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 – CRIStAL
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 $\tilde{O}(d\sqrt{H^3N})$, 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.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 LINEARITY IN REINFORCEMENT LEARNING

We consider episodic reinforcement learning (RL) in a finite horizon Markov decision process (MDP) (Puterman, 1994; Sutton, 1988), 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 $P_{\theta^\star} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ for some underlying parameter $\theta^\star \in \mathbb{R}^d$ 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 a 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 (Seber and Lee, 2012), the target mean function $f : \mathcal{X} \to \mathbb{R}$ is assumed to be of the form $f(\cdot) = \theta^\top \varphi(\cdot)$ 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 φ 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 (RKHSs) (Paulsen and Raghupathi, 2016).

Linearity in Bandits For stateless MDPs or multi-armed bandits, linear models have been widely studied in a number of works, see Abbasi-Yadkori et al. (2011); Rusmevichientong and Tsitsiklis (2010); Lattimore and Szepesvari (2016), as well as Durand et al. (2017); Filippi et al. (2010) to cite a few, exploiting the connection with linear regression. In particular, the construction of finite-time confidence ellipsoids for the unknown vector parameter θ 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 Srinivas 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, when assuming some smoothness or regularity on the rewards and dynamics such as in Ortner and Ryabko (2012), and more recently Domingues et al. (2020). Building on similar tools for regret minimization in MDPs, and combining them with a linearity assumption instead, in Yang and Wang (2019) and Jin et al. (2019), the authors suggest to exploit linearity in the context of MDPs. For the rewards that are real-valued, one can directly use standard linear regression model using features $\varphi^R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. However an MDP also involves the transitions, which do not take their values in $\mathbb{R}$ but in the probability distributions over $\mathcal{S}$. The authors suggest to introduce a bi-linear model of the transition. Namely, they consider a model of the form $P(\cdot|\cdot) = \psi(\cdot)^\top M \varphi(\cdot)$, where $\psi : \mathcal{S} \to \mathbb{R}^p$ and $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^q$ are known feature functions, and M is a $p \times q$ matrix of unknown parameters. This makes it convenient since each $P(\cdot|s,a)$ can be seen as a linear model with feature function ψ and vector parameter $M\varphi(s,a) \in \mathbb{R}^p$, while each $P(s'|\cdot)$ can be seen as a linear model with feature function φ and vector parameter $\psi(s')^\top M \in \mathbb{R}^q$. Hence, popular tools from linear regression can be considered.

Bilinear Exponential Families Unfortunately, such a direct bilinear assumption of the transition probabilities falls short of capturing many rich classical models such as the linear quadratic regulator (LQR) and the factored MDPs. On the other hand, statistics 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[1]. More precisely, we assume the following bilinear exponential family model[2]:

$$P_\theta(s'|s,a) = h(s',s,a)\exp\big(\psi(s')^\top M_\theta \varphi(s,a) - Z_{s,a}(\theta)\big)\,,$$
$$Z_{s,a}(\theta) = \log \int_{\mathcal{S}} \exp\big(\psi(s')^\top M_\theta \varphi(s,a)\big) h(s',s,a) ds'\,, \quad (1)$$

at every $s, s' \in \mathcal{S}, a \in \mathcal{A}$. Here h, φ and ψ are known feature functions, and M_θ is a $p \times q$ matrix of unknown parameters. Note that h, ψ and φ cannot depend on θ. Further, since considering all entries of the matrix M_θ as unrelated parameters may prevent one from encoding stronger structure, we consider that it is of the form $M_\theta = \sum_{i=1}^d \theta_i A_i$, where $\theta = (\theta_i)_{i \le d} \in \mathbb{R}^d$ 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_θ by considering $d = pq$ and the $(A_i)_{i \le d}$ 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 $s' \mapsto (\psi(s')^\top A_i \varphi(s,a))_{i \le d}$ and unknown parameter $\theta \in \mathbb{R}^d$, while for each s', we get a linear model with feature function $(s,a) \mapsto (\psi(s')^\top A_i \varphi(s,a))_{i \le d}$ and the same unknown parameter. We detail popular classes of MDP dynamics that fit this formulation in Section 4.

Learning Model and Regret The learning agent interacts with the MDP in episodes and, at each episode t, a trajectory $(s_1^t, a_1^t, r_1^t, \ldots, s_H^t, a_H^t, r_H^t, s_{H+1}^t)$ is generated. Here a_h^t denotes the action taken at state s_h^t, $r_h^t = R(s_h^t, a_h^t)$ denotes the immediate reward, and $s_{h+1}^t \sim P_{\theta^\star}(\cdot|s_h^t, a_h^t)$ denotes the random next state. The initial state s_1^t is assumed to be fixed and history independent. The actions are chosen following some policy $\pi = (\pi_1, \ldots, \pi_H)$, where each π_h is a mapping from the state space $\mathcal{S}$ into the action space $\mathcal{A}$. The agent would like to find a policy π that maximizes the long term expected reward starting from every state $s \in \mathcal{S}$ and every step $h \in [H]$, defined as

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

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

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

[1] This is also analogous to the logistic model where we linearize not raw probabilities but their log-odds (Hosmer Jr et al., 2013).

