Bilinear Exponential Family of MDPs: Frequentist Regret Bound with Tractable Exploration & Planning

Reda Ouhamma, Debabrota Basu, Odalric Maillard

Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, F-59000
reda.ouhamma@gmail.com, debabrota.basu@inria.fr, odalric.maillard@inria.fr

Abstract

We study the problem of episodic reinforcement learning in continuous state-action spaces with unknown rewards and transitions. Specifically, we consider the setting where the rewards and transitions are modeled using parametric bilinear exponential families. We propose an algorithm, that a) uses penalized maximum likelihood estimators to learn the unknown parameters, b) injects a calibrated Gaussian noise in the parameter of rewards to ensure exploration, and c) leverages linearity of the bilinear exponential family transitions with respect to an underlying RKHS to perform tractable planning. We provide a frequentist regret upper-bound for our algorithm which, in the case of tabular MDPs, is order-optimal with respect to H and K, where H is the episode length and K is the number of episodes. Our analysis improves the existing bounds for the bilinear exponential family of MDPs by square root of H and removes the handcrafted clipping deployed in existing RLSVI-type algorithms.

1 Introduction

Reinforcement Learning (RL) is a well-studied and popular framework for sequential decision making, where an agent aims to compute a *policy* that allows her to maximize the accumulated reward over a horizon by interacting with an *unknown* environment (Sutton and Barto 2018).

Episodic RL. In this paper, we consider the episodic finite-horizon MDP formulation of RL, in short *Episodic RL* (Osband, Russo, and Van Roy 2013; Azar, Osband, and Munos 2017; Dann, Lattimore, and Brunskill 2017). Episodic RL is a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbb{P}, r, K, H \rangle$, where the state (resp. action) space $\mathcal{S}$ (resp. $\mathcal{A}$) might be continuous. In episodic RL, the agent interacts with the environment in episodes consisting of H steps. Episode k starts by observing state s_1^k. Then, for $t = 1, \dots H$, the agent draws action a_t^k from a (possibly time-dependent) policy $\pi_t(s_t^k)$, observes the reward $r(s_t^k, a_t^k) \in [0, 1]$, and transits to a state $s_{t+1}^k \sim \mathbb{P}(. \mid s_t^k, a_t^k)$ according to the transition function $\mathbb{P}$. The performance of a policy π is measured by the total expected reward V_1^π starting from a state $s \in \mathcal{S}$, the value function and the state-action value functions at step $h \in [H]$ are defined as

$$V_h^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=h}^{H} r(s_t, a_t) \mid s_h = s\right],$$

$$\text{and} \quad Q_h^\pi(s, a) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=h}^{H} r(s_t, a_t) \mid s_h = s, a_h = a\right].$$

Here, computing the policy leading to maximization of cumulative reward requires the agent to strategically control the actions in order to learn the transition functions and reward functions as precisely as required. This tension between learning the unknown environment and reward maximization is quantified as *regret*: the typical performance measure of an episodic RL algorithm. *Regret* is defined as the difference between the *expected cumulative reward* or *value* collected by the optimal agent that knows the environment and the expected cumulative reward or value obtained by an agent that has to learn about the unknown environment. Formally, the regret over K episodes is

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

Key Challenges. *The first challenge in episodic RL is to tackle the exploration–exploitation trade-off.* This is traditionally addressed with the *optimism principle* that either carefully crafts optimistic upper bounds on the value functions (Azar, Osband, and Munos 2017), or maintains a posterior on the parameters to perform posterior sampling (Osband, Russo, and Van Roy 2013), or perturbs the value function estimates with calibrated noise (Osband, Van Roy, and Wen 2016). Though the first two approaches induce theoretically optimal exploration, they might not yield tractable algorithms for large/continuous state-action spaces as they either involve optimization in the optimistic set or maintaining a high-dimensional posterior. Thus, *we focus on extending the third approach of* Randomized Least-Square Value Iteration (RLSVI) *framework, and inject noise only in rewards to perform tractable exploration.*

The second challenge, which emerges *for continuous state-action spaces*, *is to learn a parametric functional approximation of either the value function or the rewards and*

transitions in order to perform planning and exploration. Different functional representations (or models), such as linear (Jin et al. 2020), bilinear (Du et al. 2021), and bilinear exponential families (Chowdhury, Gopalan, and Maillard 2021), are studied in literature to develop optimal algorithms for episodic RL with continuous state-action spaces. Since the linear assumption is restrictive in real-life -where non-linear structures are abundant-, generalized representations have obtained more attention recently (Chowdhury, Gopalan, and Maillard 2021; Li et al. 2021; Du et al. 2021; Foster et al. 2021). The BEF model is of special interest as it is expressive enough to represent tabular MDPs (discrete state-action), factored MDPs (Kearns and Koller 1999), and linearly controlled dynamical systems (such as Linear Quadratic Regulators (Abbasi-Yadkori and Szepesvári 2011)) as special cases (Chowdhury, Gopalan, and Maillard 2021). Thus, in this paper, *we study the BEF of MDPs, i.e. the episodic RL setting where the rewards and transition functions can be modeled with bilinear exponential families.*

The third challenge is to perform tractable planning[1] *given the perturbation for exploration and the model class.* Existing work (Osband and Van Roy 2014; Chowdhury, Gopalan, and Maillard 2021) assumes an oracle to perform planning and yield policies that aren't explicit. The main difficulty in such planning approaches is calculating $\int \mathbb{P}(s' \mid s,a)V_h(s)$ for all (s,a) pairs. This is not trivial unless the transition is assumed to be linear and decouples s' from (s,a), which is not known to hold except for tabular MDPs. This challenge received attention recently, *e.g.* (Du et al. 2019) asks when misspecified linear representations are enough for a polynomial sample complexity in several settings. (Shariff and Szepesvári 2020; Lattimore, Szepesvari, and Weisz 2020; Van Roy and Dong 2019) provide positive answers for certain linear settings. In this paper, *we aim to design a tractable planner for the BEF representation.*

In this paper, we aim to address the following question that encompasses the three challenges:

Can we design an algorithm with **tractable exploration** and **planning** for the *bilinear exponential family of MDPs* yielding a **near-optimal frequentist regret bound**?

Contributions. We address this question in three folds.

1. *Formalism:* We assume that neither rewards nor transitions are known, previous efforts on the bilinear exponential family of MDPs assumed knowledge of rewards. This makes the addressed problem harder, practical, and more general. We also observe that though the transition model can represent non-linear dynamics, it implies a linear behavior (see Section 2) in a Reproducing Kernel Hilbert Space (RKHS). This observation contributes to the tractability of planning.

