Exponential Family Model-Based Reinforcement Learning via Score Matching

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose a 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. SMRL achieves $\tilde{O}(d\sqrt{H^3T})$ regret, where H is the length of each episode and T is the total number of interactions.

1 Introduction

This paper studies the regret minization 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 [28, 34, 23]. 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., 44, 21, 8, 29] or action-value functions [e.g., 25, 45] of the Markov Decision Process (MDP). However, most real world problems are *nonlinear*, and 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. [9] 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. Specifically, Chowdhury et al. study transition models 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 ψ and ϕ are known feature mappings in $\mathbb{R}^{d_\psi}$ and $\mathbb{R}^{d_\phi}$ respectively, q is a known base measure, Z_{sa} is the log partition function, and W_0 is the unknown parameter to be learned. This transition model class covers linear dynamical systems, as well as its nonlinear generalizations [27, 22] as special cases. Linear dynamical systems with quadratic rewards, i.e., the linear quadratic regulator (LQR), have received much attention recently as an important testbench for RL in unknown, complex environments [15, 35, 22]. 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 in general can be computationally

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

intractable. In addition, the estimator $\hat{W}$ cannot be computed in closed form; Chowdhury et al. leave their estimator implicitly defined in terms of the likelihood equations. As is typical in upper confidence RL (UCRL) algorithms, one constructs high probability confidence sets around the estimator. Due to the challenging modeling assumption, the confidence sets employed by Chowdhury et al. 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 [17]. 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, 21].

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 design a model-based algorithm, dubbed SMRL, which achieves regret of $\tilde{O}(d\sqrt{H^3T})$, with mild polynomial dependence on the problem constants (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 model parameters from data. It then constructs elliptic confidence regions around the estimator which contain the true parameter with high probability, and chooses policies optimistically based on such confidence regions.[1]

Our regret guarantee matches that of Exp-UCRL, the model-based algorithm proposed by Chowdhury et al.. When specialized to the nonlinear dynamical system setting with bounded costs and features, score matching and MLE are equivalent estimators. Here, the recent work of Kakade et al. [22] gives an guarantee of $\tilde{O}(\sqrt{d_\phi(d_\phi + d_\psi + H)H^2T})$ (and suggest a naive lower bound of $\Omega(\sqrt{HT})$ from results on tabular MDPs); however we stress that our analysis applies to a broader class of models. We leave providing a sharper analysis for both upper and lower bounds to future work. Broadly speaking, we view score matching and MLE as complementary estimation techniques; while MLE generally 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 4.

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 $B - A \preceq 0$. 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 $\partial_i^2 f(x) := \left(\frac{\partial^2}{\partial x_i^2} f_1(x), \ldots, \frac{\partial^2}{\partial x_i^2} f_n(x)\right)^\top \in \mathbb{R}^n$.

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.

[1] Optimistic planning is hard in the worst case, and developing fast approximation algorithms is an active area of research. This work assumes computational oracle access to such a planner.

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. For each step $h \in [H]$, the agent observes the state s_h and plays action $a_h \in \mathcal{A}$. Afterwards, they observe reward $r_h(a_h, s_h)$, 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.

Now we define the policy function, value function, and action-value function which will be central to our results. A **policy function** is a collection of functions $\pi := \{\pi_h : \mathcal{S} \mapsto \mathcal{A}\}_{h \in [H]}$ which determine the agent's strategy for interacting with the world, i.e., when presented with state s at step h, the agent will play $\pi_h(s)$. 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 [41]. Given knowledge of the MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, the agent can compute the optimal value function (and optimal action-value function) via dynamic programming [40]; 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**. For every episode $k \in [K]$, an adversary presents the agent with a state s^k_1, 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 $\pi^\star$:

$$\mathcal{R}(K) := \sum_{k=1}^{K} \left(V^{\pi^\star}_1(s^k_1) - V^{\pi^k}_1(s^k_1)\right).$$

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 $\{s^k_1\}_{k \in [K]}$. We will also denote $T := KH$ as the total number of interactions the agent makes with the world.