[2] We write probability measures assuming that they have a density mainly for convenience; the development can easily be extended to general probability transition measures.

of the agent by the cumulative (pseudo) regret accumulated over T episodes, defined as

$$\mathcal{R}(N) = \sum_{t=1}^{T} \left[V_{\theta^\star,1}^{\pi^\star}(s_1^t) - V_{\theta^\star,1}^{\pi^t}(s_1^t) \right] ,$$

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 π^t in episode t starting from some fixed initial state s_1^t. 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

$$\nabla_i Z_{s,a}(\theta) = \mathbb{E}_{s,a}^{\theta}[\psi(s')]^\top A_i \varphi(s,a) ,$$
$$\nabla_{i,j}^2 Z_{s,a}(\theta) = \varphi(s,a)^\top A_i^\top \mathbb{C}_{s,a}^{\theta}[\psi(s')] \, A_j \varphi(s,a) ,$$
$$\mathrm{KL}_{s,a}(\theta,\theta') = Z_{s,a}(\theta') - Z_{s,a}(\theta) - (\theta'-\theta)^\top \nabla Z_{s,a}(\theta) ,$$

where $\mathbb{E}_{s,a}^{\theta}, \mathbb{C}_{s,a}^{\theta}$ denote the expectation and covariance operator for the probability distribution $P_\theta(\cdot|s,a)$, and $\mathrm{KL}_{s,a}(\theta,\theta')$ 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 $\{(s_t,a_t,s_t')\}_{t\le n}$, where for each t, $s_t' \sim P_{\theta^\star}(\cdot|s_t,a_t)$, and any differentiable penalty function $\mathrm{pen}(\cdot)$, a solution to the penalized maximum-likelihood problem with regularization parameter $\eta \in \mathbb{R}^+$ must satisfy

$$\theta_n \in \arg\min_{\theta\in\mathbb{R}^d} \sum_{t=1}^{n} -\log P_\theta(s_t'|s_t,a_t) + \eta \, \mathrm{pen}(\theta) \Longrightarrow \forall i \le d,$$

$$\sum_{t=1}^{n} \left(\psi(s_t') - \mathbb{E}_{s_t,a_t}^{\theta_n}[\psi(s')]\right)^\top A_i \varphi(s_t,a_t) = \eta \nabla_i \mathrm{pen}(\theta_n). \quad (2)$$

In this paper, we choose a trace-norm penalty $\mathrm{pen}(\theta) = \frac{1}{2}\|\theta\|_{\mathbb{A}}^2$, where $\mathbb{A}$ denotes the matrix with entries $\mathbb{A}_{i,j} = \mathrm{tr}(A_i A_j^\top)$, $i,j \le d$. We assume that $\mathbb{A}$ 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 $\{\mathcal{F}_t\}_{t=0}^{\infty}$ is a filtration such that for each t, (i) s_t' is $\mathcal{F}_t$-measurable, (ii) (s_t,a_t) is $\mathcal{F}_{t-1}$-measurable, and (iii) given (s_t,a_t), $s_t' \sim P_{\theta^*}(\cdot|s_t,a_t)$ according to the exponential family defined by (1). Let θ_n be the penalized MLE defined by (2), and let $Z_{s,a}(\theta)$ be strictly convex in θ for all (s,a).*[3] *Then, for any $\delta \in (0,1]$, with probability at least $1-\delta$, the following holds uniformly over all $n \in \mathbb{N}$:*

$$\sum_{t=1}^{n} \mathrm{KL}_{s_t,a_t}(\theta_n,\theta^\star) + \frac{\eta}{2}\|\theta^\star - \theta_n\|_{\mathbb{A}}^2 - \frac{\eta}{2}\|\theta^\star\|_{\mathbb{A}}^2 \le \log\left(\frac{C_{\mathbb{A},n}}{\delta}\right),$$

where $C_{\mathbb{A},n} = \dfrac{\int_{\mathbb{R}^d} \exp\left(-\frac{\eta}{2}\|\theta'\|_{\mathbb{A}}^2\right) d\theta'}{\int_{\mathbb{R}^d} \exp\left(-\sum_{t=1}^{n} \mathrm{KL}_{s_t,a_t}(\theta_n,\theta') - \frac{\eta}{2}\|\theta'-\theta_n\|_{\mathbb{A}}^2\right) d\theta'}$.

Furthermore, introducing the matrix $(G_{s,a})_{i,j} = \varphi(s,a)^\top A_i^\top A_j \varphi(s,a)$, $\forall i,j \le d$, we have

$$C_{\mathbb{A},n} \le \det\left(I + \beta\eta^{-1}\mathbb{A}^{-1}\sum_{t=1}^{n} G_{s_t,a_t}\right) ,$$

where $\beta = \sup_{\theta,s,a} \lambda_{\max}\left(\mathbb{C}_{s,a}^{\theta}[\psi(s')]\right)$.

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

[3] Strict convexity essentially amounts to a minimal representation of the exponential family (Amari, 1997). We assume it for brevity; the result holds even if $Z_{s,a}$ 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 $\inf_{\theta,s,a} \lambda_{\min}\left(\mathbb{C}^{\theta}_{s,a}[\psi(s')]\right)$.

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

$$M_n^\lambda = \exp\Big(\lambda^\top S_n - \sum\nolimits_{t=1}^{n} B_{Z_{s_t,a_t},\theta^\star}(\lambda)\Big) \ ,$$

where $(S_n)_i = \sum_{t=1}^n \big(\psi(s'_t) - \mathbb{E}^{\theta^\star}_{s_t,a_t}\left[\psi(s')\right]\big)^\top A_i \varphi(s_t,a_t)$, $\forall i \le d$. Since, by construction, $\log \mathbb{E}\left[\exp(\lambda^\top S_n)|\mathcal{F}_{n-1}\right] = \lambda^\top S_{n-1} + B_{Z_{s_n,a_n},\theta^\star}(\lambda)$, M_n^λ is a non-negative martingale such that $\mathbb{E}[M_n^\lambda]=1$. Further, we show that for any random stopping time N, $\mathbb{E}[M_N^\lambda] \le 1$.