2. *Algorithm:* We propose an algorithm `BEF-RLSVI` that extends the `RLSVI` framework to bilinear exponential families (*cf.* Section 3). `BEF-RLSVI` a) injects calibrated Gaussian noise in the rewards to perform exploration, b) leverages linearity of the transitions with respect to an underlying RKHS to perform tractable planning and c) uses penalized maximum likelihood to learn the parameters corresponding to rewards and transitions (*cf.* Section 4). To the best of our knowledge, *`BEF-RLSVI` is the first algorithm for the bilinear exponential family of MDPs with tractable exploration and planning under unknown rewards and transitions.*

3. *Analysis:* We carefully develop an analysis of `BEF-RLSVI` that yields $\tilde{\mathcal{O}}(\sqrt{d^3H^3K})$ regret which improves the existing regret bound for the BEF of MDPs with known rewards by a factor of $\sqrt{H}$ (Section 3). Our analysis (Section 5) builds on existing analyses of RLSVI-type algorithms (Osband, Van Roy, and Wen 2016), but contrary to them, we remove the need to handcraft a clipping of the value functions (Zanette et al. 2020). We also do not need to *assume* anti-concentration bounds as we can explicitly control it by the injected noise. This was not done previously except for the linear MDPs. We illustrate this comparison in Table **??**. We highlight three technical tools that we used to improve the previous analyses: 1) Using transportation inequalities instead of the simulation lemma reduces a $\sqrt{H}$ factor compared to (Ren et al. 2021), 2) Leveraging the observation that true value functions are bounded enables using an improved elliptical lemma (compared to (Chowdhury, Gopalan, and Maillard 2021)), and 3) Noticing that the norm of features can only be large for a finite amount of time allows us to forgo clipping and reduce a $\sqrt{d}$ factor from the regret compared to (Zanette et al. 2020).

2 Bilinear Exponential Family of MDPs

We introduce the BEF model (Chowdhury, Gopalan, and Maillard 2021) and extend it to parametric rewards. Then, we make an important observation of linearity.

Bilinear exponential family model. We consider both transition and reward kernels to be unknown and modeled with bilinear exponential families. Specifically,

$$\mathbb{P}\left(\tilde{s} \mid s,a\right) = \exp\left(\psi(\tilde{s})^\top M_{\theta^{\mathrm{p}}}\varphi(s,a) - Z^{\mathrm{p}}_{s,a}(\theta^{\mathrm{p}})\right), \quad (1)$$

$$\mathbb{P}\left(r \mid s,a\right) = \exp\left(r\, B^\top M_{\theta^{\mathrm{r}}}\varphi(s,a) - Z^{\mathrm{r}}_{s,a}(\theta^{\mathrm{r}})\right), \quad (2)$$

where $\varphi \in (\mathbb{R}_+^q)^{\mathcal{S}\times\mathcal{A}}$ and $\psi \in (\mathbb{R}_+^p)^{\mathcal{S}}$ are known feature functions, and $B \in \mathbb{R}^p$ is a known scaling factor. The unknown reward and transition parameters are $\theta^{\mathrm{p}}, \theta^{\mathrm{r}} \in \mathbb{R}^d$. $M_{\theta^\cdot} \stackrel{\text{def}}{=} \sum_{i=1}^d \theta^\cdot_i A_i$, where A_i is a known $p \times q$ matrix for each i. Finally, Z denotes the log partition function:

$$Z^{\mathrm{p}}_{s,a}(\theta^{\mathrm{p}}) \stackrel{\text{def}}{=} \log \int_{\mathcal{S}} \exp\left(\psi(\tilde{s})^\top M_{\theta^{\mathrm{p}}}\varphi(s,a)\right) d\tilde{s},$$

Z^{r} is defined similarly. A minor difference with the original BEF model and the one stated here is that, like (Li et al. 2021), we omit a base measure of the form $h(s,\tilde{s},a)$, all the BEF examples provided in (Chowdhury, Gopalan, and Maillard 2021) still hold with this slight restriction. We denote $V^\pi_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},h}$, (resp. $Q^\pi_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},h}$) the value (resp. state-action) value function for policy π in the MDP parameterized by $(\theta^{\mathrm{p}},\theta^{\mathrm{r}})$ at time h. A policy $\pi^\star$ is *optimal* if for all $s \in \mathcal{S}$, $V^{\pi^\star}_{\theta,h}(s) = \max_{\pi\in\Pi} V^\pi_{\theta,h}(s)$. A learning algorithm mini-

[1] By tractable planning, we mean having a planner with (pseudo-)polynomial complexity in the problem parameters, i.e. in the dimension of features, the horizon, and the number of episodes.

Algorithm	Regret	Tractable exploration	Tractable planning	Free of clipping	Model, assumptions
Thompson sampling (Ren et al. 2021)	$\sqrt{d^2H^3K}$ (Bayesian)	✗	✓	N.A	Gaussian $\mathbb{P}$ Known rewards
EXP-UCRL (Chowdhury, Gopalan, and Maillard 2021)	$\sqrt{d^2H^4K}$ (Frequentist)	✗	✗	N.A	Bilinear Exp Family (BEF) known rewards
SMRL (Li et al. 2021)	$\sqrt{d^2H^4K}$	✗	✗	N.A	BEF, known rewards
UCRL-VTR (Ayoub et al. 2020)	$\sqrt{d^2H^4K}$	✗	✗	N.A	Linear mixture model
$\mathcal{F}$-PHE-LSVI (Ishfaq et al. 2021)	$\mathrm{poly}(d_E H)\sqrt{KH}$	✓	✗	✗	Eluder dimension, Tabular
PHE-LSVI (linear-RL)	$\sqrt{d^3H^4K}$				Anti-concentration
UC-MatrixRL (Yang and Wang 2020)	$\sqrt{d^2H^5K}$	✗	✗	N.A	Linear factor MDP
OPT-RLSVI (Zanette et al. 2020)	$\sqrt{d^4H^5K}$	✓	✓	✗	Linear V
BEF-RLSVI (this work)	$\sqrt{d^3H^3K}$	✓	✓	✓	Bilinear Exp Family

Table 1: A comparison of RL Algorithms for MDPs with functional representations.

mizes the (pseudo) regret:

$$\mathcal{R}(K) \triangleq \sum_{k=1}^{K} \left(V_{\theta,1}^{\pi^\star}(s_1^k) - V_{\theta,1}^{\pi^t}(s_1^k) \right). \tag{3}$$

Linearity of transitions. We observe that the popular assumption of linear transition function is a property of the bilinear exponential family, indeed we have:

$$2\psi(s')^\top M_{\theta^{\mathrm{p}}}\varphi(s,a) = -\|(\psi(s') - M_{\theta^{\mathrm{p}}}\varphi(s,a)\|^2 + \|\psi(s')\|^2 + \|M_{\theta^{\mathrm{p}}}\varphi(s,a)\|^2,$$

notice that the quadratic term is the Radial Basis Function (RBF). More precisely, for an RBF kernel with covariance $\Sigma = I_p$ and $k(x,y) \overset{\mathrm{def}}{=} \exp\left(-\|x-y\|^2/2\right)$, we find

$$\mathbb{P}(s' \mid s,a) = \langle \phi^{\mathrm{p}}(s,a), \mu^{\mathrm{p}}(s') \rangle_{\mathcal{H}}, \tag{4}$$

where $\mathcal{H}$ is the RKHS associated with the $k(.,.)$, and

$$\mu^{\mathrm{p}}(s') = (2\pi)^{-p/2}\, k\big(\psi(s'),.\big) \exp\left(\|\psi(s')\|^2/2\right)$$

$$\phi^{\mathrm{p}}(s,a) = k\big(M_{\theta^{\mathrm{p}}}^\top \varphi(s,a),.\big) \exp\Big(\frac{\|M_{\theta^{\mathrm{p}}}^\top \varphi(s,a)\|^2}{2} - Z_{s,a}(\theta^{\mathrm{p}})\Big)$$

In Equation (4), s' is decoupled from (s,a), we see hereafter why this is crucial to reducing the complexity of planning.

