to $\mathbb{P}$. In this paper, we include $\mathbb{P}$ in the subscript since we will analyze value functions for different models; if clear from context, we will drop the subscript $\mathbb{P}$. Specifically, we have:

$$V^\pi_{\mathbb{P},h}(s) := \mathbb{E}_{\mathbb{P}}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, a_{h'}\right) \,\middle|\, s_h = s, a_{h:H} \sim \pi\right], \quad \forall s \in \mathcal{S}, h \in [H].$$

Similarly, we define the **action-value** functions $Q^\pi_{\mathbb{P},h}(s,a) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ to be the expected value of cumulative rewards starting from a state-action pair in step h, following π afterwards:

$$Q^\pi_{\mathbb{P},h}(s,a) := \mathbb{E}_{\mathbb{P}}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, a_{h'}\right) \,\middle|\, s_h = s, a_h = a, a_{h+1:H} \sim \pi\right], \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}, h \in [H].$$

An optimal policy $\pi^\star$ is defined to be the policy such that the corresponding value function $V^{\pi^\star}_{\mathbb{P},h}(s)$ is maximized at every state $s \in \mathcal{S}$ and step $h \in [H]$. Without loss of generality, it suffices to consider deterministic policies [48]. Given knowledge of the MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, the optimal value function and action-value function can be computed via dynamic programming [47]; then the optimal policy can be computed as the policy that acts greedily with respect to the optional action-value function, i.e., $\pi^\star_h(s) = \arg\max_{a\in\mathcal{A}} Q^\star_{\mathbb{P},h}(s,a)$.

In the online setting, we will measure the performance of an agent interacting with the MDP over K episodes via the notion of **regret**. In every episode $k \in [K]$, an adversary presents the agent with a state s_1^k, and the agent then chooses a policy π^k. The regret over K episodes is the expected suboptimality of the agent's choice of policy π^k compared to the optimal policy $\pi^\star$:

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

Implicit in the notation $\mathcal{R}(K)$ are the adversary's choice of initial states; our results for regret will hold for any sequence of adversarially chosen $\{s_1^k\}_{k\in[K]}$. We will also denote $T := KH$ as the total number of interactions the agent makes with the world.

2.1 Exponential family transitions

We consider the setting when the transition model class $\mathcal{P}$ is given by exponential family transitions and the the reward function $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is bounded a.s. in $[0,1]$ and known to the learner.[1]

Definition 1 (Exponential family transitions, c.f., [11])**.** *Suppose $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ and $\mathcal{A}$ is any arbitrary action set. Fix 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 $q : \mathcal{S} \to \mathbb{R}$. For any matrix $W \in \mathbb{R}^{d_\psi \times d_\phi}$, let:*

$$\mathbb{P}_W(s'|s,a) := q(s') \exp\left(\langle \psi(s'), W\phi(s,a)\rangle - Z_{sa}(W)\right), \tag{2}$$

[1] Our results extend to settings where the rewards are not known but instead lie in some class $\mathcal{R} \subseteq (\mathcal{S} \times \mathcal{A} \to \mathbb{R})$ by including an additional reward estimation procedure in our algorithm; the regret would additionally depend on the complexity of $\mathcal{R}$.

where $Z_{sa}(\cdot)$ is the log-partition function, which is completely determined once ψ, ϕ, q, and W are specified. Then we define the ***exponential family transitions*** *model class $\mathcal{P}(\psi, \phi, q)$ as:*

$$\mathcal{P}(\psi, \phi, q) := \left\{ \mathbb{P}_W : \int_{\mathcal{S}} q(s') \exp(\langle \psi(s'), W\phi(s,a) \rangle) \, ds' < \infty, \ \forall (s,a) \in \mathcal{S} \times \mathcal{A} \right\}.$$

Since ψ, ϕ, q are taken to be fixed and known to the learner, we will write the model class as $\mathcal{P}$.

Along with this assumption, we introduce a notational convention. Given some real or vector-valued measurable function $f(s')$, we will write $\mathbb{E}_{sa}^W f(s')$ to denote the expected value of f when s' is drawn from the conditional distribution $\mathbb{P}_W(\cdot|s,a)$, i.e. $\mathbb{E}_{sa}^W f(s') := \int_{\mathcal{S}} f(s') \mathbb{P}_W(s'|s,a) ds'$.

Chowdhury et al. state their results for a setting where the unknown matrix $W_0 = \sum_{i=1}^d \theta_i A_i$, where the $A_i \in \mathbb{R}^{d_\psi \times d_\phi}$ are known matrices and $\theta \in \mathbb{R}^d$ is unknown. This setting can be viewed as a nonlinear analog of the linear mixture model considered in [36, 5]. Definition 1 is a special case with $d = d_\psi \times d_\phi$ and $A_{ij} := E_{ij}$. Our results can be extended to their general setting with minor modification. Quantitatively, we would replace factors of $d_\psi \times d_\phi$ with d in both the concentration and regret guarantees, and similar to Chowdhury et al. we would introduce constants which depend on A_i. For simplicity of presentation, we study the fully unknown matrix setting.

2.2 Relationship to (non)linear dynamical systems

We now describe how Definition 1 generalizes the previously studied model class of (non)linear dynamical systems which have been explored in reinforcement learning and control theory literature.

First, we take a step back and describe linear dynamical systems (LDS), which govern the transition dynamics of the LQR problem.[2] An LDS is defined by the following transition dynamics:

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

where $s, s' \in \mathbb{R}^{d_s}$, $a \in \mathbb{R}^{d_a}$, A, B are appropriately sized parameter matrices, and $\Sigma \in \mathbb{R}^{d_s \times d_s}$ is a known covariance matrix. The problem of estimating (A, B), known as *system identification*, has a long history (see Appendix A for more details).

Recently, system identification and regret minimization have been studied for nonlinear generalizations of LDS [33, 27]. In this paper, we refer to this setting as the *nonlinear dynamical system* (or nonLDS for short).[3] The nonLDS is described by the state transition model:

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

By setting $\phi(s,a) = [s,a]^\top$ and $W_0 = [A \ B]$, we recover the classical linear dynamical system. The nonLDS (and by extension the LDS) are special cases of Definition 1. This can be seen by writing out the pdf of the multivariate Gaussian distribution to get:

$$q(s') = \frac{1}{(2\pi)^{d_s/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{\|s'\|_{\Sigma^{-1}}^2}{2} \right), \ \psi(s') = \Sigma^{-1} s', \ Z_{sa}(W_0) = \frac{\|W_0 \phi(s,a)\|_{\Sigma^{-1}}^2}{2}.$$

Lastly, note that Definition 1 is more general than that of the nonLDS, whose base measure $q(\cdot)$ and feature mapping $\psi(\cdot)$ must take a specific form given by the multivariate Gaussian. Definition 1 gives extra flexibility in the functions q, ψ, and ϕ, which can be regarded as *design choices* for the practitioner. For example, one can pick the mapping ψ the output of a neural network which captures the relevant features for the transition to s'; this is not permitted under the nonLDS setting.

3 Model estimation via score matching

In this section, we present the score matching method, the subroutine in our RL algorithm that estimates model parameters. We also introduce structural assumptions that enable us to derive a nonasymptotic concentration guarantee for the score matching estimator.

[2] Strictly speaking, our results do not handle unbounded costs, so they do not apply to the LQR problem.

[3] Kakade et al. [27] study kernelized version of this model, which they call the *kernelized nonlinear regulator*.

3.1 Background on score matching

Suppose we want to estimate the conditional density $\mathbb{P}(s'|s,a)$ of the form (2), given a dataset $\mathcal{D} = \{(s_t, a_t, s_{t+1})\}_{t\in[T]}$. MLE is the natural candidate for this estimation procedure, but it suffers from pitfalls. Solving for the MLE requires computing the log-partition function $Z_{sa}(\cdot)$. If the log-partition function is not known in closed form, it can be estimated using Markov Chain Monte Carlo methods [7, 10, 14]; however, this procedure may be computationally expensive. For some settings such as kernelized exponential families, MLE fails due to ill-posedness [21, 44].

Hyvärinen [22, 23] proposed score matching as an alternative to minimizing the log likelihood. Score matching minimizes the Fischer divergence, which is the expected squared distance between the score functions $\nabla_{s'} \log \mathbb{P}_W(s'|s,a)$. Specifically, we define the divergence between $\mathbb{P}_{W_0}$ and $\mathbb{P}_W$ for fixed (s,a) as:

$$J(\mathbb{P}_{W_0}(\cdot|s,a)\|\mathbb{P}_W(\cdot|s,a)) := \frac{1}{2}\int_{\mathcal{S}} \mathbb{P}_{W_0}(s'|s,a) \left\|\nabla_{s'} \log \tfrac{\mathbb{P}_{W_0}(s'|s,a)}{\mathbb{P}_W(s'|s,a)}\right\|^2 ds'. \tag{3}$$

Before proceeding with the exposition of the score matching estimator, we list standard regularity conditions that are required for the analysis of score matching [cf., 44, 4].

Assumption 1 (Regularity conditions)**.**

- **(A)** $\mathcal{S}$ is a non-empty open subset of $\mathbb{R}^{d_s}$ with piecewise smooth boundary $\partial\mathcal{S} := \overline{\mathcal{S}} - \mathcal{S}$, where $\overline{\mathcal{S}}$ is the closure of $\mathcal{S}$.
- **(B)** (Differentiability): $\psi(\cdot)$ is twice continuously differentiable on $\mathcal{S}$ with respect to each coordinate $i \in [d_s]$, and $\partial_i^j \psi(s)$ is continuously extensible to $\overline{\mathcal{S}}$ for all $j \in \{1,2\}, i \in [d_s]$.
- **(C)** (Boundary Condition): For all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $i \in [d_s]$, as $s' \to \partial\mathcal{S}$, we have:

$$\|\partial_i \psi(s')\|\, \mathbb{P}_{W_0}(s'|s,a) = o(\|s'\|^{1-d_s}).$$

- **(D)** (Integrability): For all $i \in [d_s]$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, let $p_{sa} := \mathbb{P}_{W_0}(\cdot|s,a)$. Then:

$$\|\partial_i \psi(s')\| \in L^2(\mathcal{S}, p_{sa}),\ \left\|\partial_i^2 \psi(s')\right\| \in L^1(\mathcal{S}, p_{sa}),\ \|\partial_i \psi(s')\|\, \partial_i \log q(s') \in L^1(\mathcal{S}, p_{sa}).$$

The key insight of Hyvärinen is that via an integration by parts trick, the divergence can be rewritten in a more amenable form. Essentially, these regularity conditions allow us to rewrite the conditional score function $J(W) := J(\mathbb{P}_{W_0}(\cdot|s,a)\|\mathbb{P}_W(\cdot|s,a))$ as:

$$J(W) = \frac{1}{2}\int_{\mathcal{S}} \mathbb{P}_{W_0}(s'|s,a)\cdot \sum_{i=1}^{d_s} \left[(\partial_i \log \mathbb{P}_W(s'|s,a))^2 + 2\partial_i^2 \log \mathbb{P}_W(s'|s,a)\right] ds' + C, \tag{4}$$

where C does not depend on the parameter W. In Appendix B.1 we provide a more formal derivation of (4) for exponential family densities as well as further discussion on Assumption 1.

Crucially, (4) *can be estimated with samples without requiring computation of the partition function*, since the partition function vanishes when taking partial derivatives with respect to s'. This gives rise to the following formulation of the **empirical score matching loss** for a dataset $\mathcal{D} = \{(s_t, a_t, s'_t)\}_{t\in[n]}$:

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

Furthermore, for any regularizer $\lambda > 0$, we can define the **empirical score matching estimator**:

$$\hat{W}_{n,\lambda} := \arg\min_W \hat{J}_n(W) + \tfrac{\lambda}{2}\|W\|_F^2. \tag{SM-E}$$

The following theorem gives a closed form expression for the empirical score matching estimator, when specialized to densities given by Definition 1.

Theorem 1. *For a dataset $\mathcal{D} = \{(s_t, a_t, s'_t)\}_{t\in[n]}$,* (SM-L) can be written as:

$$\hat{J}_n(W) = \frac{1}{2}\left\langle \mathsf{vec}\,(W)\,, \hat{V}_n \mathsf{vec}\,(W)\right\rangle + \left\langle \mathsf{vec}\,(W)\,, \hat{b}_n\right\rangle + C,$$

where:

$$\hat{V}_n := \sum_{t=1}^{n}\sum_{i=1}^{d_s} \mathsf{vec}\left(\partial_i\psi(s'_t)\phi(s_t,a_t)^\top\right)\mathsf{vec}\left(\partial_i\psi(s'_t)\phi(s_t,a_t)^\top\right)^\top \in \mathbb{R}^{d_\psi d_\phi \times d_\psi d_\phi},$$

$$\hat{b}_n := \mathsf{vec}\left(\sum_{t=1}^{n}\sum_{i=1}^{d_s}\left(\partial_i \log q(s'_t)\partial_i\psi(s'_t) + \partial_i^2\psi(s'_t)\right)\phi(s_t,a_t)^\top\right) \in \mathbb{R}^{d_\psi d_\phi},$$

and C does not depend on W. In addition, (SM-E) can be computed as:

$$\mathsf{vec}\left(\hat{W}_{n,\lambda}\right) = -(\hat{V}_n + \lambda I)^{-1}\hat{b}_n. \tag{5}$$

Theorem 1 is a typical result in score matching literature, and can be derived as a corollary of Arbel and Gretton [4, Thm. 3]. For completeness, we give a proof in Appendix B.2.