We now apply the method of mixtures technique (Peña et al., 2008) by integrating over λ. To this end, for any prior density $q(\theta)$ for θ, we define a mixture of martingales $M_n = \int_{\mathbb{R}^d} M_n^\lambda q(\theta^\star+\lambda) d\lambda$ so that $\mathbb{E}\left[M_n\right]=1$. Considering the prior density $\mathcal{N}\left(0,(\eta\mathbb{A})^{-1}\right)$, we then show that

$$M_n = \exp\Big(\sum_{t=1}^{n} B_{Z_{s_t,a_t}}(\theta^\star,\theta_n) + \frac{\eta}{2}\|\theta^\star-\theta_n\|^2_{\mathbb{A}} - \frac{\eta}{2}\|\theta^\star\|^2_{\mathbb{A}}\Big)/C_{\mathbb{A},n}.$$

We then deduce from a simple Markov inequality that

$$\mathbb{P}\left[\sum_{t=1}^{n} B_{Z_{s_t,a_t}}(\theta^\star,\theta_n) + \frac{\eta}{2}\|\theta^\star-\theta_n\|^2_{\mathbb{A}} - \frac{\eta}{2}\|\theta^\star\|^2_{\mathbb{A}} \ge \log\left(\frac{C_{\mathbb{A},n}}{\delta}\right)\right]$$
$$= \mathbb{P}\left[M_n \ge 1/\delta\right] \le \delta \cdot \mathbb{E}\left[M_n\right] \le \delta \ , \quad \forall \delta \in (0,1) \ .$$

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 REGRET MINIMIZATION IN BILINEAR EXPONENTIAL MDPS

We now introduce a low-regret algorithm inspired by the popular upper confidence RL (UCRL) strategy[4] 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 tradeoff, the proposed algorithm maintains both an empirical estimate of $\theta^\star$ as well as a high-probability confidence set. Specifically, at the start of episode t, we let $n=(t-1)H$ denote the total number of steps completed so far. The penalized maximum likelihood (PML) estimate $\theta_n \equiv \theta_{(t-1)H}$ is computed to solve the following equation

$$\sum_{\tau=1}^{t-1}\sum_{h=1}^{H}\left(\psi(s^\tau_{h+1}) - \mathbb{E}^{\theta}_{s^\tau_h,a^\tau_h}\Big[\psi(s')\Big]\right)^\top A_i \varphi(s^\tau_h,a^\tau_h) = \eta(\mathbb{A}\theta)_i \ .$$

where $\forall i \le d$, $(\mathbb{A}\theta)_i = \sum_{j=1}^d \theta_j \,\mathrm{tr}(A_i A_j^\top)$. Penalization acts as a regularizer and avoids the need for any specific initialization scheme to make the MLE well-defined. Now, for any $\delta \in (0,1]$ and $B_{\mathbb{A}} \ge \|\theta^\star\|_{\mathbb{A}}$, Theorem 1 suggests the following high probability confidence set around the PML estimate:

$$\Theta_n = \left\{\theta \in \mathbb{R}^d \Big| \sum_{\tau=1}^{t-1}\sum_{h=1}^{H} \mathrm{KL}_{s^\tau_h,a^\tau_h}(\theta_n,\theta) + \frac{\eta}{2}\|\theta-\theta_n\|^2_{\mathbb{A}} \le \beta_n(\delta)\right\}$$

where the confidence width is given by $\beta_n(\delta) \equiv \beta_{(t-1)H}(\delta) = \frac{\eta}{2}B^2_{\mathbb{A}} + \log(2C_{\mathbb{A},(t-1)H}/\delta)$. This high-probability confidence set is then used to carry out an optimistic planning step:

$$\pi^t = \mathrm{argmax}_{\pi\in\Pi} \max_{\theta\in\Theta_n} V^\pi_{\theta,1}(s^t_1) \ , \qquad (3)$$

where s^t_1 is the initial state, and $V^\pi_{\theta,1}$ is the value function under transition model is P_θ. Using the transition model (1) that we introduced, the algorithm crucially adapts the confidence sets of UCRL2 (Jaksch et al., 2010) to exploit the linear structure. We call this algorithm Exponential Family Upper Confidence RL (Exp-UCRL); its pseudocode appears in Algorithm 1.

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

Input: Matrices $A_1,\dots,A_d$, constant $B_{\mathbb{A}}$, parameters $\delta \in (0,1]$, $\eta > 0$.
for episode $t = 1,2,3,\dots$ **do**
  Set $n=(t-1)H$.
  Compute the penalized ML estimate θ_n and the confidence set Θ_n.
  Observe initial state s^t_1.
  Choose policy $\pi^t = \mathrm{argmax}_{\pi\in\Pi}\max_{\theta\in\Theta_n} V^\pi_{\theta,1}(s^t_1)$.
  for period $h = 1,2,3,\dots,H$ **do**
    Choose action $a^t_h = \pi^t_h(s^t_h)$.
    Observe reward r^t_h and next state s^t_{h+1}.
  end for
end for

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* $\mathbb{A}_{i,j} = \mathrm{tr}(A_i A_j^\top)$ *and* $(G_{s,a})_{i,j} = \varphi(s,a)^\top A_i^\top A_j \varphi(s,a)$, $\forall i,j \le d$. *Assume that* $\|\theta^\star\|_{\mathbb{A}} \le B_{\mathbb{A}}$ *and* $\left\|\mathbb{A}^{-1}G_{s,a}\right\| \le B_{\varphi,\mathbb{A}}$ *for all* (s,a). *Then, for any* $\eta > 0$ *and* $\delta \in (0,1]$, Exp-UCRL *enjoys, with probability at least* $1-\delta$, *the*

[4] or 'optimism in the face of uncertainty'

cumulative regret

$$\mathcal{R}(N) \leq 2H\sqrt{\frac{\beta}{\alpha}\left(1+\frac{\beta B_{\varphi,\mathbb{A}}H}{\eta}\right)2\beta_N(\delta)N\gamma_N}$$
$$+2H\sqrt{2N\ln(2/\delta)}+\frac{2H}{3}\frac{\beta}{\alpha}\left(1+\frac{\beta B_{\varphi,\mathbb{A}}H}{\eta}\right)\beta_N(\delta)\gamma_N\ ,$$

where $\gamma_N \equiv \gamma_N(\beta,\varphi,\mathbb{A}) = d\log\big(1+\beta\eta^{-1}B_{\varphi,\mathbb{A}}N\big)$, $\alpha = \inf_{\theta,s,a}\lambda_{\min}\big(\mathbb{C}^{\theta}_{s,a}[\psi(s')]\big)$, $\beta = \sup_{\theta,s,a}\lambda_{\max}\big(\mathbb{C}^{\theta}_{s,a}[\psi(s')]\big)$, $\beta_N(\delta) = \frac{\eta}{2}B^2_{\mathbb{A}}+\gamma_N+\log(2/\delta)$ *and* $N=TH$.