Remark 1. *To our knowledge, only (Ren et al. 2021) and this work provide examples of linear transitions in RL with continuous state-actions. The former consider Gaussian transitions with unknown mean ($f^\star(s,a)$) and known variance, i.e. an instance of the BEF model where $\psi(s') = (s', \|s'\|^2)$ and $M_\theta\varphi(s,a) = (f_\theta(s,a)/\sigma^2, -1/\sigma^2)$.*

Importance of linearity. To understand the planning challenge in RL, recall the Bellman equation:

$$Q_h^\pi(s,a) = r(s,a) + \int_{\tilde{s}\in\mathcal{S}} P(s' \mid s,a) V_{h+1}^\pi(\tilde{s}) d\tilde{s},$$

We must approximate the integral at the R.H.S.for $(s,a) \in \mathcal{S}\times\mathcal{A}$. For a tabular MDP with $|S|$ states and $|A|$ actions, we need to evaluate $(Q_h^\pi)_{h\in[H]}$, i.e. to approximate $|S|\times|A|\times H$ integrals per episode, which can be very expensive. However, if the transition model is linear (Equation (4)), then

$$Q_{\theta,h}^\pi(s,a) = r(s,a) + \left\langle \phi^{\mathrm{p}}(s,a), \int_{\mathcal{S}} \mu^{\mathrm{p}}(\tilde{s}) V_{\theta,h+1}^\pi(\tilde{s}) d\tilde{s} \right\rangle. \tag{5}$$

When $\phi^{\mathrm{p}}, \mu^{\mathrm{p}} \in \mathbb{R}^\tau$, we then obtain $Q_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},h}$ by computing τ integrals per timestep. For our model, although ϕ^{p} and μ^{p} are infinite dimensional, we show in Section 4 (§ planning) that the planning complexity is still significantly reduced.

3 BEF-RLSVI: Algorithm Design and Frequentist Regret Bound

We formally introduce the Bilinear Exponential Family Randomized Least Squares Value Iteration (BEF-RLSVI) algorithm and provide a high probability regret bound for it.

Algorithm 1: BEF-RLSVI

Input: failure rate δ, constants $\alpha^{\mathrm{p}}, \eta$ and $(x_k)_{k\in[K]} \in \mathbb{R}^+$
for episode $k = 1, 2, \ldots$ **do**
  Observe initial state s_1^k
  Sample noise $\xi_k \sim \mathcal{N}\big(0, x_k(\bar{G}_k^{\mathrm{p}})^{-1}\big)$ such that $\bar{G}_k^{\mathrm{p}} = \frac{\eta}{\alpha^{\mathrm{p}}}\mathbb{A} + \sum_{\tau=1}^{k-1}\sum_{h=1}^{H}(\varphi(s_h^\tau,a_h^\tau)^\top A_i^\top A_j \varphi(s_h^\tau,a_h^\tau))_{i,j\in[d]}$
  Perturb reward parameter: $\tilde{\theta}^{\mathrm{r}}(k) = \hat{\theta}^{\mathrm{r}}(k) + \xi_k$
  Compute $(Q^k_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}},h})_{h\in[H]}$ via Bellman-backtracking, see Algorithm 2
  for $h = 1, \ldots, H$ **do**
    Pull action $a_h^k = \arg\max_a Q_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}},h}(s_h^k, a)$
    Observe reward $r(s_h^k, a_h^k)$ and state s_{h+1}^k.
  end for
  Update the penalized ML estimators $\hat{\theta}^{\mathrm{p}}(k), \hat{\theta}^{\mathrm{r}}(k)$, see Equation (6) and Equation (7)
end for

BEF−RLSVI: Algorithm Design

BEF−RLSVI is based on RLSVI (Osband, Van Roy, and Wen 2016) except it perturb the reward parameter only. The latter is reminiscent of Thompson Sampling, yet more explicit and with a better control of the optimism probability.

In line. 4 of Algorithm. 1, BEF−RLSVI explores by a Gaussian perturbation of the reward. Contrary to optimistic approaches, this method is explicit and more efficient since it does not a involve high-dimensional optimization.

Algorithm 2: Bellman Backtracking

1: **Input** Parameters $\hat{\theta}^{\mathrm{p}}, \tilde{\theta}^{\mathrm{r}}$, initialize $\tilde{\theta} = (\tilde{\theta}^{\mathrm{r}}, \hat{\theta}^{\mathrm{p}})$ and for all $s \in \mathcal{S}, V_{H+1}(s) = 0$
2: **for** steps $h = H-1, H-2, \cdots, 0$ **do**
3: $\quad Q_{\tilde{\theta},h}(s,a) = \langle \phi^{\mathrm{p}}(s,a), \int V_{\tilde{\theta},h+1}(s')\mu^{\mathrm{p}}(s')ds' \rangle_{\mathcal{H}} + \mathbb{E}^{\tilde{\theta}^{\mathrm{r}}}_{s,a}[r]$
4: **end for**

Line 3 of Algorithm. 2 can be approximated without damaging the regret in $\mathcal{O}(pH^3K\log(HK))$ time (*cf.* § planning, Section 4). Therefore, planning is tractable.

Remark 2. *The observation of linearity (cf. Equation.* (5) *and Line 3) does not reduce BEF MDPs to linear MDPs because the former holds in an RKHS. Also, linearity is not in the representation parameter. Therefore, linear RL algorithms do not readily solve the BEF MDPs.*

BEF−RLSVI: Regret Upper-Bound

We consider the standard smoothness assumptions on the model (Chowdhury, Gopalan, and Maillard 2021; Jun et al. 2017; Lu, Meisami, and Tewari 2021).