For the rest of the paper, it is useful to derive matrix expressions for $\hat{V}_n$ and $\hat{b}_n$. We define the following functions:

$$\Phi(s,a) := [E_{11}\phi(s,a), E_{12}\phi(s,a), \ldots E_{ij}\phi(s,a), \ldots E_{d_\psi\cdot d_\phi}\phi(s,a)]^\top \in \mathbb{R}^{d_\psi d_\phi\times d_\psi},$$

$$C(s') := \sum_{i=1}^{d_s}\partial_i\psi(s')\partial_i\psi(s')^\top \in \mathbb{R}^{d_\psi\times d_\psi}, \quad \xi(s') := \sum_{i=1}^{d_s}\partial_i \log q(s')\partial_i\psi(s') + \partial_i^2\psi(s') \in \mathbb{R}^{d_\psi}.$$

In addition, we use the subscript t to denote the value of the above expressions on sample (s_t, a_t, s'_t). We succintly represent $\hat{V}_n = \sum_{t=1}^n \Phi_t C_t \Phi_t^\top$ and $\hat{b}_n = \sum_{t=1}^n \Phi_t\xi_t$.

Computational efficiency. We make a few remarks on the computation of the score matching estimator. From Theorem 1, we see that computing $\hat{W}_n$ does not require estimation of the log-partition function Z_{sa}. The objective is a *quadratic* function in W, which we can solve for via Equation (5). However, Equation (5) requires us to invert a $d_\phi d_\psi \times d_\phi d_\psi$ matrix, which takes time $O(d_\phi^3 d_\psi^3)$ and memory $O(d_\phi^2 d_\psi^2)$. This can be disappointing from a practical perspective, where the dimensionality of ϕ and ψ can be large. Several additional considerations may remedy this. First, using the representer theorem, it is possible to show that $\hat{W}$ is the solution of a linear system of $n \cdot d_S$ variables, thus taking time $O(n^3 d_S^3)$ and space $O(n^2 d_S^2)$ [4, Thm. 1]. One can further reduce the dependence on n using Nyström approximations [46]. Second, if we are in the structured setting where $W_0 = \sum_{i=1}^d \theta_i A_i$, where $\theta \in \mathbb{R}^d$ is unknown but the matrices $A_i \in \mathbb{R}^{d_\psi \times d_\phi}$ are known. Theorem 1 can be adapted to this setting, and solving for $\hat{\theta}_n$ will take time $O(d^3)$ and space $O(d^2)$.

3.2 Concentration guarantee

We provide concentration guarantees for score matching under some structural assumptions:

Assumption 2 (Structural scaling)**.**

- **(A)** For any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $s' \sim \mathbb{P}_{W_0}(\cdot|s,a)$: we have $\xi(s')$ is B_ψ-subgaussian.
- **(B)** For any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $s' \sim \mathbb{P}_{W_0}(\cdot|s,a)$: we have $C(s')W_0\phi(s,a)$ is B_c-subgaussian.
- **(C)** For any $s' \in \mathcal{S}$: $\alpha_1 I \preceq C(s') \preceq \alpha_2 I$, where $\alpha_2 \geq \alpha_1 > 0$.
- **(D)** For any $(s,a) \in \mathcal{S}\times\mathcal{A}$: $\mathbb{E}_{sa}^{W_0}\psi(s')\psi(s')^\top - \mathbb{E}_{sa}^{W_0}\psi(s')\mathbb{E}_{sa}^{W_0}\psi(s')^\top \leq \kappa I$.

The conditions in Assumption 2 are mostly adapted from prior work [44, 4, 11], with suitable modifications to accomodate our non-i.i.d. setting. Notably, Assumption 2 holds for nonLDS (when $\Sigma = \sigma^2 I$) with $B_\psi = \sigma^{-6}, B_c = 0, \alpha_1 = \alpha_2 = \sigma^{-4}$ and $\kappa = \sigma^{-2}$. Due to space considerations, we defer further discussion on Assumption 2 to Appendix B.3.

We can prove the following concentration guarantee.

Theorem 2. *Suppose Assumptions 1 and 2 hold. Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration such that (s_t, a_t) is $\mathcal{F}_t$ measurable, s'_t is $\mathcal{F}_{t+1}$ measurable, and $s'_t \sim \mathbb{P}_{W_0}(\cdot|s_t, a_t)$.*

For any $\delta \in (0,1)$ and $\lambda > 0$, let:

$$\beta_n := \sqrt{\tfrac{2(B_\psi + B_c)}{\alpha_1^2}} \cdot \sqrt{\log \tfrac{\det(\lambda^{-1}\hat{V}_n + I)^{1/2}}{\delta}} + \sqrt{\lambda}\,\|W_0\|_F\,.$$

With probability at least $1-\delta$, the score matching estimators of (SM-E) satisfy:

$$\left\|\mathsf{vec}\left(\hat{W}_{n,\lambda}\right) - \mathsf{vec}\left(W_0\right)\right\|_{\hat{V}_n + \lambda I} \leq \beta_n, \quad \text{for all } n \in \mathbb{N}.$$

Theorem 2 is a *self-normalized* concentration guarantee, since the parameter error is rescaled by a data-dependent term $\hat{V}_n + \lambda I$. The proof is provided in Appendix B.4. The proof relies on the method of mixtures argument developed in the linear bandit literature [see, e.g., 2, 29].

4 Algorithm and main result

In this section, we present our main results, which introduce the Score Matching for RL (SMRL) algorithm (Algorithm 1) and provide regret guarantees.

4.1 Algorithm specification

Our algorithm works as follows. In each episode $k = 1, 2, \ldots, K$, we compute a elliptic confidence set $\mathcal{W}_k$ centered at our score matching estimator. In particular, we consider the $n := (k-1)H$ state transitions $\mathcal{D} = \{s_t, a_t, s'_t\}_{t=1}^{n}$ the agent has observed up until the beginning of episode k and run the score matching estimator to get the prediction $\hat{W}_k := \arg\min_W \hat{J}(W) + \frac{\lambda}{2}\|W\|_F^2$, via (Equation (5). In discussing our RL algorithm and its regret guarantees, we choose to index $\hat{W}$ and $\hat{V}$ by k rather than n to emphasize that these quantities are computed once per episode. We also drop the subscript λ because it is fixed across the run of the algorithm.

Let $B_\star$ is some known upper bound on $\|W_0\|_F$. We define the confidence set

$$\mathcal{W}_k := \left\{ W \in \mathbb{R}^{d_\psi \times d_\phi} : \left\|\mathsf{vec}\left(\hat{W}_k\right) - \mathsf{vec}\left(W\right)\right\|_{\hat{V}_k + \lambda I} \leq \beta_k \right\}, \tag{6}$$

where

$$\beta_k := \sqrt{\tfrac{2(B_\psi + B_c)}{\alpha_1^2}} \cdot \sqrt{\log \tfrac{2\det(\lambda^{-1}\hat{V}_k + I)^{1/2}}{\delta}} + \sqrt{\lambda}B_\star.$$

Once the agent computes the confidence set $\mathcal{W}_k$, they observe a new state s_1^k and compute an optimistic policy π^k (line 5-6), which is the optimal policy with respect to the "best model" in $\mathcal{W}_k$. As long as $W_0 \in \mathcal{W}_k$, the optimistic planning procedure gives us an overestimate of the true value function $V^\star_{\mathbb{P},1}(s_1^k)$, ensuring sufficient exploration of the MDP. Lastly, the agent runs policy π^k on the MDP to collect a new trajectory of data, which is added to the dataset $\mathcal{D}$.

4.2 Computational complexity

Algorithm 1 has two main components: model estimation (line 9) via score matching and optimistic planning (line 6). We have already discussed in Section 3 that the model estimation can be computed efficiently. Planning is a different story. Even planning with a *known model*, i.e., solving the problem $\pi^k = \arg\max_\pi V^\pi_{\mathbb{P}_W,1}(s_1^k)$, is already challenging without imposing further structure. However, it can be approximated with model predictive control [34, 49]. Furthermore, even with access to a planning oracle, *optimistic planning* is known to be NP-hard in the worst case [15]. In this work, we assume computational oracle access to the optimistic planner that solves (line 6) and leave developing efficient approximation algorithms to future work. One alternative to optimistic planning is to employ posterior sampling methods in conjunction with (approximate) planning oracles; the Bayesian regret can be theoretically analyzed using well-established techniques [e.g., 37, 11].

Algorithm 1 Score Matching for RL (SMRL)

Input: Regularizer λ and constants $B_\psi, B_c, B_\star, \kappa, \alpha_1$.
Initialize: starting confidence set $\mathcal{W}_1 = \mathbb{R}^{d_\psi \times d_\phi}$, confidence widths $\{\beta_k\}_{k\geq 1}$, dataset $\mathcal{D} = \emptyset$.
for episode $k = 1, 2, 3, \cdots, K$ **do**
  Planning:
  Observe initial state s_1^k
  Choose the optimistic policy: $\pi^k = \arg\max_\pi \max_{W \in \mathcal{W}_k} V^\pi_{\mathbb{P}_W, 1}(s_1^k)$
  Execution:
  Execute π^k to get a trajectory $\{s_h^k, a_h^k, r_h^k, s_{h+1}^k\}_{h\in[H]}$, and add it to $\mathcal{D}$.
  Solve for score matching estimator $\hat{W}_k = \arg\min_W \hat{J}(W) + \frac{\lambda}{2}\|W\|_F^2$ via (5)
  Update confidence set $\mathcal{W}_{k+1}$ via (6)

4.3 Regret guarantee

We now provide our main result, which is a $\sqrt{T}$-regret guarantee on the performance of SMRL.

Theorem 3 (SMRL Regret Guarantee)**.** *Suppose Assumptions 1 and 2 hold. Set $\lambda := 1/B_\star^2$ and fix $\delta \in (0, 1)$. Then with probability at least $1 - \delta$:*

$$\mathcal{R}(K) \leq C\sqrt{\gamma_{K+1} \cdot \left(\tfrac{\kappa(B_\psi + B_c)}{\alpha_1^3}\left(\gamma_{K+1} + \log {}^1\!/\!_\delta\right) + \tfrac{\kappa}{\alpha_1} + H\right)} \cdot \sqrt{H^2 T},$$

where $C > 0$ is an absolute constant and $\gamma_{K+1} := \log\det(\lambda^{-1}\hat{V}_{K+1} + I)$. If $\|\phi(s, a)\| \leq B_\phi$ for all (s, a), then $\mathcal{R}(K) \leq \tilde{O}(d_\psi d_\phi \cdot \sqrt{H^3 T})$, where the $\tilde{O}$ hides log factors and $\mathrm{poly}(\kappa, B_\psi, B_c, \alpha_1^{-1})$.

The proof is presented in Appendix C. A few remarks are in order. Our regret guarantee depends on the number of model parameters $d_\psi \cdot d_\phi$ and not on the state and action space sizes, thus making our algorithm sample-efficient in large-scale environments where $|\mathcal{S}|$ and $|\mathcal{A}|$ are infinite. Additionally, it is easy to redo the analysis when the parameter matrix is structured, i.e., $W_0 = \sum_{i=1}^d \theta_i A_i$, to see that the regret guarantee depends on d instead of $d_\psi \times d_\phi$. Thus, we can recover the same regret guarantee of $\tilde{O}(d\sqrt{H^3 T})$ that Chowdhury et al. provide.

On the more technical side, in Theorem 3, we require ϕ to be a bounded feature mapping, which linear dynamical systems do not satisfy in general (recall $\phi = [s, a]^\top$, and s, a can have unbounded norm). We need this to provide a bound on a certain "information gain" quantity $\gamma_k = \log\det(\lambda^{-1}\hat{V}_k + I)$ [cf., 43, 27]; however, the bounded ϕ assumption can be substantially weakened because our proof only requires $\sum_{h=1}^H \|\phi_h\|^2$ to be bounded in every episode with high probability. In particular, if one restricts to controllable policies which do not blow up norm of the state [e.g., 13], then the information gain term can be bounded.

5 Score matching vs maximum likelihood estimation

In this section, we provide a detailed comparison of score matching with maximum likelihood approaches. First we compare for exponential family transitions of Definition 1; then we specialize our comparison for the nonLDS setting. Lastly, we provide numerical evidence to demonstrate a setting where (a variant of) SMRL is superior.

5.1 General comparison for exponential family transitions

Score matching and MLE can be viewed as complementary techniques for density estimation; we highlight the relative pros and cons of SMRL vs Exp-UCRL.

In general, Exp-UCRL can be applied to more settings than score matching, due to the fact that score matching requires regularity conditions (Assumption 1) that are needed for the derivation of (4). In particular, we require $\mathcal{S}$ to be a Euclidean space and the feature vector $\psi : \mathcal{S} \to \mathbb{R}^{d_\psi}$ to be a twice-differentiable mapping. In this sense, the scope of SMRL is more limited than that of Exp-UCRL. For example, while tabular and factored MDPs can be modeled as exponential family transitions via the softmax parameterization,[4] we cannot prove regret guarantees for SMRL due to the differentiability requirement. Since the MLE estimator of Chowdhury et al. can be computed in $\mathrm{poly}(S, A)$ time, in the tabular and factored MDP settings we would prefer to run Exp-UCRL.