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 α scales only as $1/\sqrt{\alpha}$ 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 1) 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 $\tilde{O}\big(d\sqrt{H^3N}\big)$ 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 $\mathcal{S}$ and $\mathcal{A}$, 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 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 $O\big(d\sqrt{N}\big)$ 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 $P(s'|s,a) = \psi(s')^{\top}M\varphi(s,a)$, and propose a model-based algorithm with regret $\tilde{O}\big(\sqrt{d^3H^4N}\big)$ 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 $\tilde{O}\big(\sqrt{d^3H^3N}\big)$. 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 factored MDPs.[5] We consider a more natural and flexible transition model, involving exponential families, than these prior works. Though it is incomparable 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 $\tilde{O}\big(\sqrt{HSAN}\big)$ (Gheshlaghi Azar et al., 2017; Zanette and Brunskill, 2019; Efroni et al., 2019) whereas the best known regret for model-free learning is $\tilde{O}\big(\sqrt{H^3SAN}\big)$ (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 $O\big(d^{3/2}\big)$; however, the ap-

[5] 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).

parent $\sqrt{d}$ factor improvement in our result is just a consequence of what bound we assume on the scale of the features. Under a similar assumption, the prior bounds can also be shown to be linear in the number of parameters d like ours. The more important question is whether this linear dependency is optimal or not. A linear scaling of worst-case regret is well-known for linear stochastic bandits (Abbasi-Yadkori et al., 2011), which are a special case of the MDPs studied in this work with an additional linear reward structure and with the episode length set equal to one. We note, however, that while an MDP has state transitions, the bandits do not, and a naive adaptation of existing linear bandit algorithms to the linear MDP setting would give a regret exponential in episode length H. As for model-based RL, Osband and Van Roy (2014a) analyze the regret guarantee for any given class of transition functions. Chowdhury and Gopalan (2019) make a smoothness assumption compatible with a reproducing kernel Hilbert space and prove regret bound for this transition model. In the case of linearly parametrized (with d unknown parameters) transition models, both these bound reduce to $\tilde{O}\big(d\sqrt{H^2N}\big)$. However, the above works make a restrictive assumption of the transitions being deterministic with a controllable amount of noise. We make 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 $\varepsilon = \sqrt{H/t}$ at episode t, we can ensure that this adds only an additional $O(\sqrt{N})$ 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 $\theta^\star \in \mathbb{R}^d$ 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 $\tilde{\theta}_n \sim \mu_n$, where $\mu_n = \mathbb{P}(\theta^\star \in \cdot | \mathcal{H}_n)$ denotes the posterior distribution of $\theta^\star$, given the history of transitions $\mathcal{H}_n = \{(s_h^\tau, a_h^\tau, s_{h+1}^\tau)_{\tau<t, h\le H}\}$. Then, we execute the optimal policy for the MDP whose transition model is parameterized by $\tilde{\theta}_n$: $\pi^t = \operatorname{argmax}_{\pi\in\Pi} V^\pi_{\tilde{\theta}_n,1}(s_1^t)$, where s_1^t is the initial state and and $V^\pi_{\tilde{\theta}_n,1}$ is the value function when the transition model is $P_{\tilde{\theta}_n}$. 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 $\mathbb{E}[\mathcal{R}(N)]$, where the expectation is taken with respect to the randomness in $\theta^\star$, in the state transitions and in the algorithm. The complete proof is deferred to Appendix D.

Theorem 3 (Bayes regret for Exp-PSRL)**.** *Let $\theta^\star \sim \mu$. Then, the Bayes regret of* Exp-PSRL *is*

$$\mathbb{E}\left[\mathcal{R}(N)\right] \le 2H\sqrt{\frac{\beta}{\alpha}\left(1+\frac{\beta B_{\varphi,\mathbb{A}} H}{\eta}\right) 2\beta_N(1/N) N\gamma_N} + \frac{2H}{3}\frac{\beta}{\alpha}\left(1+\frac{\beta B_{\varphi,\mathbb{A}} H}{\eta}\right)\beta_N(1/N)\gamma_N + 1 \ ,$$

where $B_{\mathbb{A}}$, $B_{\varphi,\mathbb{A}}$, α, β, γ_N and $\beta_N(\cdot)$ are as given in Theorem 2.

Note that the regret bound depends on the prior distribution via the norm bound $\|\theta^\star\|_{\mathbb{A}} \le B_{\mathbb{A}}$, 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:

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

where $(s_0, a_0, s_0') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and $n_0 \in \mathbb{N}^+$ is a scalar. We can think of the prior as incorporating n_0 "virtual" observations of (s_0, a_0, s_0'). Now, given n samples $(s_t, a_t, s_t')_{t\le n}$, we obtain the joint likelihood

$$\mathcal{L}_n(\theta) \propto \exp\left(\sum_{t=1}^{n}\left(\sum_{i=1}^{d}\theta_i \psi(s_t')^\top A_i \varphi(s_t,a_t) - Z_{s_t,a_t}(\theta)\right)\right).$$

Then the posterior density takes the form

$$\mu_n(\theta) \propto \exp\left(\sum_{t=1}^{n+n_0}\left(\sum_{i=1}^{d}\theta_i \psi(s_t')^\top A_i \varphi(s_t,a_t) - Z_{s_t,a_t}(\theta)\right)\right).$$

where we set $(s_t, a_t, s_t') = (s_0, a_0, s_0')$ for all $n < t \le n + n_0$. 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

$$s' = Gs + Ha + \zeta,$$

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 $\zeta \sim \mathcal{N}(0, \Sigma_{s,a})$ 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 $\theta = [G, H] \in \mathbb{R}^{m_1 \times (m_1+m_2)}$. For this model the conditional density of the next state s' given s, a is the multivariate normal density

$$P_\theta(s'|s,a) = \frac{\exp\left(-\frac{1}{2} s'^\top \Sigma_{s,a}^{-1} s'\right)}{(2\pi)^{\frac{m_1}{2}} |\Sigma_{s,a}|^{\frac{1}{2}}} \exp\left(s'^\top \Sigma_{s,a}^{-1} \theta \begin{bmatrix} s \\ a \end{bmatrix} \right) \times \exp\left(- \begin{bmatrix} s \\ a \end{bmatrix}^\top \frac{\theta^\top \Sigma_{s,a}^{-1} \theta}{2} \begin{bmatrix} s \\ a \end{bmatrix} \right).$$

Identifying the exponent in the second multiplicand above with $\psi(s')^\top M_\theta \varphi(s,a)$ yields the natural parametric form $\psi(s') = s' \in \mathbb{R}^{m_1}$, $M_\theta = I \otimes \mathrm{vec}(\theta)^\top \in \mathbb{R}^{m_1 \times m_1^2 (m_1+m_2)}$ and $\varphi(s,a) = \left[\mathrm{vec}\left(\left(\Sigma_{s,a}^{-1}\right)_1 [s^\top, a^\top] \right)^\top, \ldots, \mathrm{vec}\left(\left(\Sigma_{s,a}^{-1}\right)_{m_1} [s^\top, a^\top] \right)^\top \right]^\top \in \mathbb{R}^{m_1^2 (m_1+m_2)}$, where for 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 $A_i \in \mathbb{R}^{m_1 \times m_1^2 (m_1+m_2)}$ has $(j, (j-1)d+i)$-th entry equal to 1 for all $j \le m_1$ and all other entries equal to zero. Therefore, in this case $\mathbb{A} = m_1 I$ and thus, equation (2) specifies to

$$\sum_{t=1}^n (s'_t - Gs_t - Ha_t)^\top \left(\Sigma_{s_t,a_t}^{-1}\right)_i (s_t)_j = \eta m_1 G_{i,j} \text{ and}$$
$$\sum_{t=1}^n (s'_t - Gs_t - Ha_t)^\top \left(\Sigma_{s_t,a_t}^{-1}\right)_i (a_t)_k = \eta m_1 H_{i,k} ,$$

for all $i, j \le m_1$ and $k \le m_2$. Further, for each state-action pair (s,a), $\mathbb{C}_{s,a}^\theta [\psi(s')] = \Sigma_{s,a}$, and thus α and β from Theorem 2 are $\lambda_{\min}(\Sigma_{s,a})$ and $\lambda_{\max}(\Sigma_{s,a})$, respectively. Applying Theorem 2 to this MDP yields

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

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 $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ so that each $x \in \mathcal{X}$ is a state-action pair (s,a). Let the state space $\mathcal{S}$ and the joint state-action space $\mathcal{X}$ are factorized as Cartesian product of some finite sets: $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_m$ and $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. For each state coordinate $i \in [m]$, the parents of i, $\mathrm{par}_i \subseteq [n]$ are the subset of state-action coordinates that directly influence i. For a state $s \in \mathcal{S}$, the value of s on the i-th coordinate is denoted by $s(i)$ with $s(i) \in \mathcal{S}_i$. Similarly for each state-action pair $x \in \mathcal{X}$, the value of x for a subset of coordinates par_i is denoted by $x(\mathrm{par}_i)$ with $x(\mathrm{par}_i) \in \mathcal{X}(\mathrm{par}_i)$, where $\mathcal{X}(\mathrm{par}_i) = \bigotimes_{j \in \mathrm{par}_i} \mathcal{X}_j$. For ease of representation, we enumerate $\mathcal{S}_i = \{1, \ldots, |\mathcal{S}_i|\}$ and $\mathcal{X}(\mathrm{par}_i) = \{1, \ldots, |\mathcal{X}(\mathrm{par}_i)|\}$. In factored MDPs, the transition dynamics P factorize according to the parent relationships:

$$P_\theta(s'|s,a) \equiv P_\theta(s'|x) = \prod_{i=1}^m P_\theta^i (s'(i)|x(\mathrm{par}_i)) ,$$

where each P_θ^i is a conditional probability table (CPT) with $|\mathcal{S}_i|$ rows and $|\mathcal{X}(\mathrm{par}_i)|$ columns. Following classical parametrization of discrete distributions as exponential family (Amari, 1997), the next state probabilities are given by

$$P_\theta^i (j_i|l_i) = \begin{cases} \frac{\exp(\theta^i_{j_i,l_i})}{1+\sum_{j_i<|\mathcal{S}_i|} \exp(\theta^i_{j_i,l_i})}, & j_i < |\mathcal{S}_i|, l_i \le |\mathcal{X}(\mathrm{par}_i)| \\ \frac{1}{1+\sum_{j_i<|\mathcal{S}_i|} \exp(\theta^i_{j_i,l_i})}, & j_i = |\mathcal{S}_i|, l_i \le |\mathcal{X}(\mathrm{par}_i)| \end{cases}.$$

The transition model involves a total of $d = \sum_{i=1}^m |\mathcal{S}_i| \cdot |\mathcal{X}(\mathrm{par}_i)|$ real-valued parameters $\theta^i_{j_i,l_i}$ with $\sum_{i=1}^m |\mathcal{X}(\mathrm{par}_i)|$ of them equal to zero.

