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

Anonymous Author(s)
Affiliation
Address
email

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, BEF-RLSVI, 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 exponential family with respect to an underlying RKHS to perform tractable planning. We further provide a frequentist regret analysis of BEF-RLSVI that yields an upper bound of $\tilde{\mathcal{O}}(\sqrt{d^3H^3K})$, where d is the dimension of the parameters, 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 $\sqrt{H}$ and removes the handcrafted clipping deployed in existing RLSVI-type algorithms. Our regret bound is order-optimal with respect to H and K.

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 [SB18].

Episodic RL. In this paper, we consider the episodic finite-horizon MDP formulation of RL, in short *Episodic RL* [ORVR13, AOM17, DLB17]. 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

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

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

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

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 key 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 (or state-action value) functions [AOM17], or maintains a posterior on the parameters to perform posterior sampling [ORVR13], or perturbs the value (or state-action value) function estimates with calibrated noise [OVRW16]. 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 [JYWJ20], bilinear [DKL+21], and bilinear exponential families [CGM21], 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 [CGM21, LLS+21, DKL+21, FKQR21]. The bilinear exponential family model is of special interest as it is expressive enough to represent tabular MDPs (discrete state-action), factored MDPs [KK99], linear MDPs [JYWJ20], linearly controlled dynamical systems (such as Linear Quadratic Regulators [AYS11]) as special cases [CGM21]. Thus, in this paper, *we study the bilinear exponential family of MDPs, i.e. the episodic RL setting where the rewards and transition functions can be modelled with bilinear exponential families.*

The third challenge is to perform tractable planning[1] *given the perturbation for exploration and the model class.* Existing work [OVR14, CGM21] assumes an oracle to perform planning and yield policies that aren't explicit. The main difficulty in such planning approaches is that dynamic programming requires 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. Much ink has been spilled about this challenge recently, *e.g.* [DKWY19] asks when misspecified linear representations are enough for a polynomial sample complexity in several settings. [SS20, LSW20, VRD19] provide positive answers for specific linear settings. In this paper, *we aim to address this issue by designing a tractable planner for the bilinear exponential family representation.*

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

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

Our Contributions. Our contributions to this question are three-fold.

1. *Formalism:* We assume that neither rewards nor transitions are known, whereas existing efforts on the bilinear exponential family of MDPs assume 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 Reproducible 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 (see Section 3). `BEF-RLSVI` a) injects calibrated Gaussian noise in the rewards to perform exploration, b) leverages the linearity of the transition model with respect to an underlying RKHS to perform tractable planning and c) uses penalized maximum likelihood estimators to learn the parameters corresponding to rewards and transitions (see 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.*

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

Table 1: A comparison of RL Algorithms for continuous state-actions with functional representations.

Algo	Regret	Tractable exploration	Tractable planning	Free of clipping	Model, assumptions
Thompson sampling [RZSD21]	$\sqrt{d^2H^3K}$ (Bayesian)	✗	✓	N.A	Gaussian $\mathbb{P}$ Known rewards
LSVI-PHE [ICN+21]	$\sqrt{d^3H^4K}$ (Freq.)	✓	✓	✗	Generalized V approx Tabular, anti-concentration
OPT-RLSVI [ZBB+20]	$\sqrt{d^4H^5K}$ (Freq.)	✓	✓	✗	Linear V
EXP-UCRL [CGM21]	$\sqrt{d^2H^4K}$ (Freq.)	✗	✗	N.A	Bilinear Exp family known rewards
BEF-RLSVI This work	$\sqrt{d^3H^3K}$ (Freq.)	✓	✓	✓	Bilinear Exp family

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 bilinear exponential family of MDPs with known reward by a factor of $\sqrt{H}$ (Section 3.2). Our analysis (Section 5) builds on existing analyses of RLSVI-type algorithms [OVRW16], but contrary to them, we remove the need to handcraft a clipping of the value functions [ZBB+20]. 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 1. 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 [RZSD21], 2) Leveraging the observation that true value functions are bounded enables using an improved elliptical lemma (compared to [CGM21]), 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 [ZBB+20].