Among models given by Definition 1 where *both* score matching and MLE can be applied, score matching is preferred because the estimator can be computed in closed form as the solution to a ridge regression problem, and elliptic confidence sets can be constructed around it using Theorem 2. For the MLE, this is not possible in general. Chowdhury et al. implicitly define the estimator as the solution to the likelihood equations, and their confidence set is constructed in a complicated fashion, in terms of sums of KL divergences taken over the dataset. Thus, while we are unable to claim overall computational tractability of Algorithm 1 due to the computational difficulty of optimistic planning, score matching enables us to estimate model parameters efficiently, an improvement from Exp-UCRL.

We now compare the regret guarantee of Theorem 3 with previous results; the detailed calculations are deferred to Appendix D.1. We achieve the same order-wise guarantee as Chowdhury et al.(Thm. 2) of $\tilde{O}(d_\phi d_\psi \cdot \sqrt{H^3 T})$. In terms of problem constants, both bounds depend on $\sqrt{\kappa}$, but we (1) require the constants B_ψ and B_c, (2) replace dependence on strict convexity of the log partition function with the parameter α_1.

5.2 Comparison with prior work for nonLDS

Now we compare our results for SMRL with the results for Exp-UCRL (Chowdhury et al.) and LC^3 (Kakade et al.) for the nonLDS problem with bounded and known rewards. For simplicity we will take the transition noise to be $\mathcal{N}(0, \sigma^2 I_{d_s})$. We will also assume that $\|W_0\|_F \leq B_\star$ and that the feature vectors are bounded as $\|\phi(s,a)\| \leq B_\phi$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. All three are similar UCRL-style algorithms, and we compare the parameter estimation, confidence sets, and regret guarantee.

Estimation and confidence set construction. For nonLDS, score matching and MLE are equivalent estimators (see Proposition 10 for a formal statement). Thus, in all three algorithms, the parameter estimation *procedure* is identical, up to rescaling of regularization parameter λ. To further facilitate comparison, we will hereafter fix the λ of each algorithm such that the parameter estimation is the same as LC^3 (for any fixed dataset). Our choices are detailed in Appendix D.2.

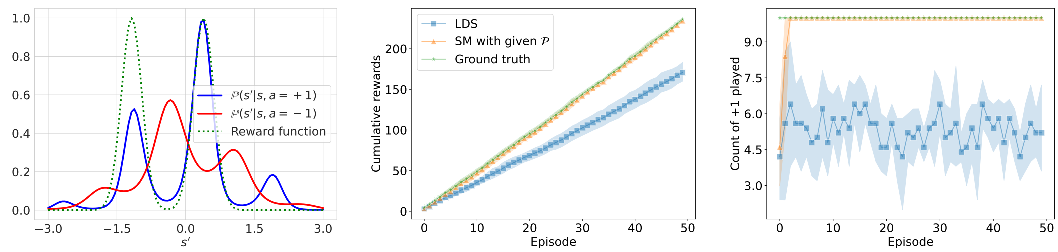
Once we have fixed the parameter λ for each algorithm, the main distinction lies in the confidence set construction. While all three algorithms essentially utilize the same optimistic planning procedure, optimistic planning depends on the confidence sets constructed in each episode. The chosen policies and the resulting trajectories will be different in all three algorithms. The confidence sets constructed for each paper are essentially the tightest self-normalized bound one can prove, so it is hard to directly compare the confidence sets from paper to paper due to the difference in analyses. Generally speaking, SMRL uses Frobenius norm bounds (Theorem 2), Exp-UCRL uses a mixture of both Frobenius and spectral [11, Sec. 3.1], and LC^3 uses only spectral norm bounds [27, Eq. 3.2].

Regret guarantee. In terms of the regret guarantee, Theorem 3 gives us a regret guarantee of $\tilde{O}\big(\sqrt{d_\phi d_\psi \cdot (\sigma^4 d_\phi d_\psi + H) H^2 T}\big)$, while a bound of $\tilde{O}\big(\sqrt{d_\phi^2 d_\psi^2 (1 + \sigma^{-2} B_\star^2 B_\phi^2 H) H^2 T}\big)$ can be derived for Exp-UCRL. Note that the latter bound depends polynomially on the scale of W_0 and ϕ. Kakade et al. (Remark 3.5) give a bound for LC^3 of $\tilde{O}(\sqrt{d_\phi (d_\phi + d_\psi + H) H^2 T})$, without polynomial dependence on σ^2 and the scale of W_0 and ϕ. We conjecture that the σ^2 dependence is an artifact of our analysis, but it is less clear whether the dependence on d_ϕ, d_ψ can be improved.

5.3 Experiments on synthetic MDP

We demonstrate end-to-end benefits of using score matching in a (highly stylized) synthetic MDP; see Figure 1. In our constructed MDP, the transition function is multimodal; the action choice affects the location of the modes of the next state density. The reward is constructed so that $a = +1$ leads to higher reward than $a = -1$ at most states. To enable fair comparison, we *fix* a simple random sampling shooting planner [39] and evaluate three model estimation procedures: score matching with

[4]There is a mild technical issue, since Definition 1 cannot capture transitions with probability 0, so we must assume that the support of the transitions is known in advance. See the paper [11] for more details.



(a) MDP transition and reward (b) Cumulative rewards (c) Planner action choices

Figure 1: Comparing SM vs fitting an LDS for a synthetic MDP, with $\mathcal{S} = \mathbb{R}$, $\mathcal{A} = \{+1, -1\}$, $H = 10$, initial state distribution $\text{Unif}([-1, +1])$, $\mathbb{P}(s'|s, a) = \exp(-s'^{1.7}/1.7) \cdot \exp(\sin(4s')(s + a))$, and $r(s, a) = \exp(-10(s - \pi/8)^2) + \exp(-10(s + 3\pi/8)^2)$. (a) plots $\mathbb{P}$ for a single starting state $s = 0.5$ for $a = +1$ and $a = -1$; the reward r is superimposed. Taking $a = +1$ is more likely to transition to states with high reward. (b) plots cumulative reward for fixed planner with varying model estimation: SM with the given $\mathcal{P}$, fitting an LDS, and a baseline with the ground truth model. (c) plots the number of steps in every episode where $a = +1$ is picked by the planner. In (b) and (c), shaded areas correspond to 95% confidence intervals.

the given class $\mathcal{P}$, fitting an LDS via MLE, and a baseline where planner is supplied the ground truth $\mathbb{P}$. (For this simple one-dimensional RL task, one can also numerically compute the MLE with the given $\mathcal{P}$. However, this approach does not scale to RL tasks with high-dimensional states.) Fitting an LDS does poorly because the LDS density is not expressive enough to differentiate between $a = +1$ and $a = -1$, while score matching estimates the density well, so the planner quickly learns to pick $a = +1$. Our experiments suggest that modeling the transition $\mathbb{P}$ via the richer Definition 1 can yield end-to-end benefits for RL tasks. Further experimental details can be found in Appendix F.

6 Conclusion

In this paper, we show $\sqrt{T}$-regret guarantees for a reinforcement learning setting when the state transition model is an exponential family model, a challenging nonlinear setting. Under this modeling assumption, the commonly employed MLE may be intractable; we bypass such issues by proposing to learn the model via the score matching method.

We conclude with a few possible directions for future work.

- *Model Misspecification:* Proving guarantees for SMRL when the underlying transition $\mathbb{P}$ do not lie in the model class $\mathcal{P}$ but instead is well-approximated by $\tilde{\mathbb{P}} \in \mathcal{P}$ is an interesting direction.
- *Arbitrary State Spaces:* A key limitation of the score matching estimator is that it requires that the state space $\mathcal{S}$ must be a subset of the Euclidean space $\mathbb{R}^{d_s}$ and the feature mapping ψ to be twice differentiable; therefore it cannot handle arbitrary state spaces. One important direction is extending the score matching algorithm to discrete state spaces such as tabular/factored MDPs through a suitable modification of the estimation procedure [e.g., 23, 31].
- *Kernelization:* We would like to extend our guarantees to the *kernel conditional exponential family* (KCEF) setting of Arbel and Gretton [4], i.e., when the conditional model is $\mathbb{P}_f(s'|s, a) := q(s') \cdot \exp(\langle f, \Gamma_{sa} k(s', \cdot)\rangle - Z_{sa}(f))$, where f lies in some vector valued Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, $k(s', \cdot)$ lies in an RKHS $\mathcal{H}_S$, and $\Gamma_{sa} : \mathcal{H}_S \to \mathcal{H}$ is an operator that depends on (s, a). This generalizes our finite dimensional setting and is a special case of the conditional family where the inner product is $\langle f, \phi(s, a, s')\rangle$, studied by Canu and Smola [9]. In the KCEF setting, MLE becomes computationally intractable; yet score matching can be kernelized and efficiently computed, and fast approximation methods exist [46]. Our theory does not hold for the KCEF because the parameter $\alpha_1 = 0$. Instead, one might be able to adapt the range-space assumption from the paper [4] to the non-i.i.d. setting.

Acknowledgments and Disclosure of Funding

This work is supported by funding from the Institute for Data, Econometrics, Algorithms, and Learning (IDEAL). We thank Pritish Kamath, Danica J. Sutherland, Akshay Krishnamurthy, and Wen Sun for helpful discussions. Part of this work was done while GL, ZW, and ZY were participating in the Simons Program on the Theoretical Foundations of Reinforcement Learning in Fall 2020.

References

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

[2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320, 2011.

[3] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

[4] Michael Arbel and Arthur Gretton. Kernel conditional exponential family. In *International Conference on Artificial Intelligence and Statistics*, pages 1337–1346, 2018.

[5] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[6] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[7] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[8] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[9] Stéphane Canu and Alex Smola. Kernel methods and the exponential family. *Neurocomputing*, 69(7-9):714–720, 2006.

[10] Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pages 33–40. PMLR, 2005.

[11] 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.

[12] Adam D Cobb, Atılım Güneş Baydin, Andrew Markham, and Stephen J Roberts. Introducing an explicit symplectic integration scheme for riemannian manifold hamiltonian monte carlo. *arXiv preprint arXiv:1910.06243*, 2019.

[13] Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only sqrtt regret. In *International Conference on Machine Learning*, pages 1300–1309. PMLR, 2019.

[14] Bo Dai, Zhen Liu, Hanjun Dai, Niao He, Arthur Gretton, Le Song, and Dale Schuurmans. Exponential family estimation via adversarial dynamics embedding. *arXiv preprint arXiv:1904.12083*, 2019.

[15] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback.

[16] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

[17] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, 2018.

[18] Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In *Conference on Learning Theory*, pages 1554–1557. PMLR, 2020.

[19] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

[20] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[21] Kenji Fukumizu. Exponential manifold by reproducing kernel hilbert spaces. *Algebraic and Geometric mothods in statistics*, pages 291–306, 2009.

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

[23] Aapo Hyvärinen. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

[24] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[25] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[26] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

[27] 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.

[28] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[29] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[30] Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

[31] Siwei Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2009.

[32] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, 2019.

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

[34] David Q Mayne. Model predictive control: Recent developments and future promise. *Automatica*, 50(12):2967–2986, 2014.

[35] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[36] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[37] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

[38] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386, 2016.

[39] Anil V Rao. A survey of numerical methods for optimal control. *Advances in the Astronautical Sciences*, 135(1):497–528, 2009.

[40] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *NIPS*, pages 2256–2264, 2013.

[41] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[42] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

[43] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[44] Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *The Journal of Machine Learning Research*, 18(1):1830–1888, 2017.

[45] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933. PMLR, 2019.

[46] Danica J. Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with nyström kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pages 652–660. PMLR, 2018.

[47] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[48] Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

[49] Nolan Wagener, Ching-An Cheng, Jacob Sacks, and Byron Boots. An online learning approach to model predictive control. *arXiv preprint arXiv:1902.08967*, 2019.

[50] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

[51] Lin F Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019.

[52] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work is mostly theoretical, and does not have any immediate negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Experiments ran on laptop.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Related work

In this section we discuss several related works on provably efficient reinforcement learning.

In the tabular setting, our theoretical understanding of RL is fairly complete [24, 38, 6, 16, 3]. The best possible regret is $\tilde{\Theta}(\sqrt{H^2SAT})$ [24, 6], where S denotes the state space size and A denotes the action space size. The minimax lower bound shows that unless stronger assumptions are placed on the MDP, we must incur regret that scales as $\Omega(\sqrt{SA})$, which can be exponentially large in real-world problems.

There has been considerable theoretical effort in understanding RL with specific model-based or model-free assumptions. Prior work often posits linearity assumptions in order to design algorithms which replace the dependence on S, A with some notion of intrinsic dimensionality [51, 26, 8, 30, 52, 36]. In particular, Yang and Wang [51] consider a bilinear transition model; in comparison, our transition model given by Definition 1 is *log-bilinear*. In nonlinear settings, a line of work establishes regret guarantees or system identification guarantees for the LQR problem with unknown dynamics [1, 17, 32, 13, 42] as well as its nonlinear generalization [33, 27]. The transitions in the LQR are *nonlinear*, but the corresponding value functions are linear in an appropriate basis. Our work studies a generalization of the transition dynamics in the aforementioned works; both the transition dynamics and the value functions can be nonlinear.