2.1 Exponential Family Transition Model

In this paper, we propose a model-based approach to online RL, meaning that in every episode, the agent will explicitly estimate the model $\mathbb{P}$ in order to pick their policy π^k. We consider the following family of transition models from [9]. *Therefore, the goal of the RL algorithm will be to achieve sublinear (in T) regret when run on MDPs satisfying Assumption 1.*

Assumption 1. Suppose $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ and $\mathcal{A}$ is any arbitrary action set. Let *feature mappings* $\psi : \mathcal{S} \mapsto \mathbb{R}^{d_\psi}$ and $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_\phi}$, as well as *base measure* $q : \mathcal{S} \to \mathbb{R}$ be known to the learner. The state transition measures are conditional exponential family models, parameterized by an unknown matrix $W_0 \in \mathbb{R}^{d_\psi \times d_\phi}$:

$$\mathbb{P}_{W_0}(s'|s,a) = q(s') \exp\left(\langle \psi(s'), W_0 \phi(s,a) \rangle - Z_{sa}(W_0)\right), \tag{2}$$

where

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

Here, $Z_{sa}(\cdot)$ is the *log-partition function*, which is completely determined once ψ, ϕ, and q are specified. In addition, we assume that the reward function $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is bounded a.s. in $[0, 1]$ and known to the learner.

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

Discussion. The conditional exponential family model naturally arises in the well-known statistical framework of generalized linear models (GLMs), which we briefly review. GLMs model situations where the expected response $Y \in \mathbb{R}$, is conditionally related to the covariates X as $\mathbb{E}[Y|X] = \mu(\theta^\top \phi_X)$, where $\mu : \mathbb{R} \to \mathbb{R}$ is the so-called *link function*, $\theta \in \mathbb{R}^d$ is some unknown parameter, and ϕ is a feature mapping for X. To define a GLM, one first defines a probability distribution which belongs to the *canonical exponential family*, i.e. the probability density has the form:

$$p_\eta(y) = q(y) \cdot \exp(y \cdot \eta - Z(\eta)),$$

where $q(\cdot)$ is a real valued function, η is a real-valued parameter, and $Z(\cdot)$ is a twice differentiable function. Then, the GLM associated with the exponential family model is determined by $\mathbb{P}_\theta(y|x) := p_{\theta^\top \phi_x}(y) = q(y) \exp(y \cdot \theta^\top \phi_x - Z(\theta^\top \phi_x))$; it can be shown that $\mathbb{E}[Y|X] = \mu(\theta^\top \phi_X) = Z'(\eta)|_{\eta = \theta^\top \phi_X}$. Therefore, Assumption 1 can be viewed as a multivariate generalization of the GLM, where the response is a function of the next state $\psi(s')$ and the covariates are current state-action pair (s, a).

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 [29, 5]. Assumption 1 is a special case with $d = d_\psi \times d_\phi$ and the $A_{ij} := E_{ij}$. Our results can be extended to this 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.

As a side note, a *generalization* of Assumption 1 is possible by assuming that $\mathbb{P}_{W_0}(s'|s,a) = q(s') \cdot \exp\left(\langle \psi(s', s, a), \theta \rangle - Z_{sa}(\theta)\right)$, where $\psi : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_\psi}$ is a known feature mapping and $\theta \in \mathbb{R}^{d_\psi}$ is an unknown parameter. This can be viewed as a nonlinear generalization of the "linear kernel MDP" studied in [8, 46]. To the best of our knowledge, there are no results in RL which cover this setting.

The known and bounded rewards assumption is fairly standard in model-based reinforcement learning, see e.g., [44, 5, 9]. Our results easily 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 guarantee would additionally depend on the complexity of $\mathcal{R}$, i.e., via a complexity measure such as eluder dimension. For example, if the reward is unknown but is linear in the state-action feature vector, i.e., $\mathbb{E}[r(s,a)] = \langle \theta^\star, \phi(s,a) \rangle$, the regret would increase by an additive factor of $\tilde{O}(d\sqrt{T})$. The main focus of this paper is to understand the statistical and computational complexity associated with the more difficult task of estimating state transition measures; as such, we omit the extension to unknown rewards.

2.2 Relationship to (Non)linear Dynamical Systems

We now describe how Assumption 1 generalizes previous models of 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:

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

where $s, s' \in \mathbb{R}^{d_s}$, $u \in \mathbb{R}^{d_u}$, 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, nonlinear generalizations of LDS have also been proposed and studied in terms of system identification and regret minimization [27, 22]. 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. We note that nonLDS (and by extension the LDS) are special cases of Assumption 1. By writing out the probability density function of the multivariate gaussian distribution, one can read off the functions (details omitted due to space):

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

Lastly, we note that Assumption 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 distribution. Assumption 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 feature mapping ψ to be some polynomial in s', or even the output of a neural network; this is not permitted under the nonLDS setting. The contribution of this paper (as well as Chowdhury et al.) is to show provably efficient guarantees for a class of problems which subsumes nonLDS.

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.