2 Bilinear exponential family of MDPs

In this section, we introduce the bilinear exponential family model coined in [CGM21] and extend it to parametric rewards. Then, we state a novel observation about linearity of this representation.

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

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

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

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 $\theta^{\mathrm{p}}, \theta^{\mathrm{r}} \in \mathbb{R}^d$. $M_{\theta^\cdot} \overset{\text{def}}{=} \sum_{i=1}^d \theta^\cdot_i A_i$, where A_i is a known $p \times q$ matrix for each i. Finally, Z denotes the log partition function:

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

Z^{r} is defined similarly. We denote $V^\pi_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},h}$, respectively $Q^\pi_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},h}$, the value, respectively 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 minimizes the (pseudo-)regret defined as:

$$\mathcal{R}(K) \triangleq \sum_{k=1}^K \left(V^{\pi^\star}_{\theta,1}(s^k_1) - V^{\pi^t}_{\theta,1}(s^k_1)\right). \tag{3}$$

Linearity of transitions. Now, we state an observation about the bilinear exponential family and discuss how it helps with the challenge of planning in episodic RL. Specifically, the popular assumption of linearity of the transition kernel is a direct consequence of our model. Indeed,

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

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

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

where $\mathcal{H}$ is the RKHS associated with the kernel, $\mu^{\mathrm{p}}(s') = (2\pi)^{-p/2}\, k\left(\psi(s'), .\right)\, \exp\left(\|\psi\left(s'\right)\|^2/2\right)$, and $\phi^{\mathrm{p}}(s,a) = k\left(M_{\theta^{\mathrm{p}}}^\top \varphi(s,a), .\right)\ \exp\left(\|M_{\theta^{\mathrm{p}}}\varphi(s,a)\|^2/2 - Z_{s,a}(\theta^{\mathrm{p}})\right)$. Equation (4) shows that s' is decoupled from (s,a), we see hereafter why this is crucial to reducing the complexity of planning.

Remark. *Up to our knowledge, [RZSD21] is the only work providing an example of linear transition kernel for RL with continuous state-action spaces. They consider Gaussian transitions with an unknown mean ($f^\star(s,a)$) and known variance (σ^2). Actually, linear $f^\star$ is a special case of the bilinear exponential family 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 can obtain $Q_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},h}$ by computing τ integrals per timestep, reducing the state-action space complexity to τ only. 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

In this section, we formally introduce the Bilinear Exponential Family Randomized Least-Squares Value Iteration (BEF-RLSVI) algorithm along with a high probability upper-bound on its regret.

3.1 BEF-RLSVI: algorithm design

BEF-RLSVI is based on RLSVI [OVRW16] framework with the distinction that we only perturb the reward parameters and not all the parameters of the value function. RLSVI algorithms are reminiscent of Thompson Sampling, yet more tractable with better control over the probability to be optimistic.

Algorithm 1 BEF-RLSVI

1: **Input:** failure rate δ, constants $\alpha^{\mathrm{p}}, \eta$ and $(x_k)_{k\in[K]} \in \mathbb{R}^+$
2: **for** episode $k = 1, 2, \ldots$ **do**
3: Observe initial state s_1^k
4: Sample noise $\xi_k \sim \mathcal{N}\left(0, x_k(\bar{G}_k^{\mathrm{p}})^{-1}\right)$ such that
$$\bar{G}_k^{\mathrm{p}} = \tfrac{\eta}{\alpha^{\mathrm{p}}}\mathbb{A} + \textstyle\sum_{\tau=1}^{k-1}\sum_{h=1}^{H} (\varphi(s_h^\tau, a_h^\tau)^\top A_i^\top A_j \varphi(s_h^\tau, a_h^\tau))_{i,j\in[d]}$$
5: Perturb reward parameter: $\tilde{\theta}^{\mathrm{r}}(k) = \hat{\theta}^{\mathrm{r}}(k) + \xi_k$
6: Compute $(Q_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}},h}^k)_{h\in[H]}$ via Bellman-backtracking, see Algorithm 2
7: **for** $h = 1, \ldots, H$ **do**
8: Pull action $a_h^k = \arg\max_a Q_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}},h}(s_h^k, a)$
9: Observe reward $r(s_h^k, a_h^k)$ and state s_{h+1}^k.
10: **end for**
11: Update the penalized ML estimators $\hat{\theta}^{\mathrm{p}}(k), \hat{\theta}^{\mathrm{r}}(k)$, see Equation (6) and Equation (8)
12: **end for**