Another line of work studies RL with general function approximation. Some authors study model-free RL by placing structural assumptions on the value function, i.e., assuming bounded Bellman rank [25, 18] or bounded eluder dimension [50]. While in principle it is possible to apply model-free algorithms with general function approximation to our setting, the induced class of value functions generated by our assumption can be complex. More relevant to our work are model-based approaches [45, 37, 5]. In contrast to our work, these papers do not focus on the computational tractability of model estimation. Sun et al. [45] prove PAC-learning guarantees for model-based RL by relying on a complexity measure called the *witness rank*. They analyze exponential family models under the assumption that the Hessian of the log-partition function has lower and upper-bounded eigenvalues. Our setting can be viewed as a special case of theirs with linear test functions $\mathcal{F} = \{(s, a, s') \mapsto \langle W\phi(s, a), \psi(s')\rangle : W \in \mathbb{R}^{d_\psi \times d_\phi}\}$. Their results are not directly comparable to ours, since their final bound depends on the log cardinality of the model class $\mathcal{M}$ and the test function class $\mathcal{F}$, as well as a linear dependence on the number of actions; in contrast, our bound scales with the dimensionality $d_\psi \cdot d_\phi$, with no dependence on the number of actions. Osband and Van Roy [37] prove Bayesian regret guarantees for posterior sampling in terms of the *eluder dimension* [40] of the model class; their results are limited due to requirements of global Lipschitzness on the future value functions and subgaussianity of the transitions. Ayoub et al. [5] introduce a model-based algorithm which learns a general class of transition models $\mathcal{P}$ via a technique called *value-targeted regression*. Their regret guarantees depend on the eluder dimension of a constructed Q function class $\mathcal{F}_{\mathcal{P}}$. Lastly, Foster et al. [20] introduce a *Decision Estimation Coefficient* and show that it provides upper and lower bounds for interactive decision making.

B Score matching details

In this section, we provide details on the score matching estimator and prove Theorem 1 and Theorem 2. We use results from prior work [22, 44, 4] and defer further discussion to those papers.

B.1 Background on score matching

We recall some notation.

The model we consider under Definition 1 is:

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

where

$$W_0 \in \mathcal{W} := \left\{ W \in \mathbb{R}^{d_\psi \times d_\phi} : \int_{\mathcal{S}} q(s') \exp(\langle \psi(s'), W\phi(s, a)\rangle)\, ds' < \infty,\ \forall (s, a) \in \mathcal{S} \times \mathcal{A} \right\}.$$

Let us define the population score matching loss for a fixed $(s, a) \in \mathcal{S} \times \mathcal{A}$ pair as:

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

Hereafter, we will assume standard regularity conditions (Assumption 1) which enable us to employ integration by parts to simplify the score matching objective, [e.g., 4, Appendix A.4]. We pause to make a few remarks about Assumption 1 and compare to previous work [4].

- It is fair to note that these regularity assumptions limit the applicability of score matching. Several exponential family densities whose parameters can be estimated via MLE cannot be estimated via score matching. For example, exponential distributions violate the boundary condition because the pdf does not decay to 0 as we approach 0. Other distributions like the gamma distribution violate the integrability conditions. For *nonnegative distributions*, follow up work by Hyvärinen [23] provides a modified score matching estimator, which seems to work well empirically for density estimation of nonnegative gaussians; to the best of our knowledge, finite sample concentration guarantees for this estimator do not exist.
- Our distributional assumption is a special case of the kernel conditional exponential family introduced by Arbel and Gretton [4]; so everywhere they have $k(s', \cdot)$ we replace it with $\psi(s')$.
- We generally work with the conditional versions of the score matching loss J_{sa}, while Arbel and Gretton [4] work in the i.i.d. setting, where they assume the existence of a joint distribution $p(x, y) = \pi(x) \cdot p(y|x)$. Thus, their theorems are stated for the averaged loss $J(W) := \int_{\mathcal{S}} J_x(W)\pi(x)$.

In addition, we will define the following quantities, which can be viewed as the population versions of various quantities which appear in Theorem 1:

$$\begin{aligned}
\overline{V}_{sa} &:= \mathbb{E}_{sa}^{W_0} \left[\sum_{i=1}^{d_s} \mathsf{vec}\left(\partial_i \psi(s') \phi(s,a)^\top\right) \mathsf{vec}\left(\partial_i \psi(s') \phi(s,a)^\top\right)^\top \right] \in \mathbb{R}^{d_\psi d_\phi \times d_\psi d_\phi}, \\
\overline{\xi}_{sa} &:= \mathbb{E}_{sa}^{W_0} \left[\sum_{i=1}^{d_s} \partial_i \log q(s'_t) \partial_i \psi(s'_t) + \partial_i^2 \psi(s'_t) \right] \in \mathbb{R}^{d_\psi d_\phi}, \\
\overline{C}_{sa} &:= \mathbb{E}_{sa}^{W_0} \left[\partial_i \psi(s'_t) \partial_i \psi(s'_t)^\top \right] \in \mathbb{R}^{d_\psi \times d_\psi}.
\end{aligned}$$

Now we introduce the main theorem for score matching, a version of which was shown in Arbel and Gretton [4, Theorem 3].

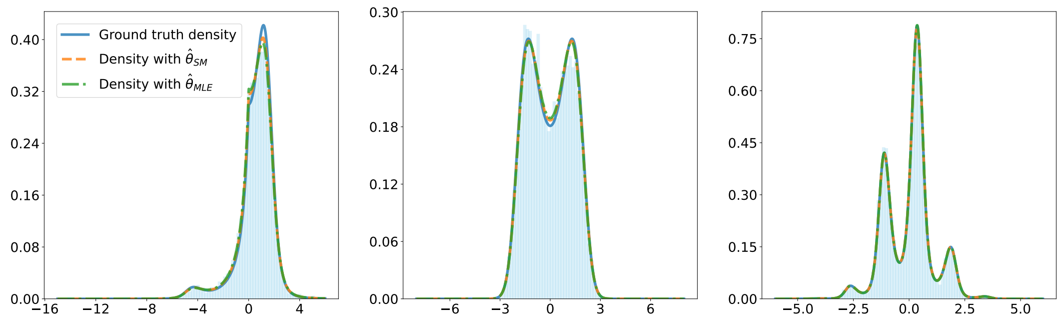
Theorem 4. *Under Assumption 1, the following are true for all* $(s, a) \in \mathcal{S} \times \mathcal{A}$.

1. $J_{sa}(W) < \infty$ *for all* $W \in \mathcal{W}$.
2. *For all* $W \in \mathcal{W}$, $J_{sa}(W) = \frac{1}{2} \left\langle \mathsf{vec}\left(W - W_0\right), \overline{V}_{sa} \mathsf{vec}\left(W - W_0\right) \right\rangle$.
3. *Alternatively:*

$$J_{sa}(W) = \frac{1}{2} \left\langle \mathsf{vec}\left(W\right), \overline{V}_{sa} \mathsf{vec}\left(W\right) \right\rangle + \left\langle \mathsf{vec}\left(W\right), \overline{\xi}_{sa} \right\rangle + J(\mathbb{P}_{W_0}(\cdot|s,a) \| q(\cdot)),$$

4. $\overline{\xi}_{sa} = -\overline{C}_{sa} W \phi(s, a)$.

We do not formally prove this because all that has changed is some of the notation from Arbel and Gretton [4], as well as specializing their results to the finite-dimensional setting. In order to derive the alternative expression in Theorem 4, part 3, we must use the integration by parts trick which was first introduced by Hyvärinen [22] and requires the boundary condition **(C)** and integrability conditions **(D)**. Theorem 4, part 3 is useful as it gives us an expression for the score matching estimator that does not require us to estimate the log partition function; as we show in Theorem 1, the empirical loss is simply a quadratic function in W.



(a) $\alpha = 1, \psi(s) = \exp(\sin(s))$ (b) $\alpha = 2, \psi(s) = s\sin(s)$ (c) $\alpha = 1.7, \psi(s) = \sin(4s)$

Figure 2: Performance of score matching vs. MLE for three unconditional 1D densities. We consider densities $q(s) \cdot \exp(\psi(s') \cdot \theta_0 - Z(\theta))$, where $q(s) = \exp(-s^\alpha/\alpha)$ for some $\alpha > 0$, $\psi(\cdot)$ is allowed to vary, and the parameter is set as $\theta_0 = 1$. For each density we sample 10^4 points with Hamiltonian Monte Carlo, using the package `hamiltorch` [12]. The MLE does not have a closed form and is approximated numerically. All densities violate Assumption 2(C), and possibly other assumptions.

B.2 Proof of Theorem 1

For notational convenience, we will denote $\phi_t := \phi(s_t, a_t)$ and $\psi_t := \psi(s'_t)$. We write the score matching loss as:

$$\begin{aligned}\hat{J}_n(W) &:= \frac{1}{2}\sum_{t=1}^{n}\sum_{i=1}^{d_s}\left\{(\partial_i \log \mathbb{P}_W(s'_t|s_t,a_t))^2 + 2\partial_i^2 \log \mathbb{P}_W(s'_t|s_t,a_t)\right\} \\ &= \frac{1}{2}\sum_{t=1}^{n}\sum_{i=1}^{d_s}\left\{\left(\partial_i \log q(s'_t) + \partial_i \psi_t^\top W \phi_t\right)^2 + 2\left(\partial_i^2 \log q(s'_t) + \partial_i^2 \psi_t^\top W\phi_t\right)\right\} \\ &= \frac{1}{2}\sum_{t=1}^{n}\sum_{i=1}^{d_s}\left(\partial_i \psi_t^\top W\phi_t\right)^2 + \sum_{t=1}^{n}\sum_{i=1}^{d_s}\left(\partial_i \log q(s'_t)\partial_i\psi_t^\top W\phi_t + \partial_i^2\psi_t^\top W\phi_t\right) + C,\end{aligned}$$

where C only contains terms that depend on $\partial_i q(s')$ and $\partial_i^2 q(s')$. The first part follows after using the trace trick identities:

$$\begin{aligned}a^\top M b &= \mathsf{Tr}(Mba^\top) = \left\langle \mathsf{vec}\,(M)\,, \mathsf{vec}\,\left(ab^\top\right)\right\rangle \\ (a^\top M b)^2 &= \left\langle \mathsf{vec}\,(M)\,, \mathsf{vec}\,\left(ab^\top\right)\mathsf{vec}\,\left(ab^\top\right)^\top \mathsf{vec}\,(M)\right\rangle,\end{aligned}$$

and applying linearity to move the sums inside the inner product operators.

The second part is simply the standard form of ridge regression estimator.

B.3 Discussion of Assumption 2

We give some additional background on Assumption 2. Roughly speaking, one can view the Φ_t as the covariates and ξ_t as the response. The equation Equation (SM-E) bears strong resemblance to the ridge regression solution from the standard least squares setup; however, the main difference is that the "covariance" matrix $\hat{V}_n$ includes an additional term C_t which captures the curvature of the ψ mapping. **(A)** and **(B)**: are needed to control the regression error term $\hat{b}_n + \hat{V}_n \mathsf{vec}(W_0) = \sum_t \Phi_t(\xi_t + C_t W_0 \phi_t)$. **(C)**: This assumption is a strengthening of a certain assumption in the i.i.d. setting [44], which assumes that $\mathbb{E}\left[C(s')\right] \geq \alpha_1 I$. In the non-i.i.d. setting there is no fixed distribution that we draw s' from, so we replace it with an almost sure bound $C(s') \succeq \alpha_1 I$. On a technical level, **(C)** is required to change the matrix norm in the proof of Theorem 2. **(D)**: this assumption essentially states that covariance of ψ is bounded when drawn from the conditional distribution $\mathbb{P}(\cdot|s,a)$. It is used to analyze the regret guarantee to bound the KL divergence of the estimated density and the ground truth in terms of the parameter estimation error, and is also made in Chowdhury et al. [11].

Experimentally, we find that a much richer class of densities can be estimated via score matching than required by Assumption 2, suggesting that Theorem 2 can be shown under weaker conditions (see Figure 2); we leave this to future work.

B.4 Proof of Theorem 2

Fix any $n \in \mathbb{N}$. Using the definition of $\hat{W}_{n,\lambda}$ from (5), we can write:

$$\begin{aligned}\left\|\mathsf{vec}\left(\hat{W}_{n,\lambda}\right) - \mathsf{vec}\left(W_0\right)\right\|_{\hat{V}_n+\lambda I} &\leq \left\|\hat{b}_n + \hat{V}_n \mathsf{vec}\left(W_0\right)\right\|_{(\hat{V}_n+\lambda I)^{-1}} + \lambda \left\|\mathsf{vec}\left(W_0\right)\right\|_{(\hat{V}_n+\lambda I)^{-1}} \\ &\leq \left\|\hat{b}_n + \hat{V}_n \mathsf{vec}\left(W_0\right)\right\|_{(\hat{V}_n+\lambda I)^{-1}} + \sqrt{\lambda} \left\|W_0\right\|_F, \end{aligned} \tag{7}$$

where the first inequality uses triangle inequality and the second inequality uses the fact that $\hat{V}_n + \lambda I \succeq \lambda I$.