Now for any $s' = (j_1, \ldots, j_m)$ and $x = (x_1, \ldots, x_n)$ such that $x(\mathrm{par}_i) = l_i, \forall i \in [m]$, we have

$$P_\theta(s'|s,a) = \exp\left(\sum_{i=1}^m \theta^i_{j_i,l_i} - Z_{s,a}(\theta) \right) ,$$

where $Z_{s,a}(\theta) = \sum_{i=1}^m \log\left(1 + \sum_{j_i<|\mathcal{S}_i|} \exp(\theta^i_{j_i,l_i})\right)$. Identifying the first term in the exponent above with $\psi(s')^\top M_\theta \varphi(s,a)$ yields the natural parametric form with $M_\theta = \mathrm{diag}(M_\theta^1, \ldots, M_\theta^m)$ being a block diagonal matrix, where each sub-block $M_\theta^i = [\theta^i_{j_i,l_i}]_{j_i,l_i} \in \mathbb{R}^{|\mathcal{S}_i| \times |\mathcal{X}(\mathrm{par}_i)|}$ is composed of the parameters of P_θ^i. The state features are $\psi(s') = [\psi^1(j_1)^\top, \ldots, \psi^m(j_m)^\top]^\top$, where each $\psi^i(j_i) = \mathbb{1}_{j_i}$, the indicator vector of length $|\mathcal{S}_i|$. The state-

[6] We assume that the process noise covariance $\Sigma_{s,a}$ depends on s, a but is known.

action features are $\varphi(s,a)=[\varphi^1(l_1)^\top,\dots,\varphi^m(l_m)^\top]^\top$, where each $\varphi^i(l_i)=\mathbb{1}_{l_i}$, the indicator vector of length $|\mathcal{X}(\mathrm{par}_i)|$. We can further express $M_\theta=\sum_{i=1}^m\sum_{j_i=1}^{|\mathcal{S}_i|}\sum_{l_i=1}^{|\mathcal{X}(\mathrm{par}_i)|}\theta^i_{j_i,l_i}\mathbb{1}^i_{j_i,l_i}$, where $\mathbb{1}^i_{j_i,l_i}$ is the one-hot indicator block diagonal matrix (of suitable size) whose (j_i,l_i)-th entry of the i-th sub-block is 1 and all other entries are 0. Therefore, in this case $\mathbb{A}=I$ and thus, equation (2) specifies to

$$\frac{\sum_{k_i\in\mathcal{S}_i\setminus j_i}\exp(\theta^i_{k_i,l_i})}{\sum_{k_i\in\mathcal{S}_i}\exp(\theta^i_{k_i,l_i})}\sum_{t=1}^n\mathbb{I}\{s'_t(i)=j_i,x_t(\mathrm{par}_i)=l_i\}=\eta\theta^i_{j_i,l_i}$$

(with $\theta^i_{|\mathcal{S}_i|,l_i}=0$), $\forall i\le m$, $j_i<|\mathcal{S}_i|$ and $l_i\le|\mathcal{X}(\mathrm{par}_i)|$.

Further, for the state-action pair $x=(x_1,\dots,x_n)$ such that $x(\mathrm{par}_i)=l_i,\forall i\le m$, where $l_i\le|\mathcal{X}(\mathrm{par}_i)|$, the covariance matrix $\mathbb{C}^\theta_{s,a}[\psi(s')]$ is block diagonal with each sub-block also being diagonal. The j_i-th diagonal entry, $j_i\le|\mathcal{S}_i|$, of the i-th sub-block, $i\le m$, is equal to $u^i_{j_i,l_i}=\frac{\exp(\theta^i_{j_i,l_i})\sum_{k_i\in\mathcal{S}_i\setminus j_i}\exp(\theta^i_{k_i,l_i})}{(\sum_{k_i\in\mathcal{S}_i}\exp(\theta^i_{k_i,l_i}))^2}$. Consequently, α and β from Theorem 2 simply become $\alpha=\min_{i,j_i,l_i}u^i_{j_i,l_i}$ and $\beta=\max_{i,j_i,l_i}u^i_{j_i,l_i}\le\frac14$. Now, applying Theorem 2 to this MDP yields

Corollary 2 (Factored MDP regret)**.** *For factored MDPs, the cumulative regret of* Exp-UCRL *is* $\mathcal{R}(N)=\tilde{O}\big(\sum_{i=1}^m|\mathcal{S}_i||\mathcal{X}(\mathrm{par}_i)|\sqrt{H^3N\log(1/\delta)}\big)$ *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_l,a_l), $1\le l\le m_1m_2$, and following classical parametrization of discrete distributions as exponential family, the next state probabilities are given by $P_\theta(s'_i|s_l,a_l)=\frac{\exp(\theta_{i,l})}{1+\sum_{k=1}^{m_1-1}\exp(\theta_{k,l})}$ for $1\le i\le m_1-1$, and $P_\theta(s'_i|s_l,a_l)=\frac{1}{1+\sum_{k=1}^{m_1-1}\exp(\theta_{k,l})}$ for $i=m_1$. This model involves $m_1^2m_2$ real-valued parameters $\theta_{i,l}$, $i\in[m_1],l\in[m_1m_2]$, m_1m_2 of which being equal to 0 (we have $\forall l,\theta_{m_1,l}=0$), and the rest are unknown. The conditional probability mass function of the next state satisfies $P_\theta(s'_i|s_l,a_l)=\exp(\theta_{i,l}-\log Z_{s_l,a_l}(\theta))=\frac{1}{Z_{s_l,a_l}(\theta)}\exp\left(\mathbb{1}_i^\top M_\theta\mathbb{1}_l\right)$, where $M_\theta=[\theta_{i,l}]_{i,l}$, and $\mathbb{1}_x$ is the indicator vector (of suitable length) whose x-th entry is 1 and all other entries are 0. Further, $M_\theta=\sum_{j=1}^{m_1}\sum_{r=1}^{m_1m_2}\theta_{j,r}\mathbb{1}_{j,r}$, where $\mathbb{1}_{x,y}$ 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 $\mathbb{A}=I$ and thus, equation (2) specifies to

$$\frac{\sum_{k\ne i}\exp(\theta_{k,l})}{\sum_{k=1}^{m_1}\exp(\theta_{k,l})}\sum_{t=1}^n\mathbb{I}\{s'_t=s_i,s_t=s_l,a_t=a_l\}=\eta\theta_{i,l},$$

(with $\theta_{m_1,l}=0$), for all $i\in[m_1],l\in[m_1m_2]$.

Further, for each state-action pair $l\in[m_1m_2]$, $\mathbb{C}^\theta_{s_l,a_l}[\psi(s')]$ is a diagonal matrix with entry (i,i), $i\in[m_1]$ equal to $\frac{\exp(\theta_{i,l})\sum_{k\ne i}\exp(\theta_{k,l})}{\left(\sum_k\exp(\theta_{k,l})\right)^2}$. Consequently, α and β from Theorem 2 simply become

$$\begin{cases}\alpha &= \min_{l,i}\frac{\exp(\theta_{i,l})\sum_{k\ne i}\exp(\theta_{k,l})}{\left(\sum_k\exp(\theta_{k,l})\right)^2}, \quad\text{and}\\ \beta &= \max_{l,i}\frac{\exp(\theta_{i,l})\sum_{k\ne i}\exp(\theta_{k,l})}{\left(\sum_k\exp(\theta_{k,l})\right)^2}\le\frac14.\end{cases}$$

Now, applying Theorem 2 to this MDP yields

Corollary 3 (Tabular regret)**.** *For tabular Markov decision processes, the cumulative regret of* Exp-UCRL *is* $\mathcal{R}(N)=\tilde{O}\big(S^2A\sqrt{H^3N\log(1/\delta)}\big)$ *with probability at least* $1-\delta$*.*

Remark 4. *We remark that our regret bounds for tabular and factored MDPs are worse than the respective best known bounds (Gheshlaghi Azar et al. (2017) for the tabular model and Osband and Van Roy (2014b) for the factored model); by more refined analyses specialized to these models, the regret bound of our algorithm can be improved using techniques similar to the mentioned works. However, we emphasize that our algorithm and analysis tackle a more general setting; thus, recovering the optimal regret bounds for the tabular and factored models are not the main focus of this work.*

Remark 5 (Minimal exponential family)**.** *In our example of tabular MDPs, we illustrated, out of simplicity, a minimal representation with the support of each transition being the full state space. A minimal representation essentially amounts to assuming that the support of each transition is known (locally at each* (s,a)*, dimension = support*-1*). This is more flexible than assuming that we can transit to all states with a positive probability. Assuming the knowledge of the support is a restriction, but we feel that it is not stringent. For instance, in grid-world MDPs, the supports are indeed known ahead of time. If the support is unknown, then the exponential family need not be minimal. However, an equivalent minimal representation exists. We leave as future work the task of trying to remove the minimality assumption.*

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 $\hat\theta_n$ denotes the most optimistic realization from the confidence ellipsoid Θ_n, i.e., $V^{\pi_t}_{\hat\theta_n,1}(s^t_1)\ge V^{\pi}_{\theta,1}(s^t_1),\ \forall\pi\in\Pi,\ \forall\theta\in\Theta_n$. Therefore, as long as the true parameter $\theta^\star\in\Theta_n$ with high probability, $V^{\pi_t}_{\hat\theta_n,1}(s^t_1)$ gives an optimistic estimate of the value $V^{\pi^\star}_{\theta^\star,1}(s^t_1)$ of the episode. An application of Theorem 1 implies that with probability at least $1-\delta/2$, $\theta^\star\in\Theta_n$ across all episodes and thus, in turn, the cumulative regret $\mathcal{R}(N)\le\sum_{t=1}^T V^{\pi_t}_{\hat\theta_n,1}(s^t_1)-V^{\pi_t}_{\theta^\star,1}(s^t_1)$.

Step 2: Bellman Recursion, Transportation and Martingale Control For any $\theta \in \mathbb{R}^d$, $\pi \in \Pi$ and $V : \mathcal{S} \to \mathbb{R}$, we define the Bellman operator $\forall h \leq H$ as

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

Now, by the Bellman equation, we have $\forall h \leq H$,

$$V^{\pi}_{\theta,h}(s) = \mathcal{T}^{\pi}_{\theta,h}\left(V^{\pi}_{\theta,h+1}\right)(s), \quad (\text{with } V^{\pi}_{\theta,H+1}(s) := 0).$$

Applying the Bellman equation recursively, the cumulative regret can be upper bounded as $\mathcal{R}(N) \leq \sum_{t \leq T, h \leq H} \Big(\mathcal{T}^{\pi_t}_{\hat{\theta}_n,h}\big(V^{\pi_t}_{\hat{\theta}_n,h+1}\big)(s^t_h) - \mathcal{T}^{\pi_t}_{\theta^\star,h}\big(V^{\pi_t}_{\hat{\theta}_n,h+1}\big)(s^t_h) + m^t_h \Big)$, where $\{m^t_h\}_{t,h}$ is a martingale difference sequence satisfying $|m^t_h| \leq 2H$. Therefore, by the Azuma-Hoeffding inequality (Boucheron et al., 2013), with probability at least $1-\delta/2$, we obtain $\sum_{t,h} m^t_h \leq 2H\sqrt{2N\ln(2/\delta)}$. Since, by design, $V^{\pi_t}_{\hat{\theta}_n,h+1}(s) \leq H$, we can now control the Bellman differences using transportation inequalities (Lemma 1 in Appendix A) as

$$\mathcal{T}^{\pi_t}_{\hat{\theta}_n,h}\big(V^{\pi_t}_{\hat{\theta}_n,h+1}\big)(s^t_h) - \mathcal{T}^{\pi_t}_{\theta^\star,h}\big(V^{\pi_t}_{\hat{\theta}_n,h+1}\big)(s^t_h)$$
$$\lessapprox H\Big(\sqrt{\mathrm{KL}_{s^t_h,a^t_h}(\theta_n,\hat{\theta}_n)} + \sqrt{\mathrm{KL}_{s^t_h,a^t_h}(\theta_n,\theta^\star)} + \mathrm{KL}_{s^t_h,a^t_h}(\theta_n,\theta^\star)\Big).$$

Step 3: Controlling Sum of KL Divergences We first approximate the KL divergence b/w $P_\theta(\cdot|s,a)$ and $P_{\theta'}(\cdot|s,a)$ using curvature properties of the log-partition function as $\frac{\alpha}{2}\|\theta'-\theta\|^2_{G_{s,a}} \leq \mathrm{KL}_{s,a}(\theta,\theta') \leq \frac{\beta}{2}\|\theta'-\theta\|^2_{G_{s,a}}$. This follows from a second-order Taylor approximation of the log-partition function together with the definition of β. Then, for any $\theta \in \Theta_n$, we obtain $\forall (s,a)$, $\mathrm{KL}_{s,a}(\theta_n,\theta) \leq (\beta/\alpha)\cdot\beta_n(\delta)\left\|\overline{G}_n^{-1}G_{s,a}\right\|$, where $\overline{G}_n := G_n + \alpha^{-1}\eta\mathbb{A}$ and $G_n := \sum_{\tau=1}^{t-1}\sum_{h=1}^{H} G_{s^\tau_h,a^\tau_h}$. Now $\left\|\overline{G}_n^{-1}G_{s,a}\right\| \leq \frac{\alpha}{\eta}\left\|\mathbb{A}^{-1}G_{s,a}\right\| \leq \beta B_{\varphi,\mathbb{A}}/\eta$, $\forall (s,a)$ and $\overline{G}_{n+H} = \overline{G}_n + \sum_{h=1}^{H} G_{s^t_h,a^t_h}$. Therefore, we deduce that $\left\|\overline{G}_n^{-1}G_{s,a}\right\| = \left\|\big(I + \overline{G}_n^{-1}\sum_{h\leq H} G_{s^t_h,a^t_h}\big)\overline{G}_{n+H}^{-1}G_{s,a}\right\| \leq (1 + \beta B_{\varphi,\mathbb{A}}H/\eta)\left\|\overline{G}_{n+H}^{-1}G_{s,a}\right\|$. Using spectral properties of (p.s.d.) matrices, we can now show that

$$\sum_{t,h}\left\|\overline{G}_{n+H}^{-1}G_{s^t_h,a^t_h}\right\| \leq \log\det(I + \alpha\eta^{-1}\mathbb{A}^{-1}G_N) \ ,$$

which is further upper bounded by $\gamma_N = d\log\big(1+\beta\eta^{-1}B_{\varphi,\mathbb{A}}N\big)$. Therefore, since β_n is monotone increasing in n, we have for any $\theta \in \Theta_n$,

$$\sum_{t,h}\mathrm{KL}_{s^t_h,a^t_h}(\theta_n,\theta) \leq (\beta/\alpha)\cdot(1+\beta B_{\varphi,\mathbb{A}}H/\eta)\beta_N(\delta)\gamma_N \ .$$

Final Step: We note that, by design, $\hat{\theta}_n \in \Theta_n$ and by Theorem 1, $\theta^\star \in \Theta_n$. Further, by Cauchy-Schwartz inequality, $\sum_{t,h}\sqrt{\mathrm{KL}_{s^t_h,a^t_h}(\theta,\theta')} \leq \sqrt{N\sum_{t,h}\mathrm{KL}_{s^t_h,a^t_h}(\theta,\theta')}$. The proof now can be completed by putting all the steps together and applying a union bound.

6 CONCLUDING REMARKS

We have provided a new framework to express shared, linear, structure in large, complex MDP problems, relying on the expressive power of exponential family models. We hope this opens the door on more connections to be drawn between learning in dynamical systems on the one hand, and statistical models and guarantees on the other. Several questions emerge, including whether it is possible to sharpen the regret bounds derived here, by employing/building tighter confidence sets, and making full use of the linear structure (still open for linear bandits, see Lattimore and Szepesvari (2017); Magureanu (2018)), and whether richer information structures can be cast into this framework, e.g., partially observed MDPs (POMDPs).

Acknowledgements This work has been supported by CPER Nord-Pas-de-Calais/FEDER DATA Advanced data science and technologies 2015-2020, the French Ministry of Higher Education and Research, Inria, Scool, and the French Agence Nationale de la Recherche (ANR), under grant ANR-16-CE40-0002 (the BADASS project). SRC and AG are supported by a Google Ph.D. Fellowship.

References

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, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

SI Amari. Information geometry. *Contemporary Mathematics*, 203:81–96, 1997.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2001.

Stéphane Boucheron, Gábor Lugosi, Pascal Massart, and Michel Ledoux. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, Oxford, 2013.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 844–853. JMLR. org, 2017.

Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3197–3205, 2019.

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 30 (NIPS)*, pages 5711–5721, 2017.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Regret bounds for kernel-based reinforcement learning. *arXiv preprint arXiv:2004.05599*, 2020.

Audrey Durand, Odalric-Ambrym Maillard, and Joelle Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *arXiv preprint arXiv:1708.00768*, 2017.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pages 12203–12213, 2019.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

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

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Conference on Learning Theory*, pages 861–898. PMLR, 2015.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr): 1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored MDPs. In *IJCAI*, volume 16, pages 740–747, 1999.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. *arXiv preprint arXiv:1610.04491*, 2016.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737, 2017.

Stefan Magureanu. *Efficient Online Learning under Bandit Feedback*. PhD thesis, KTH Royal Institute of Technology, 2018.

Ronald Ortner and Daniil Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 1763–1771, 2012.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 1466–1474, 2014a.

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored MDPs. In *Advances in Neural Information Processing Systems*, pages 604–612, 2014b.

Ian Osband, Dan Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 3003–3011, 2013.

Vern I Paulsen and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge University Press, 2016.

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

Martin L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning (ICML)*, 2010.

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, 2019.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020.

Grady Williams, Andrew Aldrich, and Evangelos A Theodorou. Model predictive path integral control: From theory to parallel computation. *Journal of Guidance, Control, and Dynamics*, 40(2):344–357, 2017.

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

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.