We can see that Algorithm 1 performs exploration by a Gaussian perturbation of the reward parameter (Line 4). Contrary to optimistic approaches, this method is explicit and also 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 $\forall s, V_{H+1}(s) = 0$
2: **for** steps $h = H-1, H-2, \cdots, 0$ **do**
3: Calculate $Q_{\tilde{\theta},h}(s,a) = \mathbb{E}^{\tilde{\theta}^{\mathrm{r}}}_{s,a}[r] + \langle \phi^{\mathrm{p}}(s,a), \int V_{\tilde{\theta},h+1}(s')\mu^{\mathrm{p}}(s')ds' \rangle_{\mathcal{H}}$.
4: **end for**

We can approximate Line 3 of Algorithm 2 with $\mathcal{O}(pH^3K\log(HK))$ complexity and without harming the learning process (*cf.* § planning, Section 4). Therefore, here, planning is tractable.

3.2 `BEF-RLSVI`: regret upper-bound

We state the standard smoothness assumptions on the model [CGM21, JBNW17, LMT21].

Assumption 1. *There exist constants $\alpha^{\mathbf{p}}, \alpha^{\mathbf{r}}, \beta^{\mathbf{p}}, \beta^{\mathbf{r}} > 0$, such that the representation model satisfies:*

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \forall\theta, x \in \mathbb{R}^d \quad \alpha^{\mathbf{p}} \leq x^\top C^\theta_{s,a}[\psi] x \leq \beta^{\mathbf{p}}$$
$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \forall\theta, x \in \mathbb{R}^d \quad \alpha^{\mathbf{r}} \leq \mathbb{V}\mathrm{ar}^\theta_{s,a}(r)\, x^\top B^\top B x \leq \beta^{\mathbf{r}}$$

where $\mathbb{C}^\theta_{s,a}\left[\psi\left(s'\right)\right] \triangleq \mathbb{E}_{s'\sim\mathbb{P}_\theta|s,a}\left[\psi\left(s'\right)\psi\left(s'\right)^\top\right] - \mathbb{E}_{s'\sim\mathbb{P}_\theta|s,a}\left[\psi\left(s'\right)\right]\mathbb{E}_{s'\sim\mathbb{P}_\theta|s,a}\left[\psi\left(s'\right)^\top\right]$ *and* $\mathbb{V}\mathrm{ar}^\theta_{s,a}(r) \triangleq \left(\mathbb{E}^\theta_{s,a}\left[r^2\right] - \mathbb{E}^\theta_{s,a}\left[r\right]^2\right)$ *is the variance of the reward under θ.*

A closer look at the derivatives of the model (see Appendix D.3) tells us that previous inequalities directly imply a control over the eigenvalues of the Hessian matrices of the log-normalizers.

We now state our main result, the regret upper-bound of `BEF-RLSVI`.

Theorem 2 (Regret bound)**.** *Let $\mathbb{A} \triangleq (\mathrm{tr}(A_iA_j^\top))_{i,j\in[d]}$ 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 considering that*

1. $\max\{\|\theta^{\mathbf{r}}\|_{\mathbb{A}}, \|\theta^{\mathbf{p}}\|_{\mathbb{A}}\} \leq B_{\mathbb{A}}$, $\|\mathbb{A}^{-1}G_{s,a}\| \leq B_{\varphi,\mathbb{A}}$ and $\mathbb{E}_{\theta^{\mathbf{r}}}[r(s,a)] \in [0,1]$ for all (s,a).