It now remains to bound the first term. We can write the first term as:

$$\hat{b}_n + \hat{V}_n \mathsf{vec}\left(W_0\right) = \sum_{t=1}^{n} \Phi_t \xi_t + \Phi_t C_t \Phi_t^\top \mathsf{vec}\left(W_0\right) = \sum_{t=1}^{n} \Phi_t \underbrace{\left(\xi_t + C_t W_0 \phi_t\right)}_{=:\Delta_t}.$$

From here, we have:

$$\begin{aligned}\left\|\hat{b}_n + \hat{V}_n \mathsf{vec}\left(W_0\right)\right\|_{(\hat{V}_n+\lambda I)^{-1}} &= \left\|\sum_{t=1}^{n} \Phi_t \Delta_t\right\|_{(\sum_{t=1}^n \Phi_t C_t \Phi_t^\top + \lambda I)^{-1}} \\ &\leq \alpha_1^{-1} \left\|\sum_{t=1}^{n} \Phi_t \Delta_t\right\|_{(\sum_{t=1}^n \Phi_t \Phi_t^\top + \alpha^{-1}\lambda I)^{-1}}, \end{aligned} \tag{8}$$

using Assumption 2(iii).

Next we show that Δ_t is conditionally subgaussian with parameter $B_\psi + B_c$. Then by Theorem 4 (iv), $\{\Delta_t\}_{t=1}^\infty$ is an $\mathcal{F}_{t+1}$-adapted martingale difference sequence, since:

$$\mathbb{E}\left[\Delta_t \mid \mathcal{F}_t\right] = \mathbb{E}\left[\sum_{i=1}^{d_s} \partial_i \log q(s_t')\partial_i \psi(s_t') + \partial_i^2 \psi(s_t') + \partial_i \psi(s_t')\partial_i \psi(s_t')^\top W \phi(s_t, a_t) \;\middle|\; s_t, a_t\right] = 0.$$

Furthermore, using Assumption 2(i) and (ii) and applying Lemma 13, we see that Δ_t is a conditionally subgaussian random vector with parameter $B_\psi + B_c$. By Lemma 11, with probability at least $1-\delta$, we get that for all $n \in \mathbb{N}$:

$$\begin{aligned}\left\|\sum_{t=1}^{n} \Phi_t \Delta_t\right\|^2_{(\sum_{t=1}^n \Phi_t \Phi_t^\top + \alpha_1^{-1}\lambda I)^{-1}} &\leq 2(B_\psi + B_c) \log \frac{\det(\frac{\alpha_1}{\lambda}\sum_{t=1}^n \Phi_t \Phi_t^\top + I)^{1/2}}{\delta} \\ &\leq 2(B_\psi + B_c) \log \frac{\det(\lambda^{-1}\hat{V}_n + I)^{1/2}}{\delta}. \end{aligned} \tag{9}$$

The proof concludes by combining Eqs. 7, 8, and 9.

C Regret guarantee proof

In this section, we prove our main regret guarantee, Theorem 3.

Notation. For states and actions, we let s_h^k and a_h^k denote the state (action) that the agent observes (plays) in step h of episode k. We let $\phi_h^k := \phi(s_h^k, a_h^k)$ and $\psi_h^k := \psi(s_{h+1}^k)$. In addition, we denote Φ_{kh} and C_{kh} for the matrix expressions of Φ, C for the data point $(s_h^k, a_h^k, s_{h+1}^k)$.

For value functions, we will generally write $V_{\mathbb{P},h}^{\pi}$ to denote the value function of running policy π in the MDP with transition model $\mathbb{P}$. Since we consider parameterized transition models, sometimes we will replace $\mathbb{P}$ with the parameter W. For the superscript, we adopt the following conventions: (1) we always denote $\pi^\star$ to be the optimal policy *under the ground truth model* $\mathbb{P}_{W_0}$, and sometimes denote it as $\star$, (2) in episode k with agent policy π^k, we replace π^k with k. For example, we will read $V^k_{\hat{W},h}$ to be "the value function of running policy π^k on MDP parameterized by $\hat{W}$, at step h".

We will also define the natural filtration $\mathcal{F} = \{\mathcal{F}^k_h\}_{h\geq 1, k\geq 1}$, where:

$$\mathcal{F}^k_h := \sigma\left(\bigcup_{i\in[H],\, j\in[k-1]} \{s^j_i, a^j_i\} \,\cup\, \bigcup_{i\in[h],\, j=k} \{s^j_i, a^j_i\}\right),$$

representing all state-action pairs up to time (k, h).

C.1 Preliminary lemmas

We first introduce the auxiliary lemmas which we will use in our proof. First we introduce a recursive lemma that allows us to upper bound the regret.

Lemma 5 (Recursive lemma)**.** *Let $\tilde{W}_k = \arg\max_{W\in\mathcal{W}_k} V^k_{W,1}(s^k_1)$. Then:*

$$\sum_{k=1}^K \left(V^k_{\tilde{W}_k,1}(s^k_1) - V^k_{W_0,1}(s^k_1)\right) = \sum_{k=1}^K\sum_{h=1}^H \left[\mathbb{E}^{\tilde{W}_k}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s') - \mathbb{E}^{W_0}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s')\right] + \sum_{k=1}^K\sum_{h=1}^H m^k_h,$$

where $m^k_h = \mathbb{E}^{W_0}_{s^k_h a^k_h}\left[V^k_{\tilde{W}_k,h+1}(s') - V^k_{W_0,h+1}(s')\right] - \left(V^k_{\tilde{W}_k,h+1}(s^k_{h+1}) - V^k_{W_0,h+1}(s^k_{h+1})\right)$ is a martingale difference sequence adapted to $\mathcal{F}$ satisfying $|m^k_h| \leq 2H$.

Proof. For every $(k,h) \in [K]\times[H]$, we have

$$\begin{aligned}
&V^k_{\tilde{W}_k,h}(s^k_h) - V^k_{W_0,h}(s^k_h)\\
&= r(s^k_h, a^k_h) + \mathbb{E}^{\tilde{W}_k}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s') - r(s^k_h, a^k_h) - \mathbb{E}^{W_0}_{s^k_h a^k_h} V^k_{W_0,h+1}(s')\\
&= \mathbb{E}^{\tilde{W}_k}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s') - \mathbb{E}^{W_0}_{s^k_h a^k_h} V^k_{W_0,h+1}(s')\\
&= V^k_{\tilde{W}_k,h+1}(s^k_{h+1}) - V^k_{W_0,h+1}(s^k_{h+1}) + \left(\mathbb{E}^{\tilde{W}_k}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s') - \mathbb{E}^{W_0}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s')\right)\\
&\quad \underbrace{- \left(V^k_{\tilde{W}_k,h+1}(s^k_{h+1}) - V^k_{W_0,h+1}(s^k_{h+1})\right) + \mathbb{E}^{W_0}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s') - \mathbb{E}^{W_0}_{s^k_h a^k_h} V^k_{W_0,h+1}(s')}_{=:m^k_h},
\end{aligned}$$

where in the first equality we used the definition of the value function and that $a^k_h = \pi^k(s^k_h)$. Note that because $s^k_{h+1} \sim \mathbb{P}_{W_0}(\cdot|s^k_h, a^k_h)$, m^k_h is zero mean conditioned on $\mathcal{F}^k_h$, so the martingale difference sequence property holds. Also, since $V \in [0, H]$, the a.s. bound holds.

Using this formula recursively we can obtain the lemma. $\square$

The next lemma allows us to convert a bound on KL divergence to a self-normalized bound on the parameter error; a version of it is also shown in Chowdhury et al. [11].

Lemma 6 (KL Divergence Bound)**.** *Under Definition 1 and Assumption 2, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $W, W' \in \mathbb{R}^{d_\psi\times d_\phi}$, it holds that*

$$D_{\mathrm{KL}}\left(\mathbb{P}_W(\cdot|s,a)\|\mathbb{P}_{W'}(\cdot|s,a)\right) \leq \frac{\kappa}{2}\left\|\mathsf{vec}\left(W\right) - \mathsf{vec}\left(W'\right)\right\|^2_{\Phi_{s,a}\Phi^\top_{s,a}}.$$

Proof. We have

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbb{P}_W(\cdot|s,a)\|\mathbb{P}_{W'}(\cdot|s,a)\right) &= \int_{\mathcal{S}} \mathbb{P}_W(s'|s,a)\log\frac{\mathbb{P}_W(s'|s,a)}{\mathbb{P}_{W'}(s'|s,a)}\,ds'\\
&= \mathbb{E}^W_{sa}\left[\psi(s')^\top(W-W')\phi(s,a) - Z_{sa}(W) + Z_{sa}(W')\right].
\end{aligned}$$

By Taylor expansion, there exists some $\tilde{W}$ which lies between W and W' that satisfies:

$$\begin{aligned}Z_{sa}(W') = Z_{sa}(W) + \nabla_W Z_{sa}(W)^\top \mathsf{vec}\,(W' - W)\\ + \frac{1}{2}\mathsf{vec}\,(W'-W)^\top \nabla_W^2 Z_{sa}(\tilde{W})\mathsf{vec}\,(W'-W)\,,\end{aligned}$$

where $\nabla Z_{sa}(\cdot)$ and $\nabla^2 Z_{sa}(\cdot)$ are understood to be $\mathbb{R}^{d_\psi d_\phi}$ and $\mathbb{R}^{d_\psi d_\phi \times d_\psi d_\phi}$ respectively, representing the gradient and Hessian of the function $Z_{sa} : R^{d_\psi d_\phi} \to \mathbb{R}$. Lemma 12 gives us expressions for $\nabla Z_{sa}(\cdot)$ and $\nabla^2 Z_{sa}(\cdot)$, which we can plug in to get:

$$\begin{aligned}D_{\mathrm{KL}}\left(\mathbb{P}_W(\cdot|s,a)\|\mathbb{P}_{W'}(\cdot|s,a)\right) &\le \frac{1}{2}\mathsf{vec}\,(W'-W)^\top \nabla_W^2 Z_{sa}(\tilde{W})\mathsf{vec}\,(W'-W)\\ &\le \frac{\kappa}{2}\left\|\mathsf{vec}\,(W) - \mathsf{vec}\,(W')\right\|^2_{\Phi_{s,a}\Phi_{s,a}^\top}\,,\end{aligned}$$

using Assumption 2 (iv) in the last step. This proves the theorem. □

Lemma 7. *For any sequence of* $\{(\Phi_{kh}, C_{kh})\}_{k\in[K],h\in[H]}$*, we have:*

$$\sum_{k=1}^{K}\min\left\{\sum_{h=1}^{H}\left\|(\hat{V}_k+\lambda I)^{-1/2}\Phi_{kh}C_{kh}\Phi_{kh}^\top(\hat{V}_k+\lambda I)^{-1/2}\right\|, 1\right\} \le 2\log\det\left(\frac{\hat{V}_{K+1}}{\lambda}+I\right).$$

Proof. First we use the fact that for all $x\in[0,1]$, $x \le 2\log(1+x)$:

$$\begin{aligned}&\min\left\{\sum_{h=1}^{H}\left\|(\hat{V}_k+\lambda I)^{-1/2}\Phi_{kh}C_{kh}\Phi_{kh}^\top(\hat{V}_k+\lambda I)^{-1/2}\right\|, 1\right\}\\
&\le 2\log\left(1+\sum_{h=1}^{H}\left\|(\hat{V}_k+\lambda I)^{-1/2}\Phi_{kh}C_{kh}\Phi_{kh}^\top(\hat{V}_k+\lambda I)^{-1/2}\right\|\right)\\
&= 2\log\left(1+\sum_{h=1}^{H}\left\|(\hat{V}_k+\lambda I)^{-1/2}\left(\Phi_{kh}\sum_{i=1}^{d_s}\partial_i\psi_h^k\,\partial_i\psi_h^{k\top}\Phi_{kh}^\top\right)(\hat{V}_k+\lambda I)^{-1/2}\right\|\right)\\
&\le 2\log\left(1+\sum_{h=1}^{H}\sum_{i=1}^{d_s}\left\|(\hat{V}_k+\lambda I)^{-1/2}\left(\Phi_{kh}\partial_i\psi_h^k\,\partial_i\psi_h^{k\top}\Phi_{kh}^\top\right)(\hat{V}_k+\lambda I)^{-1/2}\right\|\right)\\
&= 2\log\left(1+\sum_{h=1}^{H}\sum_{i=1}^{d_s}\mathsf{Tr}\left((\hat{V}_k+\lambda I)^{-1/2}\left(\Phi_{kh}\partial_i\psi_h^k\,\partial_i\psi_h^{k\top}\Phi_{kh}^\top\right)(\hat{V}_k+\lambda I)^{-1/2}\right)\right)\\
&= 2\log\left(1+\mathsf{Tr}\left((\hat{V}_k+\lambda I)^{-1/2}\sum_{h=1}^{H}\left(\Phi_{kh}\sum_{i=1}^{d_s}\partial_i\psi_h^k\,\partial_i\psi_h^{k\top}\Phi_{kh}^\top\right)(\hat{V}_k+\lambda I)^{-1/2}\right)\right)\\
&\le 2\log\det\left(I+(\hat{V}_k+\lambda I)^{-1/2}\sum_{h=1}^{H}\left(\Phi_{kh}\sum_{i=1}^{d_s}\partial_i\psi_h^k\,\partial_i\psi_h^{k\top}\Phi_{kh}^\top\right)(\hat{V}_k+\lambda I)^{-1/2}\right)\\
&= 2\log\det\left(\hat{V}_{k+1}+\lambda I\right) - 2\log\det\left(\hat{V}_k+\lambda I\right).\end{aligned}$$

In the first equality, we use the fact that the trace of a rank 1 matrix is equality to its spectral norm. In the last inequality, we used the fact that for any PSD matrix A, $\log\det(I+A)\ge\log(1+\mathsf{Tr}(A))$. In the last line, we used the definition of $\hat{V}_k$ (recall it is computed using the first $n=(k-1)H$ samples):

$$\begin{aligned}&\log\det(\hat{V}_{k+1}+\lambda I)\\
&= \log\det\left(\hat{V}_k+\lambda I+\sum_{h=1}^{H}\Phi_{kh}C_{kh}\Phi_{kh}^\top\right)\\
&= \log\det(\hat{V}_k+\lambda I)\\
&\quad+\log\det\left(I+(\hat{V}_k+\lambda I)^{-1/2}\sum_{h=1}^{H}\left(\Phi_{kh}\sum_{i=1}^{d_s}\partial_i\psi_h^k\,\partial_i\psi_h^{k\top}\Phi_{kh}^\top\right)(\hat{V}_k+\lambda I)^{-1/2}\right).\end{aligned}$$

Therefore, telescoping the sum we have:

$$\sum_{k=1}^{K} \min \left\{ \sum_{h=1}^{H} \left\| (\hat{V}_k + \lambda I)^{-1/2} \Phi_{kh} C_{kh} \Phi_{kh}^\top (\hat{V}_k + \lambda I)^{-1/2} \right\|, 1 \right\}$$
$$\leq 2 \sum_{k=1}^{K} \log\det\left(\hat{V}_{k+1} + \lambda I\right) - \log\det\left(\hat{V}_k + \lambda I\right) = 2\log\det\left(\frac{\hat{V}_{K+1}}{\lambda} + I\right).$$

□

C.2 Regret proof

Proof of Theorem 3. By Theorem 2, the event $\mathcal{E} := \{W_0 \in \mathcal{W}_k, \forall\, k \in [K]\}$ holds with probability at least $1 - \delta/2$. Henceforth, suppose $\mathcal{E}$ holds.

In episode k, we pick $\pi^k = \arg\max_\pi \max_{W \in \mathcal{W}_k} V_{W,1}^\pi(s_1^k)$. Let us denote $\tilde{W}_k = \arg\max_{W \in \mathcal{W}_k} V_{W,1}^k(s_1^k)$, that is, $\tilde{W}$ is the "optimistic model" which π^k is greedy with respect to. Under event $\mathcal{E}$, we have $V_{W_0,1}^\star(s_1^k) \leq V_{\tilde{W}_k,1}^k(s_1^k)$, so we have:

$$\mathcal{R}(K) = \sum_{k=1}^{K} \left[V_{W_0,1}^\star(s_1^k) - V_{W_0,1}^k(s_1^k)\right] \leq \sum_{k=1}^{K} \left[V_{\tilde{W}_k,1}^k(s_1^k) - V_{W_0,1}^k(s_1^k)\right].$$

Now we invoke Lemma 5 to get

$$\mathcal{R}(K) \leq \sum_{k=1}^{K}\sum_{h=1}^{H} \left[\mathbb{E}_{s_h^k a_h^k}^{\tilde{W}_k} V_{\tilde{W}_k,h+1}^k(s') - \mathbb{E}_{s_h^k a_h^k}^{W_0} V_{\tilde{W}_k,h+1}^k(s')\right] + \sum_{k=1}^{K}\sum_{h=1}^{H} m_h^k. \tag{10}$$

Since m_h^k is a martingale difference sequence with respect to $\mathcal{F}$ that is a.s. bounded by $2H$, by Azuma-Hoeffding we have:

$$\sum_{k=1}^{K}\sum_{h=1}^{H} m_h^k \leq 2H\sqrt{2T\log\left(\frac{2}{\delta}\right)},$$

with probability at least $1 - \delta/2$. We now proceed to bound the first term of (10). For every $(k, h) \in [K] \times [H]$:

$$\mathbb{E}_{s_h^k a_h^k}^{\tilde{W}_k} V_{\tilde{W}_k,h+1}^k(s') - \mathbb{E}_{s_h^k a_h^k}^{W_0} V_{\tilde{W}_k,h+1}^k(s') \leq H \cdot \mathsf{TV}\left(\mathbb{P}_{W_0}(\cdot|s_h^k, a_h^k,) \| \mathbb{P}_{\tilde{W}_k}(\cdot|s_h^k, a_h^k)\right)$$
$$\leq H \cdot \min\left\{\sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathbb{P}_{W_0}(\cdot|s_h^k, a_h^k) \| \mathbb{P}_{\tilde{W}_k}(\cdot|s_h^k, a_h^k)\right)}, 1\right\}, \tag{11}$$

where the first inequality uses the TV bound (Lemma 14) and the second inequality is Pinsker's inequality (with the additional observation that TV distance is always bounded by 1).

Next we proceed to upper bound the KL divergence in terms of the parameter estimation error of score matching.

Using Lemma 6 and the triangle inequality, we have

$$\sqrt{D_{\mathrm{KL}}\left(\mathbb{P}_{W_0}(\cdot|s_h^k, a_h^k) \| \mathbb{P}_{\tilde{W}_k}(\cdot|s_h^k, a_h^k)\right)} \leq \sqrt{\frac{\kappa}{2}} \left\| \mathsf{vec}\left(W_0\right) - \mathsf{vec}\left(\tilde{W}_k\right) \right\|_{\Phi_{kh}\Phi_{kh}^\top}$$
$$\leq \sqrt{\frac{\kappa}{2}} \left(\left\| \mathsf{vec}\left(W_0\right) - \mathsf{vec}\left(\hat{W}_k\right) \right\|_{\Phi_{kh}\Phi_{kh}^\top} + \left\| \mathsf{vec}\left(\hat{W}_k\right) - \mathsf{vec}\left(\tilde{W}_k\right) \right\|_{\Phi_{kh}\Phi_{kh}^\top} \right), \tag{12}$$

where we recall that $\hat{W}_k$ is the output of the score matching estimator at round k, a.k.a. the center of our confidence ellipsoid. Under $\mathcal{E}$, we know that $W_0, \tilde{W}_k \in \mathcal{W}_k$.

Now we apply our concentration guarantee Theorem 2. For any $W \in \mathcal{W}_k$, we know that:

$$
\begin{aligned}
&\left\| \mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right) \right\|^2_{\Phi_{kh}\Phi_{kh}^\top} \\
&= \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right)^\top \Phi_{kh}\Phi_{kh}^\top \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right) \\
&\leq \alpha_1^{-1} \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right)^\top \Phi_{kh} C_{kh} \Phi_{kh}^\top \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right) \\
&\leq \alpha_1^{-1} \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right)^\top (\hat{V}_k + \lambda I)^{1/2} (\hat{V}_k + \lambda I)^{-1/2} \Phi_{kh} C_{kh} \Phi_{kh}^\top \\
&\quad (\hat{V}_k + \lambda I)^{-1/2} (\hat{V}_k + \lambda I)^{1/2} \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right) \\
&\leq \alpha_1^{-1} \left\| \mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right) \right\|^2_{\hat{V}_k + \lambda I} \cdot \left\| (\hat{V}_k + \lambda I)^{-1/2} \Phi_{kh} C_{kh} \Phi_{kh}^\top (\hat{V}_k + \lambda I)^{-1/2} \right\| \\
&\leq \alpha_1^{-1} \beta_k^2 \cdot \left\| (\hat{V}_k + \lambda I)^{-1/2} \Phi_{kh} C_{kh} \Phi_{kh}^\top (\hat{V}_k + \lambda I)^{-1/2} \right\| \\
&\leq \alpha_1^{-1} \beta_{K+1}^2 \cdot \left\| (\hat{V}_k + \lambda I)^{-1/2} \Phi_{kh} C_{kh} \Phi_{kh}^\top (\hat{V}_k + \lambda I)^{-1/2} \right\|, \qquad (13)
\end{aligned}
$$

where the first inequality follows by Assumption 2 (iii), the third inequality by Cauchy-Schwarz, the fourth inequality because $W \in \mathcal{W}_k$, and the fifth inequality by the fact that β_k is monotonically increasing in k.

Thus, by combinine (11), (12), and (13), we bound the first term of (10) as:

$$
\begin{aligned}
&\sum_{k=1}^{K} \sum_{h=1}^{H} \left[\mathbb{E}^{\tilde{W}_k}_{s_h^k a_h^k} V^k_{\tilde{W}_k, h+1}(s') - \mathbb{E}^{W_0}_{s_h^k a_h^k} V^k_{\tilde{W}_k, h+1}(s') \right] \\
&\leq \sum_{k=1}^{K} \sum_{h=1}^{H} H \min\left\{ \beta_{K+1} \cdot \sqrt{\frac{\kappa}{\alpha_1} \left\| (\hat{V}_k + \lambda I)^{-1/2} \Phi_{kh} C_{kh} \Phi_{kh}^\top (\hat{V}_k + \lambda I)^{-1/2} \right\|}, 1 \right\} \\
&\leq \sum_{k=1}^{K} H^{3/2} \cdot \sqrt{ \sum_{h=1}^{H} \min\left\{ \beta_{K+1}^2 \cdot \frac{\kappa}{\alpha_1} \left\| (\hat{V}_k + \lambda I)^{-1/2} \Phi_{kh} C_{kh} \Phi_{kh}^\top (\hat{V}_k + \lambda I)^{-1/2} \right\|, 1 \right\} } \\
&\leq \sum_{k=1}^{K} H^{3/2} \cdot \sqrt{ \min\left\{ \beta_{K+1}^2 \cdot \frac{\kappa}{\alpha_1} \sum_{h=1}^{H} \left\| (\hat{V}_k + \lambda I)^{-1/2} \Phi_{kh} C_{kh} \Phi_{kh}^\top (\hat{V}_k + \lambda I)^{-1/2} \right\|, H \right\} } \\
&\leq \sum_{k=1}^{K} H^{3/2} \cdot \sqrt{ \max\left\{ \beta_{K+1}^2 \cdot \frac{\kappa}{\alpha_1}, H \right\} \cdot \min\left\{ \sum_{h=1}^{H} \left\| (\hat{V}_k + \lambda I)^{-1/2} \Phi_{kh} C_{kh} \Phi_{kh}^\top (\hat{V}_k + \lambda I)^{-1/2} \right\|, 1 \right\} } \\
&\leq H^{3/2} \cdot \sqrt{ K \left(\beta_{K+1}^2 \cdot \frac{\kappa}{\alpha_1} + H \right) } \cdot \sqrt{ \sum_{k=1}^{K} \min\left\{ \sum_{h=1}^{H} \left\| (\hat{V}_k + \lambda I)^{-1/2} \Phi_{kh} C_{kh} \Phi_{kh}^\top (\hat{V}_k + \lambda I)^{-1/2} \right\|, 1 \right\} } \\
&\leq H^{3/2} \cdot \sqrt{ K \left(\beta_{K+1}^2 \cdot \frac{\kappa}{\alpha_1} + H \right) } \cdot \sqrt{ \log\det\left(\frac{\hat{V}_{K+1}}{\lambda} + I \right) },
\end{aligned}
$$

where the last inequality uses Lemma 7.

From here, we bound $\gamma_{K+1} := \log\det(\hat{V}_{K+1}/\lambda + I)$ via the trace-determinant inequality:

$$
\begin{aligned}
&\det\left(\frac{1}{\lambda} \hat{V}_{K+1} + I \right) \leq \left(\frac{1}{d_\psi d_\phi} \mathsf{Tr}\left(\frac{1}{\lambda} \hat{V}_{K+1} + I \right) \right)^{d_\psi d_\phi} \\
&\leq \left(\frac{\alpha_2}{d_\psi d_\phi \lambda} \sum_{k=1}^{K} \sum_{h=1}^{H} d_\psi \cdot \|\phi_{kh}\|^2 + 1 \right)^{d_\psi d_\phi} \leq \left(\frac{\alpha_2}{d_\phi \lambda} T B_\phi^2 + 1 \right)^{d_\psi d_\phi}.
\end{aligned}
$$

Putting it all together, we get a regret guarantee of:

$$\begin{aligned}
\mathcal{R}(K) &\leq H^{3/2} \cdot \sqrt{K\left(\beta_{K+1}^2 \cdot \frac{\kappa}{\alpha_1} + H\right)} \cdot \sqrt{\log\det\left(\frac{\hat{V}_{K+1}}{\lambda} + I\right)} + 2H\sqrt{2T\log\left(\frac{2}{\delta}\right)} \\
&\leq C\sqrt{\gamma_{K+1} \cdot \left(\frac{2\kappa(B_\psi + B_c)}{\alpha_1^3}\left(\gamma_{K+1} + \log{}^2/_\delta\right) + \frac{\kappa}{\alpha_1} + H\right)} \cdot \sqrt{H^2 T} + 2H\sqrt{2T\log{}^2/_\delta} \\
&= \tilde{O}(d_\psi d_\phi \sqrt{H^3 T}),
\end{aligned}$$

where we use the definition of β_{K+1} as well as the bound on the information gain quantity γ_{K+1}. This concludes the proof. $\square$

D Details of comparison to prior work

D.1 Comparison with Exp-UCRL

In this section, we provide further details about our comparison to Chowdhury et al. and Kakade et al.. First, we translate the quantities in Chowdhury et al. to our notation:

- $\mathbb{A}_{ij} = \mathsf{Tr}(A_i A_j^\top)$; since the A_i take the form of $E_{(k,l)}$, $\mathbb{A} = I_{d_\psi d_\phi}$.
- $G_{sa} = \Phi_{sa}\Phi_{sa}^\top$.
- $\|\theta^\star\|_{\mathbb{A}} = \|W_0\|_F$, so therefore $B_{\mathbb{A}} = B_\star$ in our notation.
- $\left\|\mathbb{A}^{-1}G_{sa}\right\| = \|G_{sa}\| = \left\|\Phi_{sa}\Phi_{sa}^\top\right\|$, so therefore $B_{\varphi,\mathbb{A}} = B_\phi^2$ in our notation.
- They require that for all θ, s, a: $\alpha \preceq \mathbb{C}_{sa}^\theta[\psi(s')] \preceq \beta$. Recall that $\mathbb{C}_{sa}^\theta[\psi(s')]$ is the covariance of $\psi(s')$, as introduced in Assumption 2 (iv). Therefore, $\beta = \kappa$ in our notation, and we do not require α to be bounded for SMRL. As they state in their paper, α controls the strict convexity of the log partition function that guarantees minimality of the exponential family; it may be possible to remove the dependence on α in their proofs.
- We use K for episodes instead of T and T for total number of interactions instead of N.