Assumption 1. *There exist constants $\alpha^{\mathrm{p}}, \alpha^{\mathrm{r}}, \beta^{\mathrm{p}}, \beta^{\mathrm{r}} > 0$, such that the representation model satisfies, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and for all $\theta, x \in \mathbb{R}^d$*

$$\alpha^{\mathrm{p}} \leq x^\top C^{\theta}_{s,a}[\psi]x \leq \beta^{\mathrm{p}}, \alpha^{\mathrm{r}} \leq \mathbb{V}\mathrm{ar}^{\theta}_{s,a}(r)\, x^\top B^\top Bx \leq \beta^{\mathrm{r}},$$

where $\mathbb{V}\mathrm{ar}^{\theta}_{s,a}(r) \triangleq \left(\mathbb{E}^{\theta}_{s,a}\left[r^2\right] - \mathbb{E}^{\theta}_{s,a}\left[r\right]^2\right)$*, and:*

$$\mathbb{C}^{\theta}_{s,a}\left[\psi\left(s'\right)\right] \triangleq \mathbb{E}_{\mathbb{P}_\theta|s,a}\left[\|\psi\left(s'\right)\|^2\right] - \left\|\mathbb{E}_{\mathbb{P}_\theta|s,a}\left[\psi\left(s'\right)\right]\right\|^2$$

The above inequalities imply a control over the eigenvalues of the Hessian matrices of the log-normalizers (*cf.* Appendix **??**). *e.g.* for the Gaussian distribution, α (resp β) is a lower (resp upper) bound on the variance.

Theorem 2 (Regret bound)**.** *Define $\mathbb{A} \triangleq (\mathrm{tr}(A_iA_j^\top))_{i,j\in[d]}$ and $G_{s,a} \triangleq (\varphi(s,a)^\top A_i^\top A_j \varphi(s,a))_{i,j\in[d]}$. Under Assumption 1 and further assuming*

1. $\max\{\|\theta^{\mathrm{r}}\|_{\mathbb{A}}, \|\theta^{\mathrm{p}}\|_{\mathbb{A}}\} \leq B_{\mathbb{A}}$, $\|\mathbb{A}^{-1}G_{s,a}\| \leq B_{\varphi,\mathbb{A}}$ *and* $\mathbb{E}_{\theta^{\mathrm{r}}}[r(s,a)] \in [0,1]$ *for all* (s,a).
2. $x_k \geq \left(H\sqrt{\frac{\beta^{\mathrm{p}}\beta^{\mathrm{p}}(K,\delta)}{\alpha^{\mathrm{p}}\alpha^{\mathrm{r}}}} + \frac{\sqrt{\beta^{\mathrm{r}}\beta^{\mathrm{r}}(K,\delta)\min\{1, \frac{\alpha^{\mathrm{p}}}{\alpha^{\mathrm{r}}}\}}}{2\alpha^{\mathrm{r}}}\right)^2$,

then for all $\delta \in (0,1]$, with probability at least $1-\delta$,

$$\mathcal{R}(K) \lesssim \sqrt{d^3H^3K}$$

where $\lesssim$ hides constant and logarithmic terms, refer to Theorem. **??** *in Appendix.* **??** *for the detailed regret bound.*

Comparison with other bounds. The closest result is from (Chowdhury, Gopalan, and Maillard 2021), it considers the same model for transitions but with known rewards. They propose a UCRL-type and PSRL-type algorithm, which achieve a $\widetilde{O}(\sqrt{d^2H^4K})$ regret. There are two notable algorithmic differences with BEF−RLSVI. First, they use intractable-optimistic upper bounds or high-dimensional posteriors, while we do explore with explicit perturbations. The second difference is in planning: while they assume access to a planning oracle, we do it explicitly with pseudo-polynomial complexity (Section 4). Moreover, we improve the regret bound by $\sqrt{H}$ thanks to an improved analysis, (*cf.* Lemma **??**). But similar to all RLSVI-type algorithms, we pick up an extra $\sqrt{d}$ (*cf.* (Abeille and Lazaric 2017)).

(Zanette et al. 2020) proposes a variant of RLSVI for continuous state-action spaces, where there are low-rank models of transitions and rewards. They show a regret bound $R(K) = \widetilde{O}(\sqrt{d^4H^5K})$, which is larger than that of BEF−RLSVI by $O(\sqrt{dH^2})$. In algorithm design, we improve on their work by removing the need to carefully clip the value function. Analytically, our model allows us to use transportation inequalities (*cf.* Lemma **??**) instead of the simulation lemma, which saves us a $\sqrt{H}$ factor.

(Ren et al. 2021) considers Gaussian transitions, i.e. $s' = f^*(s,a) + \epsilon$ such that $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$. This is a particular case of our model. They propose to use Thompson Sampling, and have the merit of being the first to have observed linearity of the value function from this transition structure. But they do not connect it to the finite dimensional approximation of (Rahimi and Recht 2007) unlike us (Section 4). Finally, they show a Bayesian regret bound of $O(\sqrt{d^2H^3K})$. This notion of regret is weaker than frequentist regret, hence this result is not directly comparable with Theorem 2.

Tightness of regret bound. A lower bound for episodic RL with continuous state-action spaces is still missing. However, for tabular RL, (Domingues et al. 2021) proves a lower bound of order $\Omega(\sqrt{H^3SAK})$. If we represent a tabular MDP in our model, we would need $d = S^2 \times A$ parameters (Section 4.3, (Chowdhury, Gopalan, and Maillard 2021)). In this case, our bound becomes $R(K) = O(\sqrt{(S^2A)^3H^3K})$, which is clearly not tight is S and A. This is understandable due to the relative generality of our setting. We are however positively surprised that **our bound is tight in terms of its dependence on H and K.**

4 Algorithm Design: Building Blocks Of BEF−RLSVI

We present necessary details about BEF−RLSVI and discuss the key algorithm design techniques.

Estimation of parameters. We estimate transitions and rewards from observations similar to EXP−UCRL (Chowdhury, Gopalan, and Maillard 2021), *i.e.* by using a penalized maximum likelihood estimator

$$\hat{\theta}^{\mathrm{p}}(k) \in \underset{\theta\in\mathbb{R}^d}{\arg\min} \sum_{\substack{1\leq t\leq k\\ 1\leq h\leq H}} -\log \mathbb{P}_\theta\left(s^t_{h+1} \mid s^t_h, a^t_h\right) + \eta\,\mathrm{pen}(\theta).$$

Here, $\mathrm{pen}(\theta)$ is the trace-norm penalty: $\mathrm{pen}(\theta) = \frac{1}{2}\|\theta\|_{\mathbb{A}}$ where $\mathbb{A} = (\mathrm{tr}(A_i A_j^\top))_{i,j}$. By properties of the exponential family, the penalized ML estimator verifies, for $i \leq d$:

$$\sum_{\substack{1\leq t\leq k\\ 1\leq h\leq H}} \left(\psi\left(s_{h+1}^t\right) - \mathbb{E}^{\hat{\theta}_k^{\mathrm{p}}}_{s_h^t, a_h^t}\left[\psi\left(s'\right)\right]\right)^\top A_i \varphi\left(s_h^t, a_h^t\right) = \eta \nabla_i \mathrm{pen}\left(\hat{\theta}_k^{\mathrm{p}}\right) \tag{6}$$