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 several pitfalls. Solving for the MLE requires computing the log-partition function $Z_{sa}(\cdot)$, which in many settings is intractable. In our application, it is important to derive closed-form solution of an estimator with high probability confidence intervals in order to apply the upper confidence bound (UCB) technique; in general, we cannot get this with the MLE estimator.[4] As we will see, via score matching we get an easy-to-compute estimator and high probability confidence sets which are ellipsoids.

Hyvärinen [17] 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:

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

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

[3] Kakade et al. [22] study a infinite dimensional version of this model, which they call the *kernelized nonlinear regulator*.

[4] The nonLDS is an example when it is actually possible to compute closed form solutions with confidence intervals via MLE, as shown in [22].

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., 37, 4].

(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 $i \in [d_s]$, $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\begin{aligned} \|\partial_i \psi(s')\| &\in L^2(\mathcal{S}, \mathbb{P}_{W_0}(\cdot|s, a)), \\ \left\|\partial_i^2 \psi(s')\right\| &\in L^1(\mathcal{S}, \mathbb{P}_{W_0}(\cdot|s, a)), \\ \|\partial_i \psi(s')\| \, \partial_i \log q(s') &\in L^1(\mathcal{S}, \mathbb{P}_{W_0}(\cdot|s, a)). \end{aligned}$$

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 some further discussion about these regularity conditions.

We make the crucial point that (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 naturally gives rise to an empirical score matching loss for a dataset $\mathcal{D} = \{(s_t, a_t, s_t')\}_{t \in [n]}$:

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

Furthermore, for any regularizer $\lambda > 0$, we can define the empirical score matching estimator $\hat{W}_{n,\lambda} := \arg\min_W \hat{J}(W) + \frac{\lambda}{2} \|W\|_F^2$. (When the value of λ is understood, we will drop it in the subscript.)

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

Theorem 1. *For a dataset $\mathcal{D} = \{(s_t, a_t, s_t')\}_{t \in [T]}$, we have:*

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

where:

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

and C does not depend on W. In addition, the score matching estimator 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 [4, Thm. 3]. For completeness, we give a proof in Appendix B.3.

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:

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

$$C(s') := \sum_{i=1}^{d_s} \partial_i \psi(s') \partial_i \psi(s')^\top \in \mathbb{R}^{d_\psi \times d_\psi}.$$

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

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 $\hat{V}_n$ and $\hat{b}_n$ as:

$$\hat{V}_n = \sum_{t=1}^{n} \Phi_t C_t \Phi_t^\top, \text{ and } \hat{b}_n = \sum_{t=1}^{n} \Phi_t \xi_t. \tag{6}$$

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 optimization problem is a *quadratic* function in W, which we can solve for via Theorem 4.

However, Theorem 4 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 is perhaps disappointing from a practical perspective, where the dimensionality of ϕ and ψ can be large. Several additional considerations may remedy this:

- 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 $O(n^2 d_S^2)$, [see e.g., 4, Thm. 1]. One can further reduce the dependence on n using Nyström approximations [39].
- As stated before, suppose we are in the 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 $O(d^2)$.

Concentration Guarantee. We are able to provide nonasymptotic concentration guarantees for the score matching estimator $\hat{W}$, under some structural assumptions.

Assumption 2.

(i) 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.

(ii) 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.

(iii) For any $s' \in \mathcal{S}$: $\alpha_1 I \preceq C(s') \preceq \alpha_2 I$.

(iv) For any $(s,a) \in \mathcal{S} \times \mathcal{A}$: $\mathbb{E}_{sa}^{W_0} \psi(s')\psi(s')^\top - \mathbb{E}_{sa}^{W_0}\psi(s')\mathbb{E}_{sa}^{W_0}\psi(s')^\top \preceq \kappa I$.

The conditions in Assumption 2 are mostly adapted from prior work [37, 4, 9], with suitable modifications to accomodate our non-i.i.d. setting. We defer further discussion on it to Appendix B.2 due to space considerations.

We can provide the following concentration guarantee.