2. noise $\xi_k \sim \mathcal{N}(0, x_k(\bar{G}^{\mathbf{p}}_k)^{-1})$ satisfies $x_k \geq \left(H\sqrt{\frac{\beta^{\mathbf{p}}\beta^{\mathbf{p}}(K,\delta)}{\alpha^{\mathbf{p}}\alpha^{\mathbf{r}}}} + \frac{\sqrt{\beta^{\mathbf{r}}\beta^{\mathbf{r}}(K,\delta)\min\{1,\frac{\alpha^{\mathbf{p}}}{\alpha^{\mathbf{r}}}\}}}{2\alpha^{\mathbf{r}}}\right)^2 \propto dH^2$,

then for all $\delta \in (0,1]$, with probability at least $1-7\delta$,

$$\mathcal{R}(K) \leq \sqrt{KH}\Bigg[\underbrace{2H\left(\sqrt{\frac{2\beta^{\mathbf{p}}}{\alpha^{\mathbf{p}}}\beta^{\mathbf{p}}(K,\delta)\gamma^{\mathbf{p}}_K} + (1+\sqrt{\gamma^{\mathbf{r}}_K})\sqrt{\log(1/\delta^2)}\right)}_{\textit{Transition concentration} \,\approx\, dH} + \underbrace{\beta^{\mathbf{r}}\sqrt{\frac{\beta^{\mathbf{r}}(n,\delta)\gamma^{\mathbf{r}}_K}{2\alpha^{\mathbf{r}}}}}_{\textit{Reward concentration} \,\approx\, d}$$
$$+ \underbrace{c\beta^{\mathbf{r}}\sqrt{x_K d\gamma^{\mathbf{r}}_K\log(dK/\delta)} + \frac{\beta^{\mathbf{r}}\sqrt{x_K d\gamma^{\mathbf{r}}_K\log(e/\delta^2)}}{\Phi(-1)}(1+\sqrt{\log(d/\delta)})}_{\textit{Noise concentration} \,\approx\, d^{3/2}H}\Bigg]$$
$$+ \sqrt{H\gamma^{\mathbf{r}}_K}\Bigg[\underbrace{\beta^{\mathbf{r}}C_d\left(\sqrt{\frac{\beta^{\mathbf{r}}(K,\delta)}{2\alpha^{\mathbf{r}}}} + c\sqrt{x_K d\log(dK/\delta)}\right)}_{\textit{Estimation error for no clipping} \,\approx\, dH}$$
$$+ \underbrace{\frac{\beta^{\mathbf{r}}d\sqrt{x_K}}{\Phi(-1)}(1+\sqrt{\log(d/\delta)})\sqrt{C_d\left(1+\frac{\alpha^{\mathbf{r}}B_{\varphi,A}H}{\eta}\right)}}_{\textit{Learning error for no clipping} \,\approx\, (dH)^{3/2}}\Bigg],$$

where for $\boldsymbol{i} \in [\boldsymbol{p}, \boldsymbol{r}]$, $\beta^{\boldsymbol{i}}(K,\delta) \triangleq \frac{\eta}{2}B^2_{\mathbb{A}} + \gamma^{\boldsymbol{i}}_K + \log(1/\delta)$, and $\gamma^{\boldsymbol{i}}_K \triangleq d\log(1+\frac{\beta^{\boldsymbol{i}}}{\eta}B_{\varphi,\mathbb{A}}HK)$. Also, $C_d \triangleq \frac{3d}{\log(2)}\log\left(1+\frac{\alpha^{\mathbf{r}}\|\mathbb{A}\|_2^2 B^2_{\varphi,\mathbb{A}}}{\eta\log(2)}\right)$, Φ is the Gaussian CDF, and c is a universal constant.

Theorem 2 entails a regret $\mathcal{R}(K) = \mathcal{O}(\sqrt{d^3H^3K})$ for `BEF-RLSVI`, where d is the number of parameters of the bilinear exponential family model, K is the number of episodes, and H is the horizon of an episode. We now clarify how this contrasts with related literature.