The above can be solved in closed form for simple distributions, like Gaussian, but it can be involved for other distribution (*cf.* Appendix **??**). For the reward, θ_r is defined similarly:

$$\hat{\theta}^{\mathrm{r}}(k) \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \sum_{\substack{1\leq t\leq k\\ 1\leq h\leq H}} -\log \mathbb{P}_\theta\left(r_t \mid s_h^t, a_h^t\right) + \eta\, \mathrm{pen}(\theta),$$

Then, for all $i \in [d]$:

$$\sum_{\substack{1\leq t\leq k\\ 1\leq h\leq H}} \left(r_t - \mathbb{E}^{\hat{\theta}_k^{\mathrm{r}}}_{s_h^t, a_h^t}\left[r\right]\right) B^\top A_i \varphi\left(s_h^t, a_h^t\right) = \eta \nabla_i \,\mathrm{pen}\left(\hat{\theta}_k^{\mathrm{r}}\right) \tag{7}$$

Exploration. A significant challenge in RL is handling exploration in continuous spaces. The majority of the literature is split between intractable, upper confidence bound-style optimism or Thompson sampling algorithms with high-dimensional posterior and guarantees only in terms of Bayesian regret. In `BEF-RLSVI`, we adopt the approach of reward perturbation motivated by the `RLSVI`-framework (Zanette et al. 2020; Osband, Van Roy, and Wen 2016). We show that perturbing the reward estimation can guarantee optimism with a constant probability, *i.e.* there exists $\nu \in (0, 1]$ such that for all $k \in [K]$ and $s_1^k \in \mathcal{S}$,

$$\mathbb{P}\left(\tilde{V}_1(s_1^k) - V_1^\star(s_1^k) \geq 0\right) \geq \nu.$$

(Zanette et al. 2020) proves that this suffices to bound the learning error. However, their method clashes with not clipping the value function, as it modifies the probability of optimism. Thus, (Zanette et al. 2020) proposes an involved clipping procedure to handle the issue of unstable values. Instead, by careful geometric analysis (*cf.* Lemma **??**), we bound the occurrences of the unstable values, and in turn, upper bound the regret without clipping. Note that unlike (Ishfaq et al. 2021), `BEF-RLSVI` does not guarantee that the estimated value function is optimistic but still is able to control the learning error (*cf.* Section 5).

Planning. Recall that with our model assumptions, we can write the state-action value function linearly (Equation (5)). Using `BEF-RLSVI`, we have at step h:

$$Q^\pi_{\hat{\theta}^{\mathrm{p}}, \tilde{\theta}^{\mathrm{r}}, h}(s, a) = \mathbb{E}_{\tilde{\theta}^{\mathrm{r}}}\left[r(s, a)\right] + \left\langle \phi^{\mathrm{p}}(s, a), \int_{\mathcal{S}} \mu^{\mathrm{p}}(\tilde{s}) V^\pi_{\hat{\theta}^{\mathrm{p}}, \tilde{\theta}^{\mathrm{r}}, h+1}(\tilde{s}) d\tilde{s} \right\rangle.$$

Then, we select the best action greedily to compute $Q_h(s, a)$. Although ϕ^{p} and ψ^{p} are infinite-dimensional, approximating them (*cf.* next paragraph) with features of dimension $\mathcal{O}(pH^2K\log(HK))$ doesn't increasing the regret. Thus, the planning is done in $\mathcal{O}(pH^3K\log(HK))$, which is pseudo-polynomial in p, H and K, *i.e.* tractable.

For details about the finite-dimensional approximation of our transition kernel, refer to Appendix **??**. Now, we highlight the schematic of a finite-dimensional approximation of ϕ^{p} and ψ^{p}. We proceed in three steps. **1)** We have with high probability $\mathbb{S}(V_{\hat{\theta}^{\mathrm{p}}, \tilde{\theta}^{\mathrm{r}}, h}) \leq dH^{3/2}$ (Section 5). **2)** If we have a uniform ϵ-approximation of $\mathbb{P}_{\theta^{\mathrm{p}}}$, we show that using it incurs at most an extra $\mathcal{O}(\epsilon dH^{5/2}K)$ regret. **3)** Finally, following (Rahimi and Recht 2007), we approximate uniformly the shift invariant kernels, here the RBF in Equation (4), within ϵ error and with features of dimensions $\mathcal{O}(p\epsilon^{-2}\log\frac{1}{\epsilon^2})$, where p is dimension of ψ. Associating these three elements and choosing $\epsilon = 1/\sqrt{(H^2K)}$, we establish our claim.

5 Theoretical Analysis: Proof Outline

To convey the novelties in our analysis, we provide a proof sketch for Theorem 2. We start by decomposing the regret into an estimation loss and a learning error, as given below

$$\begin{aligned} R(K) &= \sum_{k=1}^K (V^\star_{\theta^{\mathrm{p}}, \theta^{\mathrm{r}}, 1} - V^{\pi_k}_{\theta^{\mathrm{p}}, \theta^{\mathrm{r}}, 1})(s_{1k}) \\ &= \sum_{k=1}^K (\underbrace{V^\star_{\theta^{\mathrm{p}}, \theta^{\mathrm{r}}, 1} - V^{\pi_k}_{\hat{\theta}^{\mathrm{p}}, \tilde{\theta}^{\mathrm{r}}, 1}}_{\text{Learning}} + \underbrace{V^{\pi_k}_{\hat{\theta}^{\mathrm{p}}, \tilde{\theta}^{\mathrm{r}}, 1} - V^{\pi_k}_{\theta^{\mathrm{p}}, \theta^{\mathrm{r}}, 1}}_{\text{Estimation}})(s_{1k}). \end{aligned} \tag{8}$$

For the **estimation error**, we use smoothness arguments with concentrations of parameters up to some novelties. Regarding the **learning error**, we show that the injected noise ensures a constant probability of anti-concentration. Applying Assumption 1 and Lemma **??** leads to the upper-bound.

Bounding the Estimation Error

We further decompose the estimation error into the errors in estimating transitions and rewards.

$$\begin{aligned} V^\pi_{\hat{\theta}^{\mathrm{p}}, \tilde{\theta}^{\mathrm{r}}}(s_{1k}) - V^\pi_{\theta^{\mathrm{p}}, \theta^{\mathrm{r}}}(s_{1k}) &= \underbrace{V^\pi_{\hat{\theta}^{\mathrm{p}}, \theta^{\mathrm{r}}}(s_{1k}) - V^\pi_{\theta^{\mathrm{p}}, \theta^{\mathrm{r}}}(s_{1k})}_{\text{transition estimation}} \\ &\quad + \underbrace{V^\pi_{\hat{\theta}^{\mathrm{p}}, \tilde{\theta}^{\mathrm{r}}}(s_{1k}) - V^\pi_{\hat{\theta}^{\mathrm{p}}, \theta^{\mathrm{r}}}(s_{1k})}_{\text{reward estimation}} \end{aligned} \tag{9}$$

Transition estimation Lemma **??** yields the upper-bound $H \min\left\{1, \sum_{h=1}^H \sqrt{2\,\mathrm{KL}_{s_{hk}, a_{hk}}(\theta^{\mathrm{p}}, \hat{\theta}^{\mathrm{p}})}\right\}$, then Lemma **??** shaves a $\sqrt{H}$ compared to (Chowdhury and Gopalan 2019).