Theorem 2. *Under Assumption 1 and 2 and regularity conditions* **(A)-(D)**, *for any* $\delta \in (0,1)$ *and* $\lambda > 0$, *let:*

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

Algorithm 1 Score Matching for RL (SMRL)

1: **Input:** Regularizer λ and constants $B_\psi, B_c, B_\star, \kappa, \alpha_1$.
2: **Initialize:** starting confidence set $\mathcal{W}_1 = \mathbb{R}^{d_\psi \times d_\phi}$, confidence widths $\{\beta_k\}_{k\geq 1}$, dataset $\mathcal{D} = \emptyset$.
3: **for** episode $k = 1, 2, 3, \cdots, K$ **do**
4: **Planning:**
5: Observe initial state s_1^k
6: Choose the optimistic policy: $\pi^k = \arg\max_\pi \max_{W \in \mathcal{W}_k} V^\pi_{\mathbb{P}_W,1}(s_1^k)$
7: **Execution:**
8: 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 $\mathcal{D}$.
9: **Solve for score matching estimator $\hat{W}_k = \arg\min_W \hat{J}(W) + \frac{\lambda}{2}\|W\|_F^2$ via (5)**
10: **Update confidence set $\mathcal{W}_{k+1}$ via (7)**

With probability at least $1-\delta$, the score matching estimators $\hat{W}_n$ satisfy:

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

The proof is provided in Appendix B.4. 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 relies on the method of mixtures argument developed in the linear bandit literature [see e.g., 2, 24].

4 Algorithm and Main Result

In this section, we present our main results, which introduce Algorithm 1 and provide regret guarantees. 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 prediction:

$$\hat{W}_k := \arg\min_W \hat{J}(W) + \frac{\lambda}{2}\|W\|_F^2, \text{ using (5).}$$

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.

Next, we define the confidence set:

$$\mathcal{W}_k := \left\{W \in \mathbb{R}^{d_\psi \times d_\phi} : \left\|\mathsf{vec}\left(\hat{W}_k\right) - \mathsf{vec}\left(W\right)\right\|_{\hat{V}_k+\lambda I} \leq \beta_k\right\}, \tag{7}$$

where

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

and $B_\star$ is some known upper bound on $\|W_0\|_F$.

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 $\mathcal{W}_k$. Implementing (line 6) is known to be NP-hard in the worst case [11]; in this work, we assume computational oracle access to an optimistic planner that solves (line 6) and leave developing efficient approximation algorithms to future work. As long as $W_0 \in \mathcal{W}_k$, the optimistic planning procedure gives us an overestimate of the true value function $V^\star_{\mathbb{P},1}(s_1^k)$, helping ensure 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 $\mathcal{D}$.

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 and and regularity conditions* **(A)-(D)** *hold. Set* $\lambda := 1/B_\star^2$ *and fix* $\delta \in (0,1)$*. Then with probability at least* $1-\delta$*:*

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

where C *is an absolute constant and* $\gamma_{K+1} := \log\det(\lambda^{-1}\hat{V}_{K+1} + I)$*. If we additionally assume that* $\|\phi(s,a)\| \leq B_\phi$*, then* $R(K) \leq \tilde{O}(d_\psi d_\phi \cdot \sqrt{H^3 T})$*, 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 known to take a lower-dimensional form, i.e. if $W_0 = \sum_{i=1}^d \theta_i A_i$, to see that the regret guarantee depends on d instead of $d_\psi \cdot d_\phi$. 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., 36, 22]; 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 [see e.g., 10], then the information gain term can be bounded.

Score Matching vs MLE. 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 (the algorithm of Chowdhury et al.).

In general, Exp-UCRL can be applied to more settings than score matching, due to the fact that score matching requires regularity conditions **(A)-(D)** 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. As such, factored MDP and tabular settings considered by Chowdhury et al. cannot be analyzed via SMRL. In this sense, the scope of SMRL is more limited than that of Exp-UCRL.

Among models given by Assumption 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.

In terms of the final regret guarantee, a detailed comparison of our result with previous results is deferred to Appendix C.3 due to space considerations.

5 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. Exponential family models are a well-studied and powerful statistical framework, but are not yet well-studied within the reinforcement learning theory community. 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.

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] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

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

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

[28] 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.

[29] 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.

[30] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[31] 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.

[32] 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.

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

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

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

[42] 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.

[43] Robert WM Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447, 1974.

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

[45] 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.

[46] Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. *arXiv preprint arXiv:2006.13165*, 2020.

A Related Work

In this section we discuss several related works on provably efficient reinforcement learning and the score matching technique.

Provably Efficient Model-Based RL. In the tabular setting, our theoretical understanding of RL is fairly complete [19, 32, 6, 12, 3]. The best possible regret is $\tilde{\Theta}(\sqrt{H^2SAT})$ [19, 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 [44, 21, 8, 25, 45, 29]. In particular, [44] consider a bilinear transition model; in comparison, our transition model given by Assumption 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, 13, 26, 10, 35] as well as its nonlinear generalization [27, 22]. 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 works study model-free RL by placing structural assumptions on the value function, i.e., assuming bounded Bellman rank [20, 14] or bounded eluder dimension [42]. 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 extremely complex. More relevant to our work are model-based approaches [38, 5]. Sun et al. [38] prove PAC-learning guarantees for model-based RL by relying on a complexity measure called the *witness rank*. 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* [33, 31] of a constructed function class $\mathcal{F}_{\mathcal{P}}$. We note that in both cases, it is unclear whether the relevant complexity measures can be bounded in our setting.

Exponential Families and Score Matching. Exponential families are an important class of parametric models that cover many commonly studied distributions such as the gaussian, exponential, gamma, and Bernoulli distributions. They are an important component for generalized linear models [30, 43]. The parameter of interest can be solved for via MLE; if the normalization constant is not known in closed form, it can be calculated using Markov Chain Monte Carlo methods [7]. Alternatively, the score matching estimator was proposed by Hyvärinen [17, 18] as a generic method for estimating probability densities without needing to compute the normalization constant. For exponential families, the score matching estimator can be efficiently solved for in closed form [18, 37, 4]; in particular, Sriperumbudur et al. [37] show how score matching can be used for density estimation of kernelized exponential families, a problem setting where maximum likelihood estimation fails due to ill-posedness [16].

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 [17, 37, 4] and defer further discussion to those papers.

B.1 Background on Score Matching

We recall some notation.

The model we consider under Assumption 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:

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

We recall the following regularity assumptions, which enable us to employ integration by parts to simplify the score matching objective, [see e.g., 4, Appendix A.4].

(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 $i \in [d_s]$, $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\begin{aligned} \|\partial_i \psi(s')\| &\in L^2(\mathcal{S}, \mathbb{P}_{W_0}(\cdot|s,a)), \\ \left\|\partial_i^2 \psi(s')\right\| &\in L^1(\mathcal{S}, \mathbb{P}_{W_0}(\cdot|s,a)), \\ \|\partial_i \psi(s')\|\, \partial_i \log q(s') &\in L^1(\mathcal{S}, \mathbb{P}_{W_0}(\cdot|s,a)). \end{aligned}$$

We will make a few remarks about these conditions and compare to previous work of Arbel and Gretton [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 may violate the integrability conditions, even if the boundary conditions are met. For *nonnegative distributions*, followup work by Hyvärinen [18] 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* **(A)-(D)**, *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 [17] 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.

B.2 Discussion of Assumption 2

We give some additional background on Assumption 2. (i) and (ii): are crucial in deriving non-asymptotic guarantees for the score matching estimator because they allow for us to bound the ridge regression error in Theorem 2. Bounds on these quantities are already implicit in prior analyses of score matching for the i.i.d. samples setting [37, 4]. (iii): can be viewed as a strengthening of a certain assumption in the i.i.d. setting [37], which assumes that $\mathbb{E}\left[C(s')\right] \geq \alpha_1 I$. In non-i.i.d. setting there is no fixed distribution that we draw s' from, so we replace it with an almost sure bound $C(s') \succeq \alpha_1 I$. Note that when ψ is the identity map, (iii) always holds with $\alpha_1 = \alpha_2 = 1$. (iv): this assumption essentially states that covariance of ψ is bounded when drawn from the conditional distribution $\mathbb{P}(\cdot|s,a)$. It allows us 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 [9].

We now discuss when Assumption 2 holds. It is easy to see that Assumption 2 holds for nonLDS (take $\Sigma = \sigma^2 I$) with $B_\psi = \sigma^{-6}$, $B_c = 0$, $\alpha_1 = \alpha_2 = \sigma^{-4}$, and $\kappa = \sigma^{-2}$. Furthermore, we propose the following setting which goes beyond nonLDS and discuss whether it satisfies Assumption 2:

- $q(s') = \exp(-\alpha^{-1}\sum_{i\in[d_s]}|s'_i|^\alpha)$ for some fixed $\alpha \in (1, 2]$;
- $\psi(s') : \mathbb{R}^{d_s} \to \mathbb{R}^{d_s}$ is an elementwise function $f : \mathbb{R} \to \mathbb{R}$ with $0 < \mu \leq f' \leq L$ and $|f''| \leq D$;
- $\|W_0\| \leq B_\star$ and $\|\phi(s,a)\| \leq B_\phi$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

It is immediate that (ii) holds with $B_c = L^2 B_\star B_\phi$ and (iii) holds with $\alpha_1 = \mu^2$ and $\alpha_2 = L^2$.