Comparison with Other Bounds. The closest work to ours is [CGM21] as it considers the same model for transitions but with known rewards. They propose a `UCRL`-type and `PSRL`-type algorithm, which achieve a regret of order $\widetilde{O}(\sqrt{d^2 H^4 K})$. There are two notable algorithmic differences with our work. First, they do exploration using intractable-optimistic upper bounds or high-dimensional posteriors, while we do it with explicit perturbation. 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 a $\sqrt{H}$ factor thanks to an improved analysis, (*cf.* Lemma 18). But similar to all `RLSVI`-type algorithms, we pick up an extra $\sqrt{d}$ (*cf.* [AL17]).

[ZBB+20] 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 $R(K) = \widetilde{O}(\sqrt{d^4 H^5 K})$, 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 13) instead of the simulation lemma, which saves us a $\sqrt{H}$ factor.

[RZSD21] 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 [RR07] unlike us (Section 4). Finally, they show a Bayesian regret bound of $O(\sqrt{d^2 H^3 K})$. 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, [DMKV21] proves a lower bound of order $\Omega(\sqrt{H^3 SAK})$. If we represent a tabular MDP in our model, we would need $d = S^2 \times A$ parameters (Section 4.3, [CGM21]). In this case, our bound becomes $R(K) = O(\sqrt{(S^2 A)^3 H^3 K})$, 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` [CGM21], *i.e.* by using a penalized maximum likelihood estimator

$$\hat{\theta}^{\mathtt{p}}(k) \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{t=1}^{k} \sum_{h=1}^{H} - \log \mathbb{P}_\theta\left(s_{h+1}^t \mid s_h^t, a_h^t\right) + \eta \operatorname{pen}(\theta).$$

Here, pen(θ) is a trace-norm penalty: $\operatorname{pen}(\theta) = \frac{1}{2}\|\theta\|_{\mathbb{A}}$ and $\mathbb{A} = (\operatorname{tr}(A_i A_j^\top))_{i,j}$. By properties of the exponential family, the penalized maximum likelihood estimator verifies, for all $i \le d$:

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

Equation (6) can be solved in closed form for simple distributions, like Gaussian, but it can involve integral approximations for other distribution. We estimate the parameter for reward, *i.e.* θ_r, similarly

$$\hat{\theta}^{\mathtt{r}}(k) \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{t=1}^{k} \sum_{h=1}^{H} - \log \mathbb{P}_\theta\left(r_t \mid s_h^t, a_h^t\right) + \eta \operatorname{pen}(\theta), \tag{7}$$

$$\Longrightarrow \qquad \sum_{t=1}^{k} \sum_{h=1}^{H} \left(r_t - \mathbb{E}_{s_h^t, a_h^t}^{\hat{\theta}_k^{\mathtt{r}}}\left[r\right]\right) B^\top A_i \varphi\left(s_h^t, a_h^t\right) = \eta \nabla_i \operatorname{pen}\left(\hat{\theta}_k^{\mathtt{r}}\right) \quad \forall i \in [d]. \tag{8}$$

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 [ZBB+20, OVRW16]. 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 $s_1^k \in \mathcal{S}$,

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

[ZBB$^+$20] proves that this suffices to bound the learning error. However, their method clashes with not clipping the value function, as it modifies the probability of optimism. Thus, [ZBB$^+$20] proposes an involved clipping procedure to handle the issue of unstable values. Instead, by careful geometric analysis (*cf.* Lemma 19), we bound the occurrences of the unstable values, and in turn, upper bound the regret without clipping. Note that unlike [ICN$^+$21], `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:

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

Then, we select the best action greedily using dynamic programming to compute $Q_h(s,a)$. Although our model yields infinite dimensional ϕ^{p} and ψ^{p}, approximating them (*cf.* next paragraph) with linear features of dimension $\mathcal{O}(pH^2K\log(HK))$ is possible without 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 D.5. 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 [RR07], 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

$$R(K) = \sum_{k=1}^K (V^\star_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},1} - V^{\pi_k}_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},1})(s_{1k}) = \sum_{k=1}^K (\underbrace{V^\star_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},1} - V^{\pi_k}_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}},1}}_{learning} + \underbrace{V^{\pi_k}_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}},1} - V^{\pi_k}_{\theta^{\mathrm{p}},\theta^{\mathrm{r}},1}}_{Estimation})(s_{1k}). \quad (9)$$

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 18 leads to the upper-bound.

5.1 Bounding the estimation error

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

$$V^\pi_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}}}(s_{1k}) - V^\pi_{\theta^{\mathrm{p}},\theta^{\mathrm{r}}}(s_{1k}) = \underbrace{V^\pi_{\hat{\theta}^{\mathrm{p}},\theta^{\mathrm{r}}}(s_{1k}) - V^\pi_{\theta^{\mathrm{p}},\theta^{\mathrm{r}}}(s_{1k})}_{\text{transition estimation}} + \underbrace{V^\pi_{\hat{\theta}^{\mathrm{p}},\tilde{\theta}^{\mathrm{r}}}(s_{1k}) - V^\pi_{\hat{\theta}^{\mathrm{p}},\theta^{\mathrm{r}}}(s_{1k})}_{\text{reward estimation}} \quad (10)$$

Transition estimation Since the reward parameter is exact, the value function's span is $\leq H$. Then, using the transportation of Lemma 13 we obtain the bound $H\sum_{h=1}^H \sqrt{2\,\mathrm{KL}_{s_{hk},a_{hk}}(\theta^{\mathrm{p}},\hat{\theta}^{\mathrm{p}})}$. We notice that since the reward parameter is exact, the bound is actually $H\min\{1, \sum_{h=1}^H \sqrt{2\,\mathrm{KL}_{s_{hk},a_{hk}}(\theta^{\mathrm{p}},\hat{\theta}^{\mathrm{p}})}\}$. Using Lemma 18 under Assumption 1, we win a $\sqrt{H}$ factor compared to the analysis of [CG19].

Reward estimation Previous work uses clipping to help control this error, but in this case it can hinder the optimism probability by biasing the noise. [ZBB$^+$20] proposes an involved clipping depending on the norms $\|(A_i\varphi(s_h^k,a_h^k))_{i\in[d]}\|_{(\bar{G}_k^{\mathrm{p}})^{-1}}$, which is somewhat delicate to analyze and deploy. We remedy the situation acting solely in the proof. First let's define what we call the set of "bad rounds": $\left\{k \in [K], \exists h : \|(A_i\varphi(s_h^k, a_h^k))_{i\in[d]}\|_{(\bar{G}_k^{\mathtt{p}})^{-1}} \geq 1\right\}$, these rounds are why clipping is necessary. Thanks to Lemma 19, we know that the number of such rounds is at most $\mathcal{O}(d)$. 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. Therefore, our algorithm can forgo clipping for free.

Remark. *If it wasn't for the episodic nature of our setting, we could have used the forward algorithm to eliminate the span control issue. We refer to [Vov01, AW01] for a description of this algorithm, [OMP21] for a stochastic analysis, and Section 4 therein for an application to linear bandits.*

5.2 Bounding the learning error

To upper-bound this term of the regret, we first show that the estimated value function is optimistic with a constant probability. Then, we show that this is enough to control the learning error.

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

$$\begin{aligned}(V_{\hat{\theta}^{\mathtt{p}},\tilde{\theta}^{\mathtt{r}},1} - V^\star_{\theta^{\mathtt{p}},\theta^{\mathtt{r}},1})(s_1) &\geq (Q^\star_{\hat{\theta}^{\mathtt{p}},\tilde{\theta}^{\mathtt{r}},1} - Q^\star_1)(s_1, \pi^\star(s_1)) \\ &\geq \underbrace{V^{\pi^\star}_{\hat{\theta}^{\mathtt{p}},\theta^{\mathtt{r}}}(s_1) - V^{\pi^\star}_{\theta^{\mathtt{p}},\theta^{\mathtt{r}}}(s_1)}_{\text{first term}} + \underbrace{V^{\pi^\star}_{\hat{\theta}^{\mathtt{p}},\hat{\theta}^{\mathtt{r}}}(s_1) - V^{\pi^\star}_{\hat{\theta}^{\mathtt{p}},\theta^{\mathtt{r}}}(s_1)}_{\text{second term}} + \underbrace{V^{\pi^\star}_{\hat{\theta}^{\mathtt{p}},\tilde{\theta}^{\mathtt{r}}}(s_1) - V^{\pi^\star}_{\hat{\theta}^{\mathtt{p}},\hat{\theta}^{\mathtt{r}}}(s_1)}_{\text{third term}}\end{aligned}$$