Reward estimation Earlier works use clipping to control this error, but for RLSVI it can reduce the optimism probability. (Zanette et al. 2020) proposes an involved clipping depending on the feature norms, which is somewhat delicate to analyze. We remedy the situation by improving the proof and not the algorithm. Indeed, consider the rounds $\left\{k \in [K], \exists h : \|(A_i\varphi(s_h^k, a_h^k))_{i\in[d]}\|_{(\bar{G}_k^{\mathrm{p}})^{-1}} \geq 1\right\}$, the latter are why clipping is necessary. Thanks to Lemma **??**, we know that the number of such rounds is $\mathcal{O}(d)$. Surprisingly, it depends neither on H nor on K. We show that the "bad

rounds" incur at most $O(d^{3/2}H^2)$ regret, independent of K. Thus, we can omit clipping for free.

Remark 3. *For $H = 1$, the forward algorithm (Azoury and Warmuth 2001) handles the span issue. (Ouhamma, Maillard, and Perchet 2021) analyzes it for linear bandits.*

Bounding the Learning Error

We show that the estimated value is optimistic with a constant probability and that it's enough to control the error.

Stochastic optimism. The perturbation ensures a constant probability of optimism. Specifically,

$$\begin{aligned}(V_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}},1} - V^{\star}_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},1})(s_1) &\geq (Q^{\star}_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}},1} - Q^{\star}_1)(s_1,\pi^{\star}(s_1)) \\ &\geq \underbrace{V^{\pi^{\star}}_{\hat{\theta}^{\mathrm{p}},\theta^{\mathrm{r}}}(s_1) - V^{\pi^{\star}}_{\theta^{\mathrm{p}},\theta^{\mathrm{r}}}(s_1)}_{\text{first term}} + \underbrace{V^{\pi^{\star}}_{\hat{\theta}^{\mathrm{p}},\hat{\theta}^{\mathrm{r}}}(s_1) - V^{\pi^{\star}}_{\hat{\theta}^{\mathrm{p}},\theta^{\mathrm{r}}}(s_1)}_{\text{second term}} \\ &\quad + \underbrace{V^{\pi^{\star}}_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}}}(s_1) - V^{\pi^{\star}}_{\hat{\theta}^{\mathrm{p}},\hat{\theta}^{\mathrm{r}}}(s_1)}_{\text{third term}}\end{aligned}$$

The first and second terms are perturbation free, we handle them like the estimation error, *i.e.* using concentration arguments for $\hat{\theta}^{\mathrm{p}}$ and $\hat{\theta}^{\mathrm{r}}$. For the third term, we use transportation of rewards (Lemma **??**) and anti-concentration of ξ_k (Lemma **??**) to find, with probability at least $1-2\delta$:

$$\begin{aligned}V_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}},1}(s_1) - V^{\star}_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},1}(s_1) \geq \\ \xi_k^{\top}\mathbb{E}_{(\tilde{s}_t)_{t\in[H]}\sim\hat{\theta}^{\mathrm{p}}|s_1^k}\left[\sum_{t=1}^{H}\frac{\mathbb{V}\mathrm{ar}^{\theta^{\mathrm{r}}_j}(r)}{2}\boldsymbol{A}_{\varphi}(\tilde{s}_t)\right]B \\ - Hc(n,\delta)\left\|\sum_{h=1}^{H}\mathbb{E}_{(\tilde{s}_t)_{t\in[H]}\sim\hat{\theta}^{\mathrm{p}}|s_1^k}\left[\boldsymbol{A}_{\varphi}(\tilde{s}_t)\right]\right\|_{(\bar{G}_k^{\mathrm{p}})^{-1}}\end{aligned}$$

where $\boldsymbol{A}_{\varphi}(\tilde{s}_t) = (A_i\varphi(\tilde{s}_t,\pi^{\star}(\tilde{s}_t)))_{i\in[d]}$ and $c(n,\delta) = \left(\sqrt{\beta^{\mathrm{r}}\beta^{\mathrm{r}}(n,\delta)\min\{1,\alpha^{\mathrm{p}}/\alpha^{\mathrm{r}}\}/(2\alpha^{\mathrm{r}})} + \sqrt{\beta^{\mathrm{p}}\beta^{\mathrm{p}}(n,\delta)/\alpha^{\mathrm{p}}}\right)$. Since $\xi_k \sim \mathcal{N}(0, x_k(\bar{G}_k^{\mathrm{p}})^{-1})$, then $\mathbb{P}\left(V^{\pi}_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}},1}(s_1) \geq V^{\star}_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},1}(s_1)\right) \geq \Phi(-1)$, where Φ is the normal CDF. This is ensured by the Gaussian anti-concentration, see Lemma **??**.

From stochastic optimism to error control: Optimistic algorithms require optimism with large probability. On the contrary, `BEF-RLSVI` only requires optimism with a constant probability. Indeed, the policy is always close to a good one thanks to the decreasing estimation error. This part of the proof is similar in spirit to that of (Zanette et al. 2020).

Upper bound on $V_1^{\star}$: Define the optimism event: $\bar{O}_k \triangleq \left\{V_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}}_k,1}(s_1^k) \geq V_1^{\star}(s_1^k)\right\}$. We deduce the upper-bound

$$V_1^{\star}(s_1^k) \leq \mathbb{E}_{\bar{\xi}_k|\bar{O}_k}[V_{\hat{\theta}^{\mathrm{p}},\hat{\theta}^{\mathrm{r}}+\bar{\xi}_k,1}(s_1^k)].$$

Lower bound on $V_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}}}$: Consider $\underline{\mathrm{V}}_1(s_1^k)$ a solution of the following optimization problem

$$\begin{aligned}\min_{\xi_k}\ & V_{\hat{\theta}^{\mathrm{p}},\hat{\theta}^{\mathrm{r}}+\xi_k,1}(s_1^k) \\ \text{subject to:}\quad & \|\xi_k\|_{\bar{G}_k} \leq \sqrt{x_k d\log(d/\delta)},\end{aligned}$$

If the concentration event holds for the perturbation then:

$$\underline{\mathrm{V}}_1(s_1^k) \leq V_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}}}(s_1^k).$$

Combination: Using these upper and lower bounds, we show that with probability at least $1-\delta$,

$$\begin{aligned}V_1^{\star}(s_1^k) - V_{\hat{\theta}^{\mathrm{p}},\hat{\theta}^{\mathrm{r}}+\bar{\xi}_k,1}(s_1^k) \\ \leq \mathbb{E}_{\bar{\xi}_k|\bar{O}_k}[V_{\hat{\theta}^{\mathrm{p}},\hat{\theta}^{\mathrm{r}}+\bar{\xi}_k,1}(s_1^k) - \underline{\mathrm{V}}_1(s_1^k)] \\ \leq \frac{1}{\mathbb{P}(\bar{O}_k)}\Big(\mathbb{E}_{\bar{\xi}_k}[V_{\hat{\theta}^{\mathrm{p}},\hat{\theta}^{\mathrm{r}}+\bar{\xi}_k,1}(s_1^k) - \underline{\mathrm{V}}_1(s_1^k)] \\ - \mathbb{E}_{\bar{\xi}_k|\bar{O}_k^{\mathrm{c}}}[V_{\hat{\theta}^{\mathrm{p}},\hat{\theta}^{\mathrm{r}}+\bar{\xi}_k,1}(s_1^k) - \underline{\mathrm{V}}_1(s_1^k)]\mathbb{P}(\bar{O}_k^{\mathrm{c}})\Big)\end{aligned}$$

The last step follows from the tower rule. Note that the term inside the expectations is positive with high probability but not necessarily in expectation. We follow the lines of the estimation error analysis to complete the proof of Theorem 2. We refer to Appendix **??** for the detailed proof.

Remark 4 (similarity with linear RL)**.** *While the bilinear exponential family model is novel, our proof techniques also hold for linear MDPs. Indeed, **A)** our analysis uses transportation inequalities (Lemma 13) that elegantly bound our regret by the complexity of learning a bilinear form (the exponent of the transition model), **B)** Controlling the latter is like controlling the regret in linear RL and reduces to the analysis of linear bandits, **C)** Our analytical improvements, i.e. Lemma 19 (rendering clipping unnecessary) and Lemma 18, intervene in the the step of the analysis that is identical to Linear RL. Consequently, our contributions (Lemma 18 and Lemma 19) also hold while analyzing Linear RL algorithms.*

6 Related Works: Functional Representations of MDPs With Regret and Tractability

Our work extends the endeavor of using functional representations for regret minimization in continuous state-action MDPs. Now, we posit our contributions in existing literature.

Kernel value function representation. (Ayoub et al. 2020) studies MDPs with a linear mixtures model then extends to an RKHS setting, this generalizes our work and that of (Yang and Wang 2020). However, the paper proposes an Eluder-dimension analysis, for RKHS settings this leads to the result of (Yang and Wang 2020), *i.e.* a regret $H\log(T)^d$ higher than for `BEF-RLSVI`.Recently, (Huang et al. 2021) shows that for RKHS, Eluder dimension and the information gain are strictly equivalent, which brings in the extra factor.

General functional representation. The Eluder dimension is a complexity measure often used to analyze RL with general function space, (Huang et al. 2021) asserts that "common examples of where it is known to be small are function spaces (vector spaces)". (Dai et al. 2018) provides the first convergence guarantee for general nonlinear function representations in the Maximum Entropy RL setting, where entropy of a policy is used as a regularizer to induce exploration. Thus, the analysis cannot address episodic RL, where we have to explicitly ensure exploration with optimism. In the episodic setting, (Wang, Salakhutdinov, and

Yang 2020) leverage the UCB approach for tabular MDPs and function spaces with bounded Eluder dimension, this strategy achieves a and achieve a $\tilde{O}\left(\sqrt{d^4H^2T}\right)$ regret for linear MDPs. (Ishfaq et al. 2021) considers the same setting, proposes an RLSVI based algorithm, and achieves a $\tilde{O}(\sqrt{d^3H^4K})$ for linear MDPs. However, the latter assumes an oracle perturbing the estimation to achieve anti-concentration while maintaining a bounded covering number, which is a counter-intuitive mix of boundedness and anti-concentration. Indeed, (Zanette et al. 2020) studied the linear MDP case, and while it managed to design an ingenious clipping verifying previous assumptions, the method is extremely intricate and the proof is involved and unlikely to extend for general value function spaces. *To concertize our design, we focus on the general but explicit BEF of MDPs than any abstract representation. We also remove the requirement to clip with a novel analysis.*

Bilinear exponential family of MDPs. Exponential families are studied widely in RL theory, from bandits to MDPs (Lu, Meisami, and Tewari 2021; Korda, Kaufmann, and Munos 2013; Filippi et al. 2010; Kveton and Hauskrecht 2006), as an expressive parametric family to design theoretically-grounded model-based algorithms. (Chowdhury, Gopalan, and Maillard 2021) first studies episodic RL with Bilinear Exponential Family (BEF) of transitions, which is linear in both state-action pairs and the next-state. It proposes a regularized log-likelihood method to estimate the model parameters, and two optimistic algorithms with upper confidence bounds and posterior sampling. Due to its generality to unifiedly model tabular MDPs, factored MDPs, and linearly controlled dynamical systems, the BEF-family of MDPs has received increasing attention (Li et al. 2021). (Li et al. 2021) estimates the model parameters based on score matching that enables them to replace regularity assumption on the log-partition function with Fisher-information and assumption on the parameters. Both (Chowdhury, Gopalan, and Maillard 2021; Li et al. 2021) achieve a worst-case regret of order $\tilde{O}(\sqrt{d^2H^4K})$ for known reward. On a different note, (Du et al. 2021; Foster et al. 2021) also introduces a new structural framework for generalization in RL, called bilinear classes as it requires the Bellman error to be upper bounded by a bilinear form. Instead of using bilinear forms to capture non-linear structures, this class is not identical to BEF class of MDPs, and studying the connection is out of the scope of this paper. Specifically, *we address the shortcomings of the existing works on BEF-family of MDPs that assume known rewards, absence of RLSVI-type algorithms, and access to oracle planners.*

Other function classes. The Eluder dimension is usually difficult to compute except for simple spaces or for kernel MDPs. Recently, novel complexity measures have been proposed and they seem to capture realistic scenarios. For instance, Bellman rank (Jiang et al. 2017) captures linear quadratic regulators (like the BEF), the algorithms proposed with this complexity are however still intractable and assume unrealistic cases like finite function classes. Another novel structural framework is Bilinear classes (Du et al. 2021), which captures Factored MDPs, kernel MDPs, and linear quadratic regulators, however, like most literature, the UCB-style provided algorithm is intractable. Also, the paper proved PAC bounds and not regret minimization hence we can't compare them to our result.

Tractable planning and linearity. Planning is a major byproduct of the chosen functional representation. In general, planning can incur high computational complexity if done naïvely. Specially, (Du et al. 2019) shows that for some settings, even with a linear ϵ-approximation of the Q-function, a planning procedure able to produce an ϵ-optimal policy has a complexity at least 2^H. Thus, different works (Shariff and Szepesvári 2020; Lattimore, Szepesvari, and Weisz 2020; Van Roy and Dong 2019) propose to leverage different low-dimensional representations of value functions or transitions to perform efficient planning. Here, we take note from (Ren et al. 2021) that Gaussian transitions induce an explicit linear value function in an RKHS. And generalize this observation with the bilinear exponential. Moreover, using uniformly good features (Rahimi and Recht 2007) to approximate transition dynamics from our model enables us to design a tractable planner. We provide a detailed discussion of this approximation in Section 4. More practically, (Ren et al. 2021; Nachum and Yang 2021) use representations given by random Fourier features (Rahimi and Recht 2007) to approximate the transition dynamics and provide experiments validating the benefits of this approach for high-dimensional Atari-games. *Thus, we propose the first algorithm with tractable planning for BEF-family.*

7 Conclusion and Future Work

In the setting of episodic-RL, we study the bilinear exponential family of MDPs, a representation generalizing real examples of MDPs such as Tabular, Factored, and LQR. We propose the BEF−RLSVI algorithm, the latter explores using a Gaussian perturbation of rewards, and plans tractably (running time of $\mathcal{O}(pH^3K\log(HK))$) thanks to properties of the RBF kernel. Our proof shows that clipping can be forwent, further simplifying this RLSVI-type algorithm. Moreover, we prove a $\sqrt{d^3H^3K}$ frequentist regret bound, which improves over existing work and accommodates unknown rewards. While a lower bound for RL in continuous spaces is still missing, a lower bound for the tabular case shows that our algorithm is order-optimal in H and K. Our only assumption is smoothness of transitions and rewards, *i.e.* that their Hessians' eigenvalues are well behaved, a common assumption in literature (Chowdhury, Gopalan, and Maillard 2021; Jun et al. 2017; Lu, Meisami, and Tewari 2021).

Regarding future work, we believe that our proof techniques can be extended to rewards with bounded variance. Also, we believe that the extra $\sqrt{d}$ in our bound is the price of tractability, as it is present in all tractable algorithms such as RLSVI or LSVI-UCB. We will investigate it further. Finally, we would like to study the practical efficiency of BEF−RLSVI through experiments on tasks with continuous state-action spaces in an extended version of this work.

References

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

Abeille, M.; and Lazaric, A. 2017. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, 176–184. PMLR.

Ayoub, A.; Jia, Z.; Szepesvari, C.; Wang, M.; and Yang, L. 2020. Model-Based Reinforcement Learning with Value-Targeted Regression. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 463–474. PMLR.

Azar, M. G.; Osband, I.; and Munos, R. 2017. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 263–272. PMLR.

Azoury, K. S.; and Warmuth, M. K. 2001. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3): 211–246.

Chowdhury, S. R.; and Gopalan, A. 2019. Online learning in kernelized markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 3197–3205. PMLR.

Chowdhury, S. R.; Gopalan, A.; and Maillard, O.-A. 2021. Reinforcement Learning in Parametric MDPs with Exponential Families. In *International Conference on Artificial Intelligence and Statistics*, 1855–1863. PMLR.

Dai, B.; Shaw, A.; Li, L.; Xiao, L.; He, N.; Liu, Z.; Chen, J.; and Song, L. 2018. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, 1125–1134. PMLR.

Dann, C.; Lattimore, T.; and Brunskill, E. 2017. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30.

Domingues, O. D.; Ménard, P.; Kaufmann, E.; and Valko, M. 2021. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, 578–598. PMLR.

Du, S.; Kakade, S.; Lee, J.; Lovett, S.; Mahajan, G.; Sun, W.; and Wang, R. 2021. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, 2826–2836. PMLR.

Du, S. S.; Kakade, S. M.; Wang, R.; and Yang, L. F. 2019. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*.

Filippi, S.; Cappe, O.; Garivier, A.; and Szepesvári, C. 2010. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23.

Foster, D. J.; Kakade, S. M.; Qian, J.; and Rakhlin, A. 2021. The Statistical Complexity of Interactive Decision Making. *arXiv preprint arXiv:2112.13487*.

Huang, K.; Kakade, S. M.; Lee, J. D.; and Lei, Q. 2021. A short note on the relationship of information gain and eluder dimension. *arXiv preprint arXiv:2107.02377*.

Ishfaq, H.; Cui, Q.; Nguyen, V.; Ayoub, A.; Yang, Z.; Wang, Z.; Precup, D.; and Yang, L. 2021. Randomized Exploration in Reinforcement Learning with General Value Function Approximation. In *International Conference on Machine Learning*, 4607–4616. PMLR.

Jiang, N.; Krishnamurthy, A.; Agarwal, A.; Langford, J.; and Schapire, R. E. 2017. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, 1704–1713. PMLR.

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2137–2143. PMLR.

Jun, K.-S.; Bhargava, A.; Nowak, R.; and Willett, R. 2017. Scalable generalized linear bandits: Online computation and hashing. *arXiv preprint arXiv:1706.00136*.

Kearns, M.; and Koller, D. 1999. Efficient reinforcement learning in factored MDPs. In *IJCAI*, volume 16, 740–747.

Korda, N.; Kaufmann, E.; and Munos, R. 2013. Thompson sampling for 1-dimensional exponential family bandits. *Advances in neural information processing systems*, 26.

Kveton, B.; and Hauskrecht, M. 2006. Solving Factored MDPs with Exponential-Family Transition Models. In *ICAPS*, 114–120.

Lattimore, T.; Szepesvari, C.; and Weisz, G. 2020. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, 5662–5670. PMLR.

Li, G.; Li, J.; Srebro, N.; Wang, Z.; and Yang, Z. 2021. Exponential Family Model-Based Reinforcement Learning via Score Matching. *arXiv preprint arXiv:2112.14195*.

Lu, Y.; Meisami, A.; and Tewari, A. 2021. Low-rank generalized linear bandit problems. In *International Conference on Artificial Intelligence and Statistics*, 460–468. PMLR.

Nachum, O.; and Yang, M. 2021. Provable representation learning for imitation with contrastive fourier features. *Advances in Neural Information Processing Systems*, 34.

Osband, I.; Russo, D.; and Van Roy, B. 2013. (More) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26.

Osband, I.; and Van Roy, B. 2014. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27.

Osband, I.; Van Roy, B.; and Wen, Z. 2016. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, 2377–2386. PMLR.

Ouhamma, R.; Maillard, O.-A.; and Perchet, V. 2021. Stochastic Online Linear Regression: the Forward Algorithm to Replace Ridge. *Advances in Neural Information Processing Systems*, 34: 24430–24441.

Rahimi, A.; and Recht, B. 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Ren, T.; Zhang, T.; Szepesvári, C.; and Dai, B. 2021. A Free Lunch from the Noise: Provable and Practical Exploration for Representation Learning. *arXiv preprint arXiv:2111.11485*.

Shariff, R.; and Szepesvári, C. 2020. Efficient planning in large MDPs with weak linear function approximation. *arXiv preprint arXiv:2007.06184*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Van Roy, B.; and Dong, S. 2019. Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*.

Wang, R.; Salakhutdinov, R.; and Yang, L. F. 2020. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *arXiv preprint arXiv:2005.10804*.

Yang, L.; and Wang, M. 2020. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, 10746–10756. PMLR.

Zanette, A.; Brandfonbrener, D.; Brunskill, E.; Pirotta, M.; and Lazaric, A. 2020. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, 1954–1964. PMLR.