To show (i), we need to show that $\sum_{i\in[d_s]} \partial_i \log q(s')\partial_i\psi(s') + \partial_i^2\psi(s')$ is subgaussian. Due to the bounded derivatives of f, it suffices to show that each $\partial_i \log q(s') = -\mathrm{sign}(s'_i)\,|s'_i|^{\alpha-1}$ is subgaussian. While we are unable to mathematically prove this, it seems to hold experimentally for various choices of α and f.

To show (iv), we note that since the coordinates of ψ are independent, the expression reduces to controlling the quantity $\mathrm{Var}(f(s'_i))$, where it is understood that $s' \sim \mathbb{P}_{W_0}(\cdot|s,a)$. Using the lipschitzness of f we get that:

$$\mathrm{Var}(f(s'_i)) \leq \mathbb{E}[(f(s'_i) - f(\mathbb{E}s'_i))^2] \leq L^2\mathrm{Var}(s'_i),$$

so it suffices to show that each s'_i has bounded variance. This is a rather mild condition which seems to hold experimentally for various choices of α and f.

B.3 Proof of Theorem 1

Theorem. *For a dataset $\mathcal{D} = \{(s_t, a_t, s'_t)\}_{t\in[T]}$, we have:*

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

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, the score matching estimator can be computed as:

$$\mathsf{vec}\left(\hat{W}_{n,\lambda}\right) = -(\hat{V}_n + \lambda I)^{-1}\hat{b}_n. \tag{8}$$

Proof. 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:

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

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.4 Proof of Theorem 2

Theorem. *Under Assumption 1 and 2 and regularity conditions* **(A)-(D)**, *for any $\delta \in (0,1)$ and $\lambda > 0$, let:*

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

With probability at least $1-\delta$, the score matching estimators $\hat{W}_n$ satisfy:

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

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

$$\begin{aligned}
\left\|\mathsf{vec}\left(\hat{W}_n\right) - \mathsf{vec}\left(W_0\right)\right\|_{\hat{V}_n+\lambda I} &\le \left\|\hat{b}_n + \hat{V}_n\mathsf{vec}\left(W_0\right)\right\|_{(\hat{V}_n+\lambda I)^{-1}} + \lambda\left\|\mathsf{vec}\left(W_0\right)\right\|_{(\hat{V}_n+\lambda I)^{-1}} \\
&\le \left\|\hat{b}_n + \hat{V}_n\mathsf{vec}\left(W_0\right)\right\|_{(\hat{V}_n+\lambda I)^{-1}} + \sqrt{\lambda}\left\|W_0\right\|_F,
\end{aligned} \tag{9}$$

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

It now remains to bound the first term. Using (6), 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:

$$\begin{aligned}\left\|\hat{b}_n + \hat{V}_n \mathsf{vec}\left(W_0\right)\right\|_{(\hat{V}_n+\lambda I)^{-1}} &= \left\|\sum_{t=1}^{n} \Phi_t \Delta_t\right\|_{(\sum_{t=1}^{n} \Phi_t C_t \Phi_t^\top + \lambda I)^{-1}} \\ &\leq \alpha_1^{-1} \left\|\sum_{t=1}^{n} \Phi_t \Delta_t\right\|_{(\sum_{t=1}^{n} \Phi_t \Phi_t^\top + \alpha^{-1}\lambda I)^{-1}}, \end{aligned} \tag{10}$$

using Assumption 2(iii).

Next we show that Δ_t is conditionally subgaussian with parameter $B_\psi + B_c$. Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be the filtration generated by $\{s_t, a_t\}$. 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 12, we see that Δ_t is a conditionally subgaussian random vector with parameter $B_\psi + B_c$. By Lemma 10, with probability at least $1-\delta$, we get that for all $n \in \mathbb{N}$:

$$\begin{aligned}\left\|\sum_{t=1}^{n} \Phi_t \Delta_t\right\|^2_{(\sum_{t=1}^{n} \Phi_t \Phi_t^\top + \alpha_1^{-1}\lambda I)^{-1}} &\leq 2(B_\psi + B_c) \log \frac{\det(\frac{\alpha_1}{\lambda}\sum_{t=1}^{n} \Phi_t \Phi_t^\top + I)^{1/2}}{\delta} \\ &\leq 2(B_\psi + B_c) \log \frac{\det(\lambda^{-1}\hat{V}_n + I)^{1/2}}{\delta}. \end{aligned} \tag{11}$$

The proof concludes by combining Eqs. 9, 10, and 11.

□

C Regret Guarantee

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 $V_{\hat{W},h}^{k}$ 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 $\mathcal{F} = \{\mathcal{F}_h^k\}_{h\geq 1, k\geq 1}$, where:

$$\mathcal{F}_h^k := \sigma\left(\bigcup_{i\in[H],\ j\in[k-1]} \{s_i^j, a_i^j\} \cup \bigcup_{i\in[h],\ j=k} \{s_i^j, a_i^j\}\right),$$

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 $|m^k_h| \le 2H$.

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. □

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. [9].

Lemma 6 (KL Divergence Bound)**.** *Under Assumption 1 and 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*

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

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:

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

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 11 gives us expressions for $\nabla Z_{sa}(\cdot)$ and $\nabla^2 Z_{sa}(\cdot)$, which we can plug in to get:

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

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):

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

Therefore, telescoping the sum we have:

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

□

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 $\pi^k = \arg\max_\pi \max_{W \in \mathcal{W}_k} V^\pi_{W,1}(s^k_1)$. Let us denote $\tilde{W}_k = \arg\max_{W \in \mathcal{W}_k} V^k_{W,1}(s^k_1)$, that is, $\tilde{W}$ is the "optimistic model" which π^k is greedy with respect to. Under event $\mathcal{E}$, we have $V^\star_{W_0,1}(s^k_1) \le V^k_{\tilde{W}_k,1}(s^k_1)$, so we have:

$$\mathcal{R}(K) = \sum_{k=1}^{K} \left[V^\star_{W_0,1}(s^k_1) - V^k_{W_0,1}(s^k_1)\right] \le \sum_{k=1}^{K} \left[V^k_{\tilde{W}_k,1}(s^k_1) - V^k_{W_0,1}(s^k_1)\right].$$

Now we invoke Lemma 5 to get

$$\mathcal{R}(K) \le \sum_{k=1}^{K}\sum_{h=1}^{H} \left[\mathbb{E}^{\tilde{W}_k}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s') - \mathbb{E}^{W_0}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s')\right] + \sum_{k=1}^{K}\sum_{h=1}^{H} m^k_h. \tag{12}$$

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

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

with probability at least $1 - \delta/2$. We now proceed to bound the first term of (12). For every $(k,h) \in [K] \times [H]$:

$$\begin{aligned}\mathbb{E}^{\tilde{W}_k}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s') - \mathbb{E}^{W_0}_{s^k_h a^k_h} V^k_{\tilde{W}_k,h+1}(s') &\le H \cdot \mathrm{TV}\left(\mathbb{P}_{W_0}(\cdot|s^k_h, a^k_h,)\|\mathbb{P}_{\tilde{W}_k}(\cdot|s^k_h, a^k_h)\right) \\ &\le H \cdot \min\left\{\sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathbb{P}_{W_0}(\cdot|s^k_h, a^k_h)\|\mathbb{P}_{\tilde{W}_k}(\cdot|s^k_h, a^k_h)\right)}, 1\right\},\end{aligned} \tag{13}$$

where the first inequality uses the TV bound (Lemma 13) 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

$$\begin{aligned}\sqrt{D_{\mathrm{KL}}\left(\mathbb{P}_{W_0}(\cdot|s^k_h, a^k_h)\|\mathbb{P}_{\tilde{W}_k}(\cdot|s^k_h, a^k_h)\right)} &\le \sqrt{\frac{\kappa}{2}}\left\|\mathsf{vec}\left(W_0\right) - \mathsf{vec}\left(\tilde{W}_k\right)\right\|_{\Phi_{kh}\Phi^\top_{kh}} \\ \le \sqrt{\frac{\kappa}{2}}\left(\left\|\mathsf{vec}\left(W_0\right) - \mathsf{vec}\left(\hat{W}_k\right)\right\|_{\Phi_{kh}\Phi^\top_{kh}} + \left\|\mathsf{vec}\left(\hat{W}_k\right) - \mathsf{vec}\left(\tilde{W}_k\right)\right\|_{\Phi_{kh}\Phi^\top_{kh}}\right),\end{aligned} \tag{14}$$

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:

$$
\begin{aligned}
&\left\|\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right\|^2_{\Phi_{kh}\Phi_{kh}^\top} \\
&= \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right)^\top \Phi_{kh}\Phi_{kh}^\top \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right) \\
&\leq \alpha_1^{-1} \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right)^\top \Phi_{kh}C_{kh}\Phi_{kh}^\top \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right) \\
&\leq \alpha_1^{-1} \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right)^\top (\hat{V}_k + \lambda I)^{1/2}(\hat{V}_k + \lambda I)^{-1/2}\Phi_{kh}C_{kh}\Phi_{kh}^\top \\
&\quad (\hat{V}_k + \lambda I)^{-1/2}(\hat{V}_k + \lambda I)^{1/2} \left(\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right) \\
&\leq \alpha_1^{-1} \left\|\mathsf{vec}\left(W\right) - \mathsf{vec}\left(\hat{W}_k\right)\right\|^2_{\hat{V}_k+\lambda I} \cdot \left\|(\hat{V}_k + \lambda I)^{-1/2}\Phi_{kh}C_{kh}\Phi_{kh}^\top(\hat{V}_k + \lambda I)^{-1/2}\right\| \\
&\leq \alpha_1^{-1}\beta_k^2 \cdot \left\|(\hat{V}_k + \lambda I)^{-1/2}\Phi_{kh}C_{kh}\Phi_{kh}^\top(\hat{V}_k + \lambda I)^{-1/2}\right\| \\
&\leq \alpha_1^{-1}\beta_{K+1}^2 \cdot \left\|(\hat{V}_k + \lambda I)^{-1/2}\Phi_{kh}C_{kh}\Phi_{kh}^\top(\hat{V}_k + \lambda I)^{-1/2}\right\|, \qquad (15)
\end{aligned}
$$

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 (13), (14), and (15), we bound the first term of (12) 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:

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

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. □

C.3 Details of Comparison to Prior Work

In this section, we provide further details about our comparison to Chowdhury et al. and Kakade et al.. First, we translate the quantities in Thm. 2 of 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 [9])**.** *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:*

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

From here, it is easy to specialize the results to the nonLDS setting (with bounded rewards and feature vectors ϕ), where $\Sigma = \sigma^2 I$ and $\lambda := \sigma^2/\|W_0\|^2$ to get the bounds:

- (Our Result): $\tilde{O}\left(\sqrt{d_\phi d_\psi \cdot (\sigma^4 d_\phi d_\psi + H) H^2 T}\right)$.
- (Chowdhury et al.): $\tilde{O}\left(\sqrt{d_\phi d_\psi\left(\sigma^2 + d_\phi d_\psi\right)\left(1 + \sigma^{-4} B_\star^2 B_\phi^2 H\right) H^2 T}\right)$.
- (Kakade et al.): $\tilde{O}(\sqrt{d_\phi(d_\phi + d_\psi + H) H^2 T})$.

We conjecture that the σ^2 dependence in our regret guarantee can be removed by shaving an extra factor of α_1 to get a regret guarantee that doesn't depend polynomially on the noise level.

D Technical Results

Lemma 10 (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 $t \ge 1$:

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

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 [see e.g., 24, 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 $\delta \in (0,1)$ [24, Thm. 3.9]:

$$\mathbb{P}\left(\sup_{t\in\mathbb{N}} \overline{M}_t \ge \frac{1}{\delta}\right) \le \delta \tag{16}$$

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

$$\begin{aligned}
\overline{M}_n &= \int \exp\left\{\frac{1}{\sigma}\gamma^\top S_n - \frac{1}{2}\gamma^\top V_n \gamma\right\} f(\gamma) d\gamma \\
&= \frac{1}{(2\pi)^{m/2}\det V_0^{-1/2}} \int \exp\left\{\frac{1}{\sigma}\gamma^\top S_n - \frac{1}{2}\gamma^\top V_n \gamma - \frac{1}{2}\gamma^\top V_0 \gamma\right\} d\gamma \\
&= \frac{\exp\{\frac{1}{2\sigma^2} S_n^\top (V_n+V_0)^{-1} S_n\}}{(2\pi)^{m/2}\det V_0^{-1/2}} \int \exp\left\{-\frac{1}{2}\|\gamma - (V_n+V_0)^{-1}\frac{S_n}{\sigma}\|^2_{V_n+V_0}\right\} d\gamma \\
&= \frac{\det V_0^{1/2}}{\det(V_n+V_0)^{1/2}} \exp\left\{\frac{1}{2\sigma^2} S_n^\top (V_n+V_0)^{-1} S_n\right\}.
\end{aligned} \tag{17}$$

The last equality is due to integration of the Gaussian pdf. Therefore, by combining (16) and (17) we obtain that with probability $1-\delta$, for all $n \in \mathbb{N}$:

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

□

Lemma 11 (Log Partition Derivatives, [9])**.** *The first and second derivatives of the exponential family model in Assumption 1 are:*

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

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

Lemma 12. *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 $1/p+1/q=1$:

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

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. □

Lemma 13. *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

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

□