The first and second terms are perturbation free, we handle them similarly to the estimation error, *i.e.* using concentration arguments for $\hat{\theta}^{\mathtt{p}}$ and $\hat{\theta}^{\mathtt{r}}$. For the third term, we use transportation of rewards (Lemma 17) and anti-concentration of ξ_k (Lemma 12). We find that with probability at least $1 - 2\delta$

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

where $c(n,\delta) = \left(\sqrt{\beta^{\mathtt{p}}\beta^{\mathtt{p}}(n,\delta)/\alpha^{\mathtt{p}}} + \sqrt{\beta^{\mathtt{r}}\beta^{\mathtt{r}}(n,\delta)\min\{1, \alpha^{\mathtt{p}}/\alpha^{\mathtt{r}}\}/(2\alpha^{\mathtt{r}})}\right)$. Since $\xi_k \sim \mathcal{N}(0, x_k(\bar{G}_k^{\mathtt{p}})^{-1})$ and $x_k \geq H^2 c(n,\delta)^2$, we get $\mathbb{P}\left(V^\pi_{\hat{\theta}^{\mathtt{p}},\tilde{\theta}^{\mathtt{r}},1}(s_1) - V^\star_{\theta^{\mathtt{p}},\theta^{\mathtt{r}},1}(s_1) \geq 0\right) \geq \Phi(-1)$, where Φ is the normal CDF. This is ensured by the anti-concentration property of Gaussian random variables, see Lemma 12.

From stochastic optimism to error control: Existing algorithms require the value function to be optimistic (*i.e.* negative learning error) with large probability. Contrary to them, `BEF-RLSVI` only requires the estimated value to be optimistic with a constant probability. When it is, the learning happens. Otherwise, the policy is still close to a good one thanks to the decreasing estimation error, and the learning still happens. This part of the proof is similar in spirit to that of [ZBB+20].

Upper bound on $V_1^\star$: Draw $(\bar{\xi}_k)_{k\in[K]}$ i.i.d copies of $(\xi_k)_{k\in[K]}$ and define the event where optimism holds as $\bar{O}_k \triangleq \{V_{\hat{\theta}^{\mathtt{p}},\tilde{\theta}_k^{\mathtt{r}},1}(s_1^k) - V_1^\star(s_1^k) \geq 0\}$. This implies that $V_1^\star(s_1^k) \leq \mathbb{E}_{\bar{\xi}_k|\bar{O}_k}[V_{\hat{\theta}^{\mathtt{p}},\hat{\theta}^{\mathtt{r}}+\bar{\xi}_k,1}(s_1^k)]$.

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

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

As the injected noise concentrates, we obtain $\underline{\mathrm{V}}_1(s_1^k) \leq V_{\hat{\theta}^{\mathtt{p}},\tilde{\theta}^{\mathtt{r}}}(s_1^k)$.

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

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

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 B.2 for the detailed proof.

6 Related works: functional representations with regret and tractability

Our work extends the endeavor of using functional representations to perform optimal regret minimization in continuous state-action MDPs. We now provide a few complementary details.

General functional representation. [DSL$^+$18] 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. [WSY20] proposes a framework that leverages the optimism with confidence bound approach for general functional representations with bounded Eluder dimensions, which is a complexity measure in RL. However, knowing the Eluder dimension is crucial for the optimistic confidence bound in their algorithm. Eluder dimension is not known for MDPs except linear and tabular MDPs. *To concretize our design, we focus on the general but explicit bilinear exponential family of MDPs than any abstract representation.*

Bilinear exