Reinforcement Learning in Parametric MDPs with Exponential Families

Sayak Ray Chowdhury
Department of ECE
Indian Institute of Science
Bengaluru, India

Aditya Gopalan
Department of ECE
Indian Institute of Science
Bengaluru, India

Odalric-Ambrym Maillard
Univ. Lille, Inria, CNRS, ECL
UMR 9189 – CRISiAL
F-59000 Lille, France

Abstract

Extending model-based regret minimization strategies for Markov decision processes (MDPs) beyond discrete state-action spaces requires structural assumptions on the reward and transition models. Existing parametric approaches establish regret guarantees by making strong assumptions about either the state transition distribution or the value function as a function of state-action features, and often do not satisfactorily capture classical problems like linear dynamical systems or factored MDPs. This paper introduces a new MDP transition model defined by a collection of linearly parameterized exponential families with $d$ unknown parameters. For finite-horizon episodic RL with horizon $H$ in this MDP model, we propose a model-based upper confidence RL algorithm (Exp-UCRL) that solves a penalized maximum likelihood estimation problem to learn the $d$-dimensional representation of the transition distribution, balancing the exploitation-exploration tradeoff using confidence sets in the exponential family space. We demonstrate the efficiency of our algorithm by proving a frequentist (worst-case) regret bound that is of order $\tilde{O}(d^{1/2}H^{3/2}N)$, sub-linear in total time $N$, linear in dimension $d$, and polynomial in the planning horizon $H$. This is achieved by deriving a novel concentration inequality for conditional exponential families that might be of independent interest. The exponential family MDP model also admits an efficient posterior sampling-style algorithm for which a similar guarantee on the Bayesian regret is shown.

1 LINEARITY IN REINFORCEMENT LEARNING

We consider episodic reinforcement learning (RL) in a finite horizon Markov decision process (MDP) \cite{Puterman1994, Sutton1988}, with (possibly infinite) state and action spaces $\mathcal{S}$ and $\mathcal{A}$, respectively, reward function $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and parametric state transition distribution $P_\theta : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ for some underlying parameter $\theta^* \in \mathbb{R}^d$ and episode length $H$. Very large or infinite state and/or action spaces make RL a challenging task, especially in terms of generalising learnt knowledge across unseen states and actions. In this paper, we explore how to endow an MDP with an appropriate linear structure in order to obtain algorithms with guarantees of low regret.

Linearity is a natural structural assumption when considering a function defined on a large set. For instance, in linear regression \cite{SeberLee2012}, the target mean function $f : \mathcal{X} \to \mathbb{R}$ is assumed to be of the form $f(\cdot) = \theta^T \varphi(\cdot)$ where $\theta \in \mathbb{R}^d$ is a vector of unknown parameters and $\varphi : \mathcal{X} \to \mathbb{R}^d$ is a known feature function. Using a generic $\varphi$ function allows for great flexibility and the encoding of specific expert domain knowledge, which explains the popularity of this model in machine learning and statistics. Besides, this model is powerful in the sense that it can be extended from finite dimension $d$ to infinite dimensions by appealing to the theory of Reproducing Kernel Hilbert Spaces (RKHS) \cite{PaulsenRaghupathi2016}.

Linearity in Bandits For stateless MDPs or multi-armed bandits, linear models have been widely studied in a number of works, see \cite{AbbasiYadkori2011, RusmevichientongTsitsiklis2010, LattimoreSzepesvari2016}, as well as \cite{Durand2017, Filippi2010} to cite a few, exploiting the connection with linear regression. In particular, the construction of finite-time confidence ellipsoids for the unknown vector parameter $\theta$ in the challenging context of bandit sampling that involves random stopping
times is now a popular tool (which has been extended also to RKHS, see Srivivas et al. (2010); Durand et al. (2017); Chowdhury and Gopalan (2017)).

**Linearity in MDPs**  Several studies have considered the task of regret minimization in tabular MDPs in the *episodic* setting, with a fixed and known horizon; see, e.g., Osband et al. (2013); Gheshlaghi Azar et al. (2017); Dann et al. (2017); Efroni et al. (2019); Zanette and Brunskill (2019). The work of Gopalan and Mannor (2015) consider a generalized MDP formulation, yet their results are restricted to the case of finite-state, finite-action MDPs. For this reason, their results are not clearly stronger and more general than the ones we provide. Other approaches have been introduced to extend the popular UCRL2 approach from Jaksch et al. (2010) to handle continuous MDPs, which has benefited immensely from what is arguably one of the most popular methods to linearly parameterize families of probability distributions — exponential families (Amari, 1997). Our main proposal in this paper is to consider an exponential family formulation of the MDP transition kernel, essentially assuming log $P$, rather than $P$ to be bilinear. More precisely, we assume the following bilinear exponential family model:

$$
Z_{s,a}(\theta) = \log \int_S \exp(\psi(s) \top M \varphi(s,a)) h(s',s,a) ds',
$$

(1)

at every $s,a \in S,A$. Here $h, \varphi$ and $\psi$ are known feature functions, and $M$ is a $p \times q$ matrix of unknown parameters. Note that $h, \psi$ and $\varphi$ cannot depend on $\theta$. Further, since considering all entries of the matrix $M$ as unrelated parameters may prevent one from encoding stronger structure, we consider that it is of the form $M = \sum_i \theta_i A_i$, where $\theta = (\theta_i)_{i \leq d} \in \mathbb{R}^d$ is the vector of unknown parameters, and each $A_i$ is a known $p \times q$ matrix. We recover the case of a fully unknown matrix $M_0$ by considering $d = pq$ and the $(A_i)_{i \leq d}$ to be a one-hot encoding, but this additional flexibility enables to capture situations when several entries of the matrix must have same value. With this formulation, for each $(s,a)$, we get a linear model with feature function $s' \mapsto (\psi(s') \top A, \varphi(s,a))_{i \leq d}$ and unknown parameter $\theta \in \mathbb{R}^d$, while for each $s'$, we get a linear model with feature function $(s,a) \mapsto (\psi(s') \top A, \varphi(s,a))_{i \leq d}$ and the same unknown parameter. We detail popular classes of MDP dynamics that fit this formulation in Section 4.

**Learning Agent and Regret**  The learning agent interacts with the MDP in episodes and, at each episode $t$, a trajectory $(s_t^1, a_t^1, r_t^1, \ldots, s_{t+1}^H, a_{t+1}^H, r_{t+1}^H, s_{t+1}^H)$ is generated. Here $a_t^h$ denotes the action taken at state $s_t^h, r_t^h = R(s_t^h, a_t^h)$ denotes the immediate reward, and $s_{t+1}^h \sim P_\theta(\cdot | s_t^h, a_t^h)$ denotes the random next state. The initial state $s_t^1$ is assumed to be fixed and history independent. The actions are chosen following some policy $\pi = (\pi_1, \ldots, \pi_H)$, where each $\pi_t$ is a mapping from the state space $S$ into the action space $A$. The agent would like to find a policy $\pi$ that maximizes the long term expected reward starting from every state $s \in S$ and every step $h \in [H]$, defined as

$$
V_{\theta, h}^\pi(s) = \mathbb{E}_{\theta^*} \left[ \sum_{j=h}^{H} R(s_j, \pi_j(s_j)) | s_h = s \right]
$$

We call $V_{\theta, h}^\pi : S \rightarrow \mathbb{R}$ the value function of policy $\pi$ at step $h$. The subscript $\theta^*$ refers to the bilinear exponential family transition dynamics parameterized by $\theta^* \in \mathbb{R}^d$. We assume that the agent, while not knowing $\theta^*$, knows the matrices $A_1, \ldots, A_H$ and the reward function $R$.

A policy $\pi^*$ is said to be optimal if $V_{\theta^*, h}^\pi(s) = \max_{\pi \in \Pi} V_{\theta^*, h}^\pi(s)$ for all $s \in S$ and $h \in [H]$, where $\Pi$ is the set of all non-stationary policies. (Since the episode length is finite, such a policy exists when the action space $A$ is also finite.) We measure performance

\footnote{We write probability measures assuming that they have a density mainly for convenience; the development can easily be extended to general probability transition measures.}

\[\text{times is now a popular tool (which has been extended also to RKHS, see Srivivas et al. (2010); Durand et al. (2017); Chowdhury and Gopalan (2017)).} \]

\[\text{Linearity in MDPs} \quad \text{Several studies have considered the task of regret minimization in tabular MDPs in the *episodic* setting, with a fixed and known horizon; see, e.g., Osband et al. (2013); Gheshlaghi Azar et al. (2017); Dann et al. (2017); Efroni et al. (2019); Zanette and Brunskill (2019). The work of Gopalan and Mannor (2015) consider a generalized MDP formulation, yet their results are restricted to the case of finite-state, finite-action MDPs. For this reason, their results are not clearly stronger and more general than the ones we provide. Other approaches have been introduced to extend the popular UCRL2 approach from Jaksch et al. (2010) to handle continuous MDPs, which has benefited immensely from what is arguably one of the most popular methods to linearly parameterize families of probability distributions — exponential families (Amari, 1997). Our main proposal in this paper is to consider an exponential family formulation of the MDP transition kernel, essentially assuming log $P$, rather than $P$ to be bilinear. More precisely, we assume the following bilinear exponential family model:}

\[Z_{s,a}(\theta) = \log \int_S \exp(\psi(s) \top M \varphi(s,a)) h(s',s,a) ds',\]

(1)

\[at every s,a \in S,A. Here h, \varphi and \psi are known feature functions, and M is a p \times q matrix of unknown parameters. Note that h, \psi and \varphi cannot depend on \theta. Further, since considering all entries of the matrix M as unrelated parameters may prevent one from encoding stronger structure, we consider that it is of the form M = \sum_i \theta_i A_i, where \theta = (\theta_i)_{i \leq d} \in \mathbb{R}^d is the vector of unknown parameters, and each A_i is a known p \times q matrix. We recover the case of a fully unknown matrix M_0 by considering d = pq and the (A_i)_{i \leq d} to be a one-hot encoding, but this additional flexibility enables to capture situations when several entries of the matrix must have same value. With this formulation, for each (s,a), we get a linear model with feature function s' \mapsto (\psi(s') \top A, \varphi(s,a))_{i \leq d} and unknown parameter \theta \in \mathbb{R}^d, while for each s', we get a linear model with feature function (s,a) \mapsto (\psi(s') \top A, \varphi(s,a))_{i \leq d} and the same unknown parameter. We detail popular classes of MDP dynamics that fit this formulation in Section 4.\]

**Learning Agent and Regret**  The learning agent interacts with the MDP in episodes and, at each episode t, a trajectory (s_t^1, a_t^1, r_t^1, \ldots, s_{t+1}^H, a_{t+1}^H, r_{t+1}^H, s_{t+1}^H) is generated. Here a_t^h denotes the action taken at state s_t^h, r_t^h = R(s_t^h, a_t^h) denotes the immediate reward, and s_{t+1}^h \sim P_\theta(\cdot | s_t^h, a_t^h) denotes the random next state. The initial state s_t^1 is assumed to be fixed and history independent. The actions are chosen following some policy \pi = (\pi_1, \ldots, \pi_H), where each \pi_t is a mapping from the state space S into the action space A. The agent would like to find a policy \pi that maximizes the long term expected reward starting from every state s \in S and every step h \in [H], defined as

\[V_{\theta, h}^\pi(s) = \mathbb{E}_{\theta^*} \left[ \sum_{j=h}^{H} R(s_j, \pi_j(s_j)) | s_h = s \right],\]

We call \[V_{\theta, h}^\pi : S \rightarrow \mathbb{R}\] the value function of policy \pi at step h. The subscript \theta^* refers to the bilinear exponential family transition dynamics parameterized by \theta^* \in \mathbb{R}^d. We assume that the agent, while not knowing \theta^*, knows the matrices A_1, \ldots, A_H and the reward function R.

A policy \pi^* is said to be optimal if \[V_{\theta^*, h}^\pi(s) = \max_{\pi \in \Pi} V_{\theta^*, h}^\pi(s)\] for all s \in S and h \in [H], where \Pi is the set of all non-stationary policies. (Since the episode length is finite, such a policy exists when the action space A is also finite.) We measure performance
of the agent by the cumulative (pseudo) regret accumulated over $T$ episodes, defined as
\[
R(N) = \sum_{t=1}^{T} \left[ V_{\pi_t}^\varphi(s_t^i) - V_{\pi_t^\star}^\varphi(s_t^i) \right],
\]
where $N = TH$ is the total number of steps. Intuitively, this is a measure of the cumulative difference in values due to not knowing the optimal policy $\pi^\star$ beforehand and instead using some policy $\pi_t$ in episode $t$ starting from some fixed initial state $s_t^i$. We seek algorithms with regret that is sublinear in $N$, which demonstrates the agent’s ability to act near optimally.

**Outline and Contribution** We detail useful properties about maximum likelihood estimation for this exponential family setup in Section 2. We then derive a novel concentration inequality for exponential families, generalizing the popular method of mixtures technique for sub-Gaussian random variables. We introduce in Section 3 the Exp-UCRL strategy for efficient regret minimization in the context of MDPs with such dynamics, and provide its regret guarantee in Theorem 2. In Section 4 we show that this model enables to capture large classes of MDPs, including linear dynamical systems used in the control literature, and factored and tabular models as special case. We conclude the paper with a sketch of proof highlighting the main steps leading to Theorem 2.

## 2 EXPONENTIAL FAMILIES FOR TRANSITION DYNAMICS

The benefit of modelling transition kernels as exponential families is that one may benefit from numerous, well-known properties of exponential families, relating the log-partition function $Z_{s,a}$ to the mean, variance, maximum likelihood or Kullback-Leibler (KL) divergence. Indeed, it is easily checked (see Appendix A for completeness) that
\[
\nabla_i Z_{s,a}(\theta) = \mathbb{E}_{s,a}^\theta(\psi(s'))^\top A_i \phi(s,a),
\]
\[
\nabla_{i,j}^2 Z_{s,a}(\theta) = \phi(s,a)^\top A_i^\top C_{s,a}^\theta(\psi(s')) A_j \phi(s,a),
\]
KL$_{s,a}$($\theta, \theta'$) = $Z_{s,a}(\theta') - Z_{s,a}(\theta) - (\theta' - \theta)^\top \nabla Z_{s,a}(\theta) \] , where $\mathbb{E}_{s,a}^\theta$, $C_{s,a}^\theta$ denote the expectation and covariance operator for the probability distribution $P_{\theta}(\cdot | s,a)$, and KL$_{s,a}$($\theta, \theta'$) denotes the Kullback-Leibler divergence b/w $P_{\theta}(\cdot | s,a)$ and $P_{\theta'}(\cdot | s,a)$. For the matrix norm notation to be justified, we further require that for each $s,a$, the matrix $\nabla^2 Z_{s,a}$ that is symmetric is also positive definite. Now, considering a sequence of observations $(s_t,a_t,s_t^i)_{t \leq n}$, where for each $t$, $s_t^i \sim P_{\theta_t}(\cdot | s_t,a_t)$, and any differentiable penalty function $\text{pen}(\cdot)$, a solution to the penalized maximum-likelihood problem with regularization parameter $\eta \in \mathbb{R}^+$ must satisfy
\[
\theta_n \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{t=1}^{n} -\log P_{\theta}(s_t^i | s_t, a_t) + \eta \text{pen}(\theta) \implies \forall i \leq d,
\]
\[
\frac{1}{n} \sum_{t=1}^{n} (\psi(s_t^i) - \mathbb{E}_{s_t,a_t}^\theta(\psi(s_t^i)))^\top A_i \phi(s_t,a_t) = \eta \nabla_i \text{pen}(\theta_n). \quad (2)
\]

In this paper, we choose a trace-norm penalty $\text{pen}(\theta) = \frac{1}{2} \| \theta \|_A^2$, where $A$ denotes the matrix with entries $A_{i,j} = \text{tr}(A_i A_j^\top)$, $i,j \leq d$. We assume that $A$ is invertible. A solution to equation 2 may be obtained in closed form in some cases, e.g. when densities are Gaussian. For generic features, one should resort to specific schemes, involving Monte-Carlo computations of the integrals, see [Brooks et al., 2011]. In Section 4 we detail examples of MDPs, including a specialization of [2].

We now present the following key result, which is a novel generalization of the Laplace method for Gaussian or sub-Gaussian random variables [Peña et al., 2008] to exponential families. The complete proof is provided in Appendix B.

**Theorem 1** (Laplace concentration for Exponential families). Suppose $\{F_t\}_{t=0}^\infty$ is a filtration such that for each $t$, (i) $s_t^i \sim F_t$-measurable, (ii) $(s_t,a_t)$ is $F_t$-measurable, and (iii) given $(s_t,a_t)$, $s_t^i \sim P_{\theta_t}(\cdot | s_t,a_t)$ according to the exponential family defined by (1). Let $\theta_n$ be the penalized MLE defined by (2), and let $Z_{s,a}(\theta)$ be strictly convex in $\theta$ for all $(s,a)$. Then, for any $\delta \in (0,1)$, with probability at least $1 - \delta$, the following holds uniformly over all $n \in \mathbb{N}$:
\[
\sum_{t=1}^{n} \text{KL}_{s,a}(\theta_n, \theta^\star) + \frac{\eta}{2} \| \theta^\star - \theta_n \|_A^2 - \frac{\eta}{2} \| \theta^\star \|_A^2 \leq \log \left( \frac{C_{A,n}}{\delta} \right),
\]
where $C_{A,n} = \frac{\int_{\mathbb{R}^d} \exp \left( -\frac{\eta}{2} \| \theta \|^2 \right) d\theta}{\int_{\mathbb{R}^d} \exp \left( -\sum_{t=1}^{n} \text{KL}_{s,a}(\theta_n, \theta^\star) - \frac{\eta}{2} \| \theta^\star - \theta_n \|_A^2 \right) d\theta}$.

Furthermore, introducing the matrix $(G_{s,a})_{i,j} = \phi(s,a)^\top A_i^\top A_j \phi(s,a)$, $\forall i,j \leq d$, we have
\[
C_{A,n} \leq \det \left( I + \beta \eta^{-1} A^{-1} \sum_{t=1}^{n} G_{s_t,a_t} \right),
\]
where $\beta = \sup_{\theta,s,a} \lambda_{\text{max}} \left( C_{s,a}^\theta(\psi(s')) \right)$.

This is a rather general concentration inequality that helps to design confidence sets for adaptive regression in conditional exponential families. It generalizes many previously known results for adaptive estimation, including linear bandits [Abbasi-Yadkori et al., 2011] (since KL is the Euclidean distance), GLM bandits [Filippi et al., 2010] (via bounding from below the Hessian of the log-partition function). Importantly, it preserves the information (KL divergence) geometry

\[\text{Strict convexity essentially amounts to a minimal representation of the exponential family [Amar 1997]. We assume it for brevity; the result holds even if } Z_{s,a} \text{ is only convex.}\]
of the exponential families in the sense of measuring empirical deviations in the KL divergence, and this is crucially exploited later to make the learning algorithm not require knowledge of the minimum curvature.

Proof sketch. For the conditional exponential family, a direct computation shows that the KL divergence $b/w P_0(\cdot|s,a)$ and $P_0(\cdot|s,a)$ can be expressed as a Bregman divergence of the log-partition function $Z_{s,a}$ with the parameters reversed, i.e., $B_{Z_{s,a}}(\theta', \theta) = KL_{s,a}(\theta, \theta')$. Now, for any fixed $\lambda \in \mathbb{R}^d$, we introduce the function $B_{Z_{s,a}(\theta^* + \lambda, \lambda)}$ and define

$$M_n^\lambda = \exp \left( \lambda^T S_n - \sum_{t=1}^n B_{Z_{s,a}(\theta^* + \lambda)}(\lambda) \right),$$

where $(S_n)_i = \sum_{t=1}^n \psi(s_t') - \mathbb{E} \psi(s_t'|a_t, t, \lambda), \forall i \leq d$. Since, by construction, $\log \mathbb{E} [\exp(\lambda^T S_n)] = \lambda^T S_n - \sum_{t=1}^n B_{Z_{s,a}(\theta^* + \lambda)}(\lambda)$, $M_n^\lambda$ is a non-negative martingale such that $E[M_n^\lambda] = 1$. Further, we show that for any random stopping time $N$, $\mathbb{E}[M_n^\lambda] \leq 1$.

We now apply the method of mixtures technique [Peña et al., 2006] by integrating over $\lambda$. To this end, for any prior density $q(\theta)$ for $\theta$, we define a mixture of marginals $M_\lambda = \int_{\mathbb{R}^d} M_n^\lambda q(\theta^* + \lambda) d\lambda$ so that $\mathbb{E}[M_n] = 1$. Considering the prior density $N(0, \eta A^{-1})$, we then show that

$$M_n = \exp \left( \sum_{t=1}^n B_{Z_{s,a}(\theta^* + \theta_n)} + \frac{\eta}{2} ||\theta^* - \theta_n||_2^2 \right) / C_{\lambda, n}.$$

We then deduce from a simple Markov inequality that

$$\Pr \left[ \sum_{t=1}^n B_{Z_{s,a}(\theta^* + \theta_n)} + \frac{\eta}{2} ||\theta^* - \theta_n||_2^2 \geq \log(\frac{C_{\lambda, n}}{\delta}) \right] \geq \delta \cdot \mathbb{E}[M_n] \leq \delta, \forall \delta \in (0, 1).$$

By the properties of the martingale $M_n$, this also holds for any random stopping time $N$. The proof is completed using a stopping time construction similar to that of [Abbasi-Yadkori et al., 2011].

3 REGRETS MINIMIZATION IN BILINEAR EXPONENTIAL MDPS

We now introduce a low-regret algorithm inspired by the popular upper confidence RL (UCRL) strategy applied to our bilinear exponential family model [1], and present a regret minimization guarantee for it in Theorem 2.

3.1 Exponential Family UCRL Algorithm

In order to address the exploration-exploitation trade-off, the proposed algorithm maintains both an empir-

Algorithm 1 Exponential Family Upper Confidence RL (Exp-UCRL)

**Input**: Matrices $A_1, \ldots, A_d$, constant $B_A$, parameters $\delta \in (0, 1], \eta > 0$, for episode $t=1, 2, 3, \ldots$ do

```
Set $n = (t-1)H$.
Compute the penalized ML estimate $\hat{\theta}_n$ and the confidence set $\Theta_n$.
Observe initial state $s_1$.
Choose policy $\pi_1 = \arg \max_{\theta \in \Theta_n} V_{\hat{\theta}_n}(s_1)$.
for period $h = 1, 2, 3, \ldots, H$ do

```

A key result of this paper is the following theoretical guarantee on the regret minimization properties of Exp-UCRL.

**Theorem 2** (Regret bound for Exp-UCRL). Let $A_{ij} = \text{tr}(A_i A_j^\top)$ and $(G_{s,a})_{i,j} = \psi(s, a)^\top A_i \psi(s, a), \forall i, j \leq d$. Assume that $\|\theta\|_A \leq B_A$ and $\|A^{-1}G_{s,a}\| \leq B_A$ for all $(s, a)$. Then, for any $\eta > 0$ and $\delta \in (0, 1]$, Exp-UCRL enjoys, with probability at least $1-\delta$, the
dependence on regret bound for generalized linear bandits, where the exponential family, our result implies an improved representation of the exponential family. A minimal partition function for each $s,a$ is the number of unknown parameters.

Remark 1. In the special case of a bandit setting $(|S| = 1)$ with each action’s reward being distributed as an exponential family, our result implies an improved regret bound for generalized linear bandits, where the dependence on $\alpha$ scales only as $1/\alpha$ compared to the $1/\alpha$ scaling given in Filippi et al. (2010). (We ignore the $1/\alpha$ scaling in the lower order term of the regret bound, where the dependency with $N$ is only logarithmic.) This improvement is consistent with a conjecture of Filippi et al. (2010), and we achieve this thanks to the novel concentration inequality for conditional exponential families (Theorem 2) that directly controls deviations of estimates in the KL geometry rather than Euclidean-type metrics.

Remark 2. We prove Theorem 2 assuming a minimal representation of the exponential family. A minimal representation amounts to assuming strict convexity of the log partition function for each $(s,a)$ and thus, in turn, $\alpha > 0$. If $\alpha = 0$, then the log-partition function is not strictly convex at some state-action pair $(s,a)$; this is akin to some non-minimality in the exponential family representation (e.g., degenerate multivariate Gaussian). Assuming minimality is a restriction but is quite common when dealing with exponential family models. We believe Theorem 2 can somewhat be extended to non-minimal families, but this would require specific care and technicalities that might hinder the paper’s clarity.

Theorem 2 yields a $\hat{O}(d\sqrt{H^3N})$ regret bound in the linear exponential MDP setting, where $d$ is the number of model parameters, $H$ is the episode length and $N$ is the total time. It is worth noting that the regret bound does not explicitly depend on the size (cardinality) of $S$ and $A$, which is crucial in the large state-action space setting that entails function approximation. For simplicity of representation, we have assumed that the reward function $R$ is known. When $R$ is unknown but satisfies a linear structure with $d$ unknown parameters, our algorithm can be extended naturally with an optimistic reward estimation step at each episode, similar to that for the linear bandit setting (Abbasi-Yadkori et al., 2011). This would add an additional $O(d\sqrt{N})$ term in the regret bound. We now discuss the dependence of regret on key problem parameters against the backdrop of existing work.

On the Dependency on $H$ (Time Horizon) Yang and Wang (2019) assume a bilinear transition probability model with a matrix factorization of the form $P(s'|s,a) = \psi(s')^T M \varphi(s,a)$, and propose a model-based algorithm with regret $\hat{O}(\sqrt{dH^4N})$ in general, where $d$ is the dimension of state-action features $\varphi(s,a)$. Jin et al. (2019) study a similar class of linear MDPs with the transition probabilities being linear in state-action features $\varphi(s,a)$, proposing a model-free least-squares value iteration algorithm that achieves a regret bound of order $\hat{O}(\sqrt{dH^3N})$. A similar regret guarantee is also established by Wang et al. (2020) in the context of generalized linear MDPs. In this work, the authors essentially assume that any value function arising from an optimistic value iteration step can be represented as a generalized linear function of the features. We, however, do not need to put any prior knowledge on the value function and only need an assumption on the transition structure. We believe that it is more natural in practice to impose structural assumptions on the transition model than on (future) value functions, which are generally complex, derived functions of the reward and transition structures. Moreover, all these models are rather limited in the range of well-known MDPs that they can capture; apart from the tabular model, it is unclear if they can express classical continuous-space models like the linear dynamical system or even the factorized MDPs $^3$ We consider a more natural and flexible transition model, involving exponential families, than these prior works. Though it is incommensurable with the above works in general, we note that our regret bound reduces a $\sqrt{H}$ factor as compared to Yang and Wang (2019), while achieving the same scaling with $H$ as in Jin et al. (2019). To provide further insights on the dependency of our bound on $H$, let us consider the case of finite-horizon, tabular MDP learning. In this case, the best known regret achieved by model-based methods is $\hat{O}(\sqrt{HSAN})$ (Gheshlaghi Azar et al., 2017; Zanette and Brunskill, 2019; Efroni et al., 2019) whereas the best known regret for model-free learning is $\hat{O}(\sqrt{H^2SAN})$ (Jin et al., 2018). Our algorithm’s regret scaling with $H$ is similar to the latter, but with the advantage of being able to handle models much more general than just tabular MDPs.

On the Dependency on $d$ (Number of Unknown Parameters) We first note that the regret bounds in prior work on linear MDP models (Yang and Wang, 2019; Jin et al., 2019; Wang et al., 2020) as stated previously, depend on $d$ as $\hat{O}(d^{3/2})$; however, the ap-

$^3$For the factored MDPs, in fact, model-based algorithms are exponentially better (in terms of sample complexity) than model free methods under a certain realizability condition (Sun et al., 2019).
and Van Roy (2014a) analyze the regret guarantee for H transitions, the bandits do not, and a naive adaptation assumption over the transitions, and still achieve arguably the most natural, yet expressive enough, linear assumption over the transitions, and still achieve a similar regret scaling of the prior works.

Complexity of Optimistic Planning Exact optimistic planning as prescribed in Exp-UCRL may be computationally intractable, so it is common to assume access to an oracle which returns an ε-optimal solution to (3). Now, setting ε = √H/t at episode t, we can ensure that this adds only an additional O(√N) factor in the regret bound. We note here that the design of such approximate MDP planners or oracles for continuous state and action spaces is a subject of active research, whereas our focus in this work is chiefly on the statistical efficiency of algorithms for achieving low regret.

An alternative approach for regret minimization in MDPs, to alleviate the burden of optimistic planning, is posterior or Thompson sampling. We now introduce a low-regret posterior sampling RL strategy (Osband et al., 2013) applied to our bilinear exponential family model, where planning is needed only for a single MDP and can be done using standard techniques like model predictive path integral control (Williams et al., 2017).

3.2 Exponential Family PSRL Algorithm

We consider a Bayesian setting in which the unknown parameter θ∗ ∈ Rd of the exponential family MDP (1) is assumed to be distributed according to a (known) prior µ. At the beginning of episode t (i.e., after n = (t − 1)H total steps), we first sample a parameter θn ∼ µn, where µn = P(θ ∼ | Hn) denotes the posterior distribution of θ∗, given the history of transitions Hn = {(s0, a0, s′0), ..., (st−1, at−1, st)}. Then, we execute the optimal policy for the MDP whose transition model is parameterized by θn: πt = argmaxa∈A Vθn,1(s′t | st, at). We call this algorithm Exponential Family Posterior Sampling RL (Exp-PSRL).

Similar to prior work (Osband et al., 2013), we can bound the Bayes regret E[ R(N) ], where the expectation is taken with respect to the randomness in θ∗, in the state transitions and in the algorithm. The complete proof is deferred to Appendix D.

Theorem 3 (Bayes regret for Exp-PSRL). Let θ∗ ∼ µ. Then, the Bayes regret of Exp-PSRL is

\[ \mathbb{E}[R(N)] \leq 2H \left( \frac{\beta}{\alpha} \left( 1 + \frac{\beta B_{A,H}}{\eta} \right) 2B_{N(1/N)}N \gamma N + \frac{2H \beta}{3} \right) \right) \frac{\beta B_{\varphi,A,H}}{\eta} B_{N(1/N)}N \gamma N + 1, \]

where B_{A,\varphi,A}, B_{\varphi,A}, \alpha, \beta, \gamma N and \beta N(\cdot) are as given in Theorem 3.

Note that the regret bound depends on the prior distribution via the norm bound ∥θ∗∥A ≤ B_{A,\varphi,A}, assumed to hold almost surely with respect to the prior. The proof of this result follows the general template of Osband et al. (2013) and works for any prior distribution µ. However, the exponential family structure suggests existence of a natural conjugate prior, described below.

Conjugate Prior for Conditional Exponential Families We consider the prior distribution:

\[ \mu(\theta) \propto \exp \left( \sum_{i=1}^{d} \theta_i \psi(s_0^i) A_i \varphi(s_0, a_0) - n_0 Z_{s_0, a_0}(\theta) \right), \]

where (s0, a0, s′0) ∈ S × A × S and n0 ∈ N+ is a scalar.

We can think of the prior as incorporating "virtual" observations of (s0, a0, s′0). Now, given n samples (st, at, s′t) ≤ n, we obtain the joint likelihood

\[ \mathcal{L}_n(\theta) \propto \exp \left( \sum_{i=1}^{d} \theta_i \psi(s_i^i) A_i \varphi(s, a) - Z_{s, a}(\theta) \right). \]

Then the posterior density takes the form

\[ \mu_n(\theta) \propto \exp \left( \sum_{t=1}^{n+n_0} \sum_{i=1}^{d} \theta_i \psi(s_i^i) A_i \varphi(s, a) - Z_{s, a}(\theta) \right). \]

where we set (st, at, s′t) = (s0, a0, s′0) for all n < t ≤ n + n0. The prior is conjugate since the posterior density takes the same form as the prior.
4 EXAMPLES OF EXPONENTIAL FAMILY TRANSITION MODEL

In this section, we now detail several models of dynamics to show the flexibility of the formulation we consider. Importantly for the practitioner, we recover as special cases the large classes of linear dynamical systems from continuous control literature, as well as of factored and tabular MDPs in the computer science-RL tradition.

4.1 Linearly controlled dynamical systems

Let us consider the classical linear dynamical system evolution [Bertsekas 2001] given by

\[ s' = Gs + Ha + ζ, \]

where \( s, s' \in \mathbb{R}^{m_1} \) are the current and next states, \( a \in \mathbb{R}^{m_2} \) is the current action, \( G \in \mathbb{R}^{m_1 \times m_1}, H \in \mathbb{R}^{m_1 \times m_2} \) are matrices representing the process and \( ζ \sim \mathcal{N}(0, Σ_{s,a}) \) is iid state transition noise at the state-action pair \((s, a)\). We assume the matrices \( G \) and \( H \) to be unknown\(^6\) and we denote \( θ = [G, H] \in \mathbb{R}^{m_1 \times (m_1 + m_2)} \).

Identifying the exponent in the second multiplicant above with \( ψ(s')^T Mθφ(s, a) \) yields the natural parametric form \( \psi(s') = s' \in \mathbb{R}^{m_1} \), \( M_θ = I \otimes \text{vec}(θ) \in \mathbb{R}^{m_1 \times m_1 (m_1 + m_2)} \) and \( φ(s, a) = \text{vec}\left((Σ_{s,a})_i [s^T, a^T]\right) \in \mathbb{R}^{m_2 (m_1 + m_2)} \), where any matrix \( M \), \((M)_i\) denotes its \( i\)-th column. In this case, there are \( d = m_1 (m_1 + m_2) \) unknown parameters and the matrix \( A_i \in \mathbb{R}^{m_1 \times m_1 (m_1 + m_2)} \) has \( (j, (j-1)d+i)\)-th entry equal to \( 1 \) for all \( j \) and all \( m_1 \) and all other entries equal to zero. Therefore, in this case \( A = m_1 I \) and thus, equation 2 to

\[ \sum_{i=1}^{m_1} (s' - Gs_i - Ha_i) \begin{bmatrix} Σ_{s,a}^{-1} \\ Σ_{s,a}^{-1} \end{bmatrix} s_i = ηm_1 G_i \]

for all \( i, j \leq m_1 \) and \( k \leq m_2 \). Further, for each state-action pair \((s, a)\), \( C_{s,a}^θ[ψ(s')] = Σ_{s,a} \), and thus \( α \) and \( β \) from Theorem 2 are \( θ_{\text{min}}(Σ_{s,a}) \) and \( θ_{\text{max}}(Σ_{s,a}) \), respectively. Applying Theorem 2 to this MDP yields

**Corollary 1** (Linearly controlled dynamical systems) Under the linearly controlled dynamics, the cumulative regret of \( \text{Exp-UCRL} \) is \( \mathcal{R}(N) = O(m_1 (m_1 + m_2) \sqrt{H^3 N (\log(1/δ))}) \) with probability at least \( 1 - δ \).

**Remark 3.** Corollary 1 matches (order-wise) the bound given in [Abbasi-Yadkori and Szepesvári 2011] if we restrict their result to the bounded linearly controlled dynamical systems.

4.2 Factored MDP

We now consider the factored MDP model introduced by [Kearns and Koller 1999]. Let \( X = S \times A \) so that each \( x \in X \) is a state-action pair \((s, a)\). Let the state space \( S \) and the joint state-action space \( X \) are factorized as Cartesian product of some finite sets: \( S = S_1 \times \cdots \times S_m \) and \( X = X_1 \times \cdots \times X_n \). For each state coordinate \( i \in [m] \), the parents of \( i, \par s_i \subseteq [n] \) are the subset of state-action coordinates that directly influence \( i \). For a state \( s \in S \), the value of \( s \) on the \( i\)-th coordinate is denoted by \( s(i) \) with \( s(i) \in S_i \). Similarly for each state-action pair \( x \in X \), the value of \( x \) for a subset of coordinates \( \par x \), is denoted by \( x(\par x) \). For ease of representation, we enumerate \( S_i = \{1, \ldots, |S_i|\} \) and \( X(\par x) = \{1, \ldots, |X(\par x)|\} \). In factored MDPs, the transition dynamics \( P \) factorize according to the parent relationships:

\[ P_0(s'(s, a)|s, a) = \prod_{i=1}^{m_1} P_0^i(s'(i)|x(\par x)), \]

where each \( P_0^i \) is a conditional probability table (CPT) with \( |S_i| \) rows and \( |X(\par x)| \) columns. Following classical parameterization of discrete distributions as exponential family [Amar 1997], the next state probabilities are given by

\[ P_0^i(j_i|l_i) = \begin{cases} \exp(θ_i, l_i) / \sum_{j_i \in S_i} \exp(θ_i, j_i) & j_i < |S_i|, l_i \leq |X(\par x)| \\ 1 / \sum_{j_i \in S_i} \exp(θ_i, j_i) & j_i = |S_i|, l_i \leq |X(\par x)| \end{cases} \]

The transition model involves a total of \( d = \sum_{i=1}^m |S_i| |X(\par x)| \) real-valued parameters \( θ_{j_i, l_i} \) with \( \sum_{j_i} |X(\par x)| \) of them equal to zero.

Now for any \( x = (j_1, \ldots, j_m) \) and \( x = (x_1, \ldots, x_n) \) such that \( x(\par x) = l_i, \forall i \in [m] \), we have

\[ P_0(s' = (s, a)|s, a) = \exp \left( \sum_{i=1}^{m_1} θ_{j_i, l_i} - Σ_{s,a}(θ) \right) \]

where \( Σ_{s,a}(θ) = \sum_{i=1}^m \log \left( 1 + \sum_{j_i \in S_i} \exp(θ_{j_i, l_i}) \right) \).

Identifying the first term in the exponent above with \( ψ(s')^T Mθφ(s, a) \) yields the natural parametric form with \( M_θ = \text{diag}(M_1^θ, \ldots, M_n^θ) \) being a block diagonal matrix, where each sub-block \( M_i^θ = [θ_{j_i, l_i}, \ldots, θ_{j_i, l_i}^m] \) is composed of the parameters of \( P_0^i \). The state features are \( ψ(s') = [ψ_1(j_1)^T, \ldots, ψ_1(j_m)^T]^T \), where each \( ψ_i(j_i) = 1_{j_i} \), the indicator vector of length \( |S_i| \). The state-
action features are \( \varphi(s,a) = [\varphi^1(l_1), \ldots, \varphi^m(l_m)]^T \), where each \( \varphi^i(l_i) = 1_{[l_i]} \), the indicator vector of length \( |X(\text{par})| \). We can further express \( M_\theta = \sum_{i=1}^{m} \sum_{j=1}^{\left| X(\text{par}) \right|} \sum_{l=1}^{i} \theta^{i}_{j,l} 1_{j,l}^{i} \), where \( 1_{j,l}^{i} \) is the one-hot indicator block diagonal matrix (of suitable size) whose \((j, l, i)\)-th entry is \( i \)-th sub-block of \( 1 \) and all other entries are \( 0 \). Therefore, in this case \( A = I \) and thus, equation (2) specifies to

\[
\sum_{k_i \in S \setminus j} \exp(\theta_{k_i,j,l}) \sum_{t=1}^{n} I(s_t(i) = j, x_t(\text{par}) = l_i) = \eta \theta^{j}_{i,l},
\]

(with \( \theta^{j}_{i,l} = 0 \)), \( \forall i \leq m, j_i < |S|_l, l_i \leq |X(\text{par})|_l \). Further, for the state-action pair \( x = (x_1, \ldots, x_n) \) such that \( x(\text{par}) = l_i, \forall i \leq m, \) where \( l_i \leq |X(\text{par})|_l \), the covariance matrix \( C^\theta_{x,a} [\psi(s')] \) is block diagonal with all blocks being diagonal. The \( j_i \)-th diagonal entry, \( j_i \leq |S|_l \), of the \( i \)-th sub-block, \( i \leq m \), is equal to \( u^{j}_{i,l} = \exp(\theta^{j}_{i,l}) \sum_{k_i \in S \setminus j} \exp(\theta_{k_i,j,l}) \). Consequently, \( \alpha \) and \( \beta \) from Theorem 2 simply become

\[
\alpha = \min_{j_i \leq |S|_l, l_i} u^{j}_{i,l} \quad \text{and} \quad \beta = \max_{j_i \leq |S|_l, l_i} u^{j}_{i,l} < \frac{1}{2}\text{.}
\]

Now, applying Theorem 2 to this MDP yields

**Corollary 2 (Factored MDP regret).** For factored MDPs, the cumulative regret of Exp-UCRL is \( \mathcal{R}(N) = \tilde{O} \left( \sum_{i=1}^{\Lambda} |S|_l |X(\text{par})|_l |\sqrt{H^3 N \log(1/\delta)} \right) \) with probability at least \( 1 - \delta \).

### 4.3 Tabular MDP

We now consider a discrete transition distribution with \( S = m_1 \) states and \( A = m_2 \) actions: for each state-action pair \( (s_t, a_t) \), \( 1 \leq t \leq m_1 m_2 \), and following classical parametrization of discrete distributions as exponential family, the next state probabilities are given by

\[ P_0(s'_t | s_t, a_t) = \frac{\exp(\theta_{s_t,a_t}^i)}{1 + \sum_{i=1}^{m_1} \exp(\theta_{s_t,a_t}^i)} \text{ for } 1 \leq i \leq m_1 - 1, \]

and

\[ P_0(s'_t | s_t, a_t) = \frac{1}{1 + \sum_{i=1}^{m_1} \exp(\theta_{s_t,a_t}^i)} \text{ for } i = m_1. \]

This model involves \( m_1 m_2 \) real-valued parameters \( \theta_{s_t,a_t}^i \), \( i \in [m_1], l \in [m_1 m_2], m_1 m_2 \) of which being equal to \( 0 \) (we have \( \forall l, \theta_{s_t,a_t}^i = 0 \)), and the rest are unknown. The conditional probability mass function of the next state satisfies

\[
P_0(s'_t | s_t, a_t) = \exp(\theta_{s_t,a_t}^i) \sum_{n=0}^{\infty} \frac{(M_{\theta}^T)^n}{n!} 1_{x_t(i)} 1_{x_t(j)} \text{ where } 1_{x_t(i)} \text{ is the indicator vector of suitable length whose } x_t \text{-th entry is } 1 \text{ and all other entries are } 0. \]

Further, \( M_{\theta} = \sum_{j=1}^{m_1} \sum_{r=1}^{m_2} \theta_{s_t,a_t}^{j,r} 1_{x_t(i)}, \) \( 1_{x_t(i)} \) is the one-hot indicator matrix (of suitable size) whose \((x, y)\)-th entry is \( 1 \) and all other entries are \( 0 \). Therefore, in this case \( A = I \) and thus, equation (2) specifies to

\[
\sum_{k \neq i} \exp(\theta_{k,i}) \sum_{k=1}^{m_1 m_2} \sum_{t=1}^{n} I(s_t = s_i, s_t = s_i, a_t = a) = \eta \theta_{s_t,a_t}^i, \]

(with \( \theta_{s_t,a_t}^i = 0 \)), \( \forall i \in [m_1], l \in [m_1 m_2]. \)

Further, for each state-action pair \( l \in [m_1 m_2], C^\theta_{s_t,a_t} [\psi(s')] \) is a diagonal matrix with entry \((i, i), i \in [m_1]\) equal to \( \exp(\theta_{s_t,a_t}^i) \sum_{k \neq i} \exp(\theta_{k,i}) \). Consequently, \( \alpha \) and \( \beta \) from Theorem 2 simply become

\[
\begin{align*}
\alpha &= \min_{i \in [m_1]} \exp(\theta_{s_t,a_t}^i) \sum_{k \neq i} \exp(\theta_{k,i}) , \\
\beta &= \max_{i \in [m_1]} \exp(\theta_{s_t,a_t}^i) \sum_{k \neq i} \exp(\theta_{k,i}) \leq \frac{1}{2}.
\end{align*}
\]

### 5 PROOF SKETCH: THEOREM 2

We give, in this section, an overview of several of the key ideas behind the main regret bound (Theorem 2). The full proof is deferred to the appendix.

**Step 1: Optimism** Let us consider a fixed episode \( t \), i.e., when \( n = (t-1)H \). Let \( \theta_{s_t,a_t}^i \) denote the most optimistic realization from the confidence ellipsoid \( \Theta_n \), i.e., \( \forall s_t, a_t, V_{\Theta_n}^\pi(s_t, a_t) \geq V_{\bar{\Theta}_n}^\pi(s_t, a_t) \), \( \forall s_t \in \Pi, \forall \theta \in \Theta_n \). Therefore, as long as the true parameter \( \theta \in \Theta_n \) with high probability, \( V_{\Theta_n}^\pi(s_t, a_t) \) gives an optimistic estimate of the value \( V_{\bar{\Theta}_n}^\pi(s_t, a_t) \) of the episode. An application of Theorem 2 implies that with probability at least \( 1 - \delta/2, \forall \theta \in \Theta_n \) across all episodes and thus, in turn, the cumulative regret \( \mathcal{R}(N) \leq \sum_{t=1}^{N} V_{\Theta_n}^\pi(s_t, a_t) - V_{\Theta_n}^\pi(s_t, a_t) \).
Step 2: Bellman Recursion, Transportation and Martingale Control

For any $\theta \in \mathbb{R}^d$, $\pi \in \Pi$ and $V : S \to \mathbb{R}$, we define the Bellman operator $\mathcal{T}_h \leq H$ as

$$\mathcal{T}_{\theta, h}^\pi (V) (s) = R (s, \pi (s, h)) + \mathbb{E}_{\theta, \pi (s, h)} [V] .$$

Now, by the Bellman equation, we have $\forall h \leq H$, $V_{\theta, h}^\pi (s) = \mathcal{T}_{\theta, h}^\pi (V_{\theta, h+1}^\pi) (s)$, (with $V_{\theta, H+1}^\pi (s) := 0$).

Applying the Bellman equation recursively, the cumulative regret can be upper bounded as $R \sum_{t \leq T \leq H} \mathcal{T}_{\theta, h}^\pi (V_{\theta, h}^\pi) (s) + m_h^k$,

where $\{m_h^k\}_{t,h}$ is a martingale difference sequence satisfying $|m_h^k| \leq 2H$. Therefore, by the Azuma-Hoeffding inequality, $\mathbb{E} \left[ \sum_{t \leq T \leq H} \mathcal{T}_{\theta, h}^\pi (V_{\theta, h}^\pi) (s) + m_h^k \right] \leq H \left( \sum_{t \leq T \leq H} \mathbb{E} \left[ \mathcal{T}_{\theta, h}^\pi (V_{\theta, h}^\pi) (s) + m_h^k \right] \right)$.

Step 3: Controlling Sum of KL Divergences

We first approximate the KL divergence $b/w P_0 (| s, a)$ and $P_\theta (| s, a)$ using curvature properties of the log-partition function as $\frac{\sum_{t \leq T \leq H} \mathcal{T}_{\theta, h}^\pi (V_{\theta, h}^\pi) (s) + m_h^k}{\sum_{t \leq T \leq H} \mathcal{T}_{\theta, h}^\pi (V_{\theta, h}^\pi) (s) + m_h^k}$, using curvature properties of the log-partition function together with the definition of $\beta$. Then, for any $\theta \in \Theta, \pi \in \Pi$, we obtain $V (s, a)$, $KL_{\pi, a} (\theta, \pi) \leq (\beta / \alpha) \cdot \beta_n (\delta) \left\| G_{\Theta, a} \right\|$, where $G_{\Theta, a} := G_{\Theta, a}^\pi + \alpha^{-1} \eta \Lambda$ and $G_{\Theta, a} := \sum_{\tau=1}^{H} G_{s, a}^\pi$. Now $\left\| G_{\Theta, a} \right\| \leq \left\| \Lambda^{-1} G_{\Theta, a} \right\| \leq \sum_{\tau=1}^{H} G_{s, a}^\pi$, $\left\| G_{\Theta, a} \right\| \leq \sum_{\tau=1}^{H} G_{s, a}^\pi$.

Next, we deduce that $\left\| \gamma^{\pi} \right\| \leq \left\| \gamma^{\pi} \right\| \left\| \Lambda^{-1} G_{\Theta, a} \right\| \leq \sum_{\tau=1}^{H} G_{s, a}^\pi$. Using spectral properties of the matrices, we can now show that

$$\sum_{t \leq T \leq H} \left\| G_{\Theta, a} \right\| \leq \log \det (I + \alpha^{-1} \Lambda^{-1} G_{N} ) ,$$

which is further upper bounded by $\gamma_N = d \log (1 + \beta \eta^{-1} B_{\phi, a} N)$. Therefore, since $\beta_n$ is monotone increasing in $n$, we have for any $\theta \in \Theta, \pi \in \Pi$, $\sum_{t \leq T \leq H} KL_{s, a} (\theta, \pi) \leq (\beta / \alpha) \cdot (1 + B_{\phi, a} H / \eta) \beta_n (\delta) \gamma_N .$

Final Step: We note that, by design, $\theta_n \in \Theta_n$ and $\Theta_n$ by Theorem 1 $\gamma^* \in \Theta_n$. Further, by Cauchy-Schwartz inequality, $\sum_{t \leq T \leq H} KL_{s, a} (\theta, \pi) \leq \sqrt{N} \sum_{t \leq T \leq H} KL_{s, a} (\theta, \pi)$. The proof now can be put together using (i) the union bound.