Thus, we can restate their theorem (in our notation) as:

Theorem 8 (Thm. 2 of [11])**.** *With regularizer $\lambda > 0$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ the regret of Exp-UCRL can be bounded as:*

$$\mathcal{R}(K) \leq \tilde{O}\left(\sqrt{\frac{\kappa}{\alpha}\left(1 + \frac{\kappa B_\phi^2 H}{\lambda}\right)\left(\lambda B_\star^2 + d_\phi d_\psi\right) d_\phi d_\psi} \cdot \sqrt{H^2 T}\right).$$

In comparison, our Theorem 3 can be stated as:

Theorem 9 (Theorem 3, restated)**.** *With regularizer $\lambda > 0$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ the regret of SMRL can be bounded as:*

$$\mathcal{R}(K) \leq \tilde{O}\left(\sqrt{\left(\frac{\kappa(B_\psi + B_c)}{\alpha_1^3} d_\phi d_\psi + \lambda B_\star^2 \frac{\kappa}{\alpha_1} + H\right) d_\phi d_\psi} \cdot \sqrt{H^2 T}\right).$$

D.2 Comparison details for nonLDS

First, we formally state the equivalence between score matching and MLE for the nonLDS problem.

Proposition 10. Let $\mathbb{P}_W(s'|s,a) := (2\pi\sigma^2)^{-d/2}\exp\left(-\frac{1}{2\sigma^2}\|s' - W\phi(s,a)\|_2^2\right)$. For any dataset $\mathcal{D} = \{(s_t, a_t, s_t')\}_{t\in[n]}$:

$$\text{(SM-L)} = \frac{1}{8\sigma^4}\sum_{t=1}^n \|s_t' - W\phi(s_t,a_t)\|_2^2.$$

$$\text{(MLE)} := \arg\min_W -\sum_{t=1}^n \log\mathbb{P}_W(s_t'|s_t,a_t) = \arg\min_W \frac{1}{2\sigma^2}\sum_{t=1}^n \|s_t' - W\phi(s_t,a_t)\|_2^2.$$

Proposition 10 is straightforward from the definition of the estimators as well as the form of the gaussian density; we omit the proof.

Next, we select the regularization parameter λ for each algorithm in order for the estimation procedure to be exactly the same for a fixed dataset.

- LC^3: $\hat{W} = \arg\min_W \sum_{t=1}^n \|s'_t - W\phi(s_t, a_t)\|_2^2 + \frac{\lambda_{\text{LC}}}{2} \|W\|_F^2$ with $\lambda_{\text{LC3}} = \frac{\sigma^2}{B_\star^2}$ [27, Eq. 3.1].
- Exp-UCRL: $\hat{W} = \arg\min_W - \sum_{t=1}^n \log \mathbb{P}_W(s'_t|s_t, a_t) + \frac{\lambda_{\text{Exp}}}{2} \|W\|_F^2$, so we set $\lambda_{\text{Exp}} = \frac{1}{2B_\star^2}$ [11, Eq. 2].
- SMRL: $\hat{W} = \arg\min_W \hat{J}_n(W) + \frac{\lambda_{\text{SM}}}{2} \|W\|_F^2$, so we set $\lambda_{\text{SM}} = \frac{1}{8\sigma^2 B_\star^2}$ (Algorithm 1, line 9).

E Technical results

Lemma 11 (Concentration of Self Normalized Process)**.** *Let $\{\Delta_t\}_{t=1}^\infty$ be an $\mathbb{R}^m$-valued stochastic process with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. Let $\Delta_t|\mathcal{F}_{t-1}$ be zero-mean and σ^2-subgaussian; i.e. $\mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] = 0$, and*

$$\forall \lambda \in \mathbb{R}^d, \qquad \mathbb{E}[e^{\lambda^\top \Delta_t}|\mathcal{F}_{t-1}] \le e^{\sigma^2 \|\lambda\|^2/2}.$$

Let $\{\Phi_t\}_{t=1}^\infty$ be an $\mathbb{R}^{d\times m}$-valued stochastic process where $\Phi_t \in \mathcal{F}_{t-1}$. Assume V_0 is a $d \times d$ positive definite matrix, and let $V_t = \sum_{s=1}^t \Phi_s \Phi_s^\top$ for $t \ge 1$. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have for all $t \ge 1$:

$$\left\| \sum_{s=1}^t \Phi_s \Delta_s \right\|_{V_t^{-1}}^2 \le 2\sigma^2 \log \frac{\det(V_t + V_0)^{1/2}}{\delta \det V_0^{1/2}}$$

Proof. Let $S_t = \sum_{s=1}^t \Phi_t \Delta_t$, define

$$M_n^\gamma := \exp\left\{ \frac{1}{\sigma} \gamma^\top S_n - \frac{1}{2} \gamma^\top V_n \gamma \right\}.$$

We first prove that M_n^γ is a super-martingale adapted to filtration $\{\mathcal{F}_{t-1}\}_{t=1}^\infty$. Define

$$D_t^\gamma := \exp\left\{ \frac{1}{\sigma} \gamma^\top \Phi_t \Delta_t - \frac{1}{2} \gamma^\top \Phi_t \Phi_t^\top \gamma \right\}.$$

We have that $M_n^\gamma = M_{n-1}^\gamma D_n^\gamma$. So it suffices to show $\mathbb{E}[D_t^\gamma|\mathcal{F}_{t-1}] \le 1$. Using the subgaussian property, we have

$$\mathbb{E}[D_t^\gamma|\mathcal{F}_{t-1}] \le \exp\left\{ \sigma^2 \frac{1}{2\sigma^2} \|\gamma^\top \Phi_t\|^2 - \frac{1}{2} \gamma^\top \Phi_t \Phi_t^\top \gamma \right\} = 1.$$

Next, we use the method of mixtures [e.g., 29, Ch. 20]. Define:

$$\overline{M}_n := \int M_n^\gamma f(\gamma) d\gamma,$$

where $f(\gamma)$ is the pdf of normal distribution $\mathcal{N}(0, V_0^{-1})$. On the one hand, we can prove that $\overline{M}_n$ is a super-martingale adapted to filtration $\{\mathcal{F}_{t-1}\}_{t=1}^\infty$ since

$$\mathbb{E}[\overline{M}_n|\mathcal{F}_{n-1}] = \mathbb{E}[\mathbb{E}[M_n^\gamma]|\mathcal{F}_{n-1}] = \mathbb{E}[\mathbb{E}[M_n^\gamma|\mathcal{F}_{n-1}]] \le \mathbb{E}[M_{n-1}^\gamma] = \overline{M}_{n-1}.$$

This implies that the maximal inequality holds for any $\delta \in (0, 1)$ [29, Thm. 3.9]:

$$\mathbb{P}\left(\sup_{t\in\mathbb{N}} \overline{M}_t \ge \frac{1}{\delta} \right) \le \delta \tag{14}$$

On the other hand, we can compute $\overline{M}_n$ directly. We have

$$
\begin{aligned}
\overline{M}_n &= \int \exp\left\{\frac{1}{\sigma}\gamma^\top S_n - \frac{1}{2}\gamma^\top V_n \gamma\right\} f(\gamma) d\gamma \\
&= \frac{1}{(2\pi)^{m/2}\det V_0^{-1/2}} \int \exp\left\{\frac{1}{\sigma}\gamma^\top S_n - \frac{1}{2}\gamma^\top V_n\gamma - \frac{1}{2}\gamma^\top V_0 \gamma\right\} d\gamma \\
&= \frac{\exp\{\frac{1}{2\sigma^2} S_n^\top (V_n + V_0)^{-1} S_n\}}{(2\pi)^{m/2}\det V_0^{-1/2}} \int \exp\left\{-\frac{1}{2}\|\gamma - (V_n + V_0)^{-1}\frac{S_n}{\sigma}\|^2_{V_n+V_0}\right\} d\gamma \\
&= \frac{\det V_0^{1/2}}{\det(V_n + V_0)^{1/2}} \exp\left\{\frac{1}{2\sigma^2} S_n^\top (V_n + V_0)^{-1} S_n\right\}. \qquad (15)
\end{aligned}
$$

The last equality is due to integration of the Gaussian pdf. Therefore, by combining (14) and (15) we obtain that with probability $1-\delta$, for all $n \in \mathbb{N}$:

$$
\|S_n\|^2_{(V_n+V_0)^{-1}} \leq 2\sigma^2 \log \frac{\det(V_n + V_0)^{1/2}}{\delta \det V_0^{1/2}}.
$$

$\square$

Lemma 12 (Log Partition Derivatives, [11])**.** *The first and second derivatives of the exponential family model in Definition 1 are:*

$$
\begin{aligned}
\nabla_{(i,j)} Z_{sa}(W) &= \mathbb{E}^W_{sa}[\psi(s')]^\top E_{ij}\phi(s,a), \\
\nabla^2_{(i,j),(k,l)} Z_{sa}(W) &= \phi(s,a)^\top E_{ij}^\top \left(\mathbb{E}^W_{sa}[\psi(s')\psi(s')^\top] - \mathbb{E}^W_{sa}[\psi(s')]\mathbb{E}^W_{sa}[\psi(s')^\top\right) E_{kl}\phi(s,a),
\end{aligned}
$$

for any $(i,j),(k,l) \in [d_\psi] \times [d_\phi]$.

Lemma 13. *If $X \in \mathbb{R}^d$ is a conditionally σ_1^2-subgaussian vector and $Y \in \mathbb{R}^d$ is a σ_2^2-subgaussian vector, then $X+Y$ is a conditionally $\sigma_1^2+\sigma_2^2$-subgaussian vector.*

Proof. We have for any $v \in \mathbb{R}^d$ and $p, q \geq 1$ such that $1/p + 1/q = 1$:

$$
\mathbb{E}\left[e^{v\cdot(X+Y)} \,\middle|\, \mathcal{F}_t\right] \leq \mathbb{E}\left[e^{pv\cdot X} \mid \mathcal{F}_t\right]^{1/p} \mathbb{E}\left[e^{qv\cdot X} \mid \mathcal{F}_t\right]^{1/q} \leq e^{\|v\|_2^2\left(p\sigma_1^2 + q\sigma_2^2\right)/2} = e^{\|v\|_2^2(\sigma_1^2+\sigma_2^2)/2}.
$$

where we use Hölder's inequality in the first inequality, subgaussianity of X, Y in the second inequality, and set $p = \frac{\sigma_2}{\sigma_1} + 1$ in the equality. $\square$

Lemma 14. *Suppose function $f : \mathbb{R}^m \to [0,1]$, and p, q are two probability density functions on $\mathbb{R}^m$. Suppose s is an $\mathbb{R}^m$-valued random variable. Then it holds that*

$$
|\mathbb{E}_p f(s) - \mathbb{E}_q f(s)| \leq \mathrm{TV}(p\|q).
$$

Proof. Denote $E = \{x \in \mathbb{R}^m : p(x) > q(x)\}, F = \mathbb{R}^m \backslash E$. We have

$$
\begin{aligned}
\mathbb{E}_p f(s) - \mathbb{E}_q f(s) &= \int_{\mathbb{R}^m} f(x)\left(p(x) - q(x)\right) dx \\
&= \int_E f(x)\left(p(x)-q(x)\right) dx - \int_F f(x)\left(q(x) - p(x)\right) dx \\
\Rightarrow |\mathbb{E}_p f(s) - \mathbb{E}_q f(s)| &\leq \max\left\{\int_E f(x)\left(p(x)-q(x)\right) dx, \int_F f(x)\left(q(x)-p(x)\right) dx\right\} \\
&= \max\left\{\left|\int_E f(x)\left(p(x)-q(x)\right) dx\right|, \left|\int_F f(x)\left(p(x)-q(x)\right) dx\right|\right\} \\
&\leq \sup_{A\subset\mathbb{R}^m} \left|\int_A f(x)(p(x)-q(x))dx\right| = \mathrm{TV}(p\|q).
\end{aligned}
$$

$\square$

F Experimental details

We explain the experimental setup in greater detail. Our code can be found on the Github.

F.1 MDP construction

We recall the MDP setup from Section 5.3. The MDP has $\mathcal{S} = \mathbb{R}$, $\mathcal{A} = \{+1, -1\}$, and $H = 10$. The initial state distribution is $\mathrm{Unif}([-1, +1])$. The transition and reward are given by

$$\mathbb{P}(s'|s,a) = \exp\left(-\frac{s'^{1.7}}{1.7}\right) \cdot \exp\left(\sin(4s')(s+a)\right), \tag{16}$$

$$r(s,a) = \exp\left(-10\left(s - \frac{\pi}{8}\right)^2\right) + \exp\left(-10\left(s + \frac{3\pi}{8}\right)^2\right). \tag{17}$$

Thus, the transition can be written in the form of Definition 1 with

$$q(s') := \exp\left(-\frac{s'^{1.7}}{1.7}\right), \quad \psi(s') := \sin(4s'), \quad W_0 = [1, 1], \quad \phi(s,a) = [s, a]^\top.$$

The MDP has the property that states will mostly be confined to the region $[-3, +3]$, and the conditional density $\mathbb{P}(s'|s, a = +1)$ generally has higher mass in regions with reward than the conditional density $\mathbb{P}(s'|s, a = -1)$ does. Thus, selecting $a = +1$ is preferred over $a = -1$. (We can get the opposite to happen by setting $W_0 = [1, -1]$ instead.) This is illustrated in Figure 3 for various starting states s.

F.2 Model estimation

We consider two model estimation methods, given a dataset of transitions $\mathcal{D} = \{(s_t, a_t, s'_t)\}_{t \in [n]}$.

- We can use the score matching estimator (SM-E) for model class $\mathcal{P}(\psi, \phi, q)$ to get an estimate $\hat{W}_{\mathrm{SM}}$ for W_0 and corresponding estimated transition model $\hat{\mathbb{P}}_{\mathrm{SM}}$.
- We can treat the system as an LDS and solve the least squares problem

  $$\hat{W}_{\mathrm{LS}} = \operatorname*{arg\,min}_{W \in \mathbb{R}^2} \sum_{t=1}^{n} \|s'_t - W\phi(s_t, a_t)\|_2^2 + \frac{\lambda}{2}\|W\|_F^2.$$

  Then we estimate the transition model $\hat{\mathbb{P}}_{\mathrm{LDS}}$ as $s' = \langle \hat{W}_{\mathrm{LS}}, \phi(s,a)\rangle + \eta$, where $\eta \sim \mathcal{N}(0, \sigma^2)$ is independent noise. This approach is the estimation procedure of LC^3 with the feature mapping $\phi(s,a) = [s,a]^\top$.

In both methods we set the regularization parameter $\lambda = 0$.

We also remark that since we are working with a low-dimensional transition function, a third option is the estimation procedure of Exp-UCRL: numerically compute the MLE with the given model class $\mathcal{P}(\psi, \phi, q)$. Experimentally for the RL task, we expect this to obtain similar performance with score matching, since for this density, MLE and score matching learn very similar models. However, we do not investigate this approach because MLE is only computationally tractable for low-dimensional problems and will not scale to higher-dimensional MDPs.

In Figure 4, we plot the estimated densities for various states s under i.i.d. transition data. Since the ground truth model is well-specified in the class $\mathcal{P}(\psi, \phi, q)$ (and also not representable as an LDS), score matching learns a more accurate transition model than fitting an LDS does.

F.3 RL task evaluation

We evaluate end-to-end performance for the RL task. To isolate the benefit of better model estimation via score matching, we fix a simple random sampling shooting (RSS) planner. (Recall from Section 4.2 that we cannot implement SMRL and LC^3 as written because of the computational intractability of optimistic planning. The RSS planner is an effective, easy-to-implement alternative.) In every

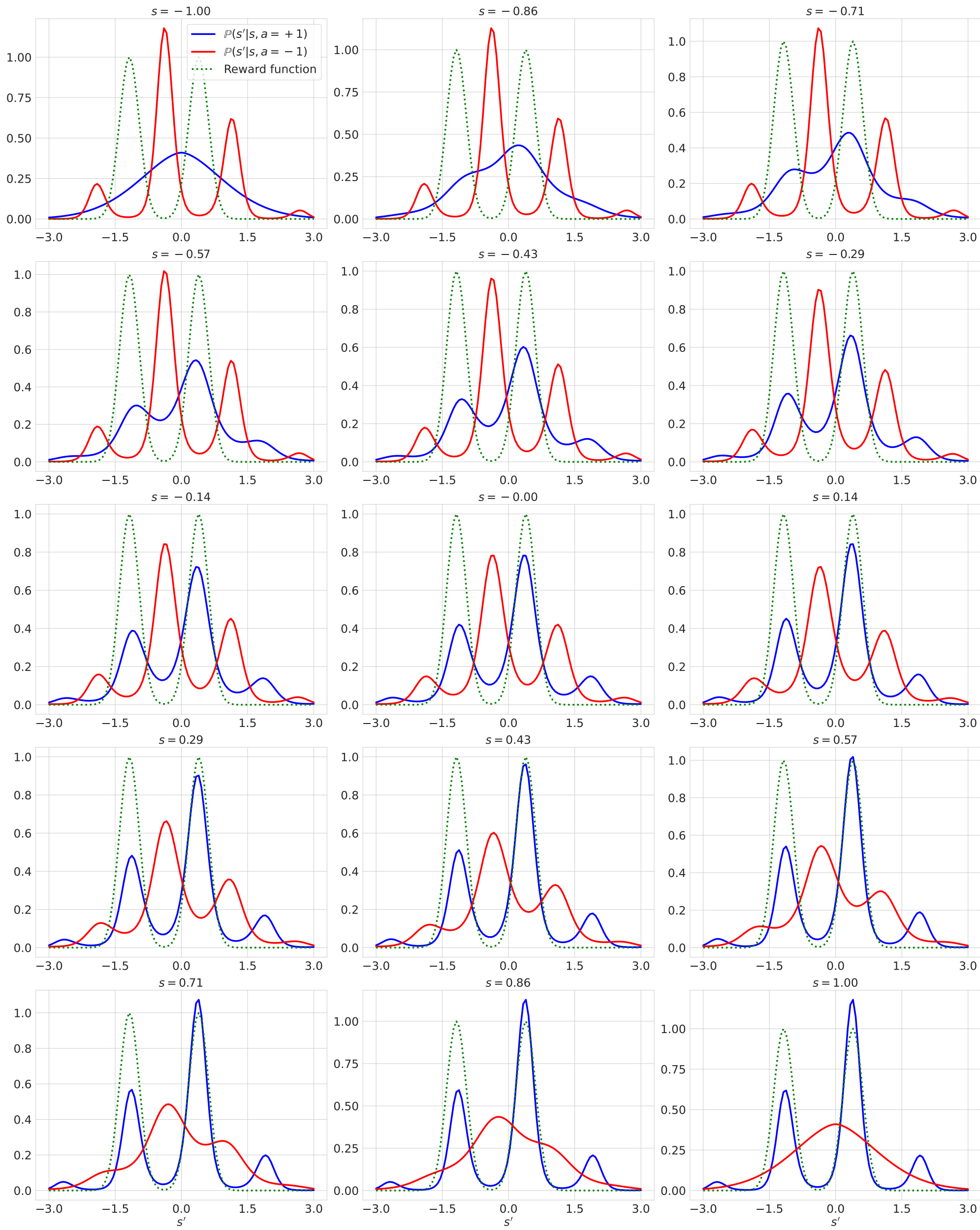


Figure 3: Plotting $\mathbb{P}(s'|s, a)$ for various start states for $a = +1$ (blue) and $a = -1$ (red); reward is superimposed (dotted green). For the plotted states $s \in [-1, +1]$, taking $a = +1$ is more likely to transition to states with higher reward than taking $a = -1$ will.

call, the RSS planner takes as input an estimated model $\hat{\mathbb{P}}$ and a reward function r and an initial state s_{init}. To evaluate the action a, it simulates a dataset of n_{rollouts} independent trajectories, each of length τ, where the initial state in each trajectory is s_{init}, action a is played in every step, and subsequent states are drawn using the estimated model $\hat{\mathbb{P}}$; the estimated value of taking action a is the average of the cumulative rewards for these trajectories. Lastly, the planner selects the action in the real environment which has highest simulated average reward. In experiments, we set the hyperparameters $n_{\text{rollouts}} := 100$ and $\tau := 5$.

To evaluate the estimation methods, we run the RSS planner in the MDP at every step $h \in [H]$. The model is re-estimated at the end of every episode using the collected transition data, using either score matching or fitting an LDS. We also compare against a baseline where the RSS planner is supplied with the ground truth model $\mathbb{P}$.

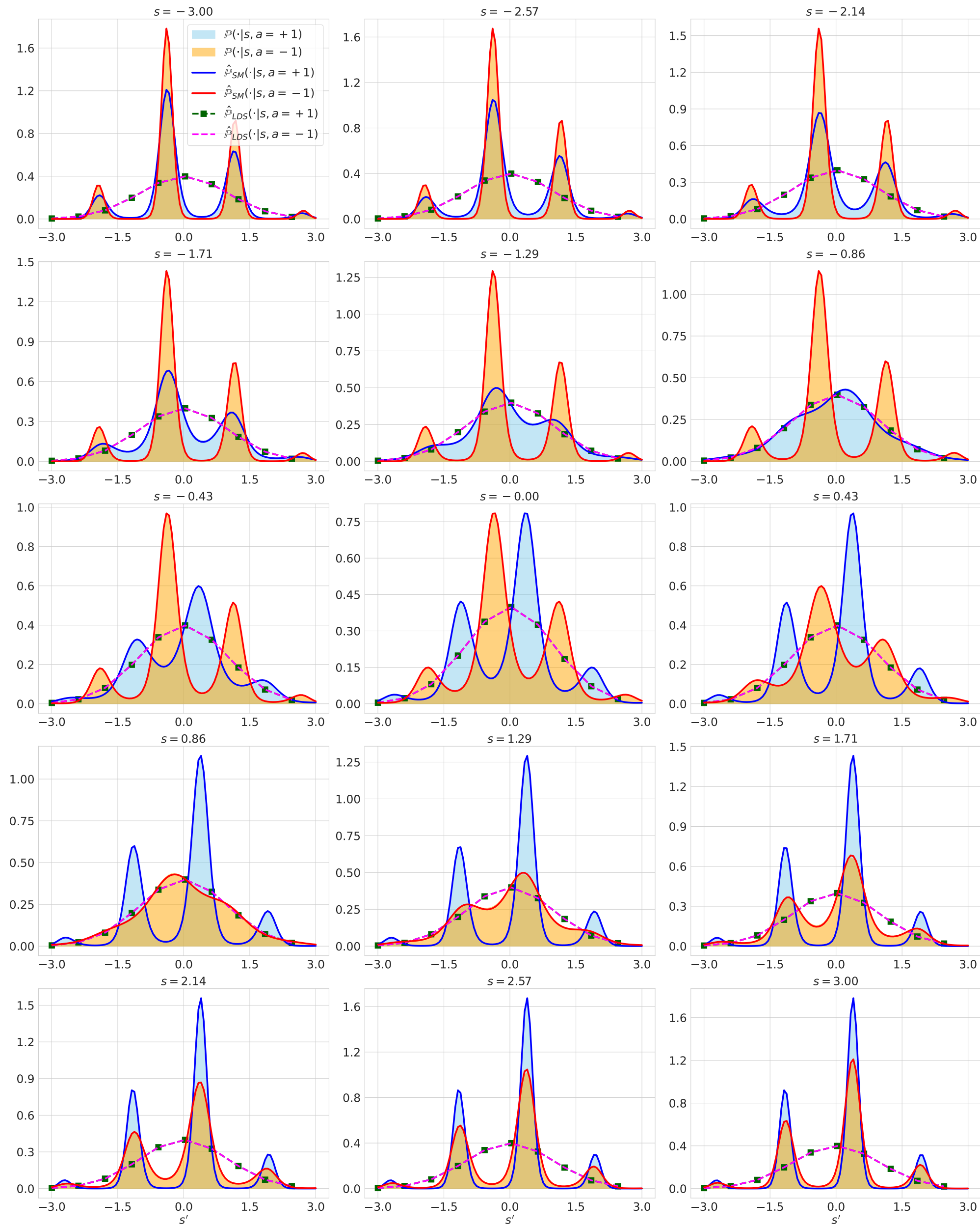


Figure 4: Plotting learned densities via score matching and fitting an LDS under i.i.d. data. We generate $n = 1000$ samples where (s, a) are sampled i.i.d. from $\mathrm{Unif}([-1, +1]) \times \mathrm{Unif}(\{-1, +1\})$ and the s' are drawn according to Equation (16). The ground truth densities for $a = +1$ and $a = -1$ are plotted as the shaded blue and orange regions respectively. Score matching learns the correct density for both $a = +1$ (blue lines) and $a = -1$ (red lines), while the LDS model is unable to differentiate between the two, and essentially learns the same model $\hat{\mathbb{P}}_{\mathrm{LDS}}(s'|s, a = +1) \approx \hat{\mathbb{P}}_{\mathrm{LDS}}(s'|s, a = -1)$ (green and magenta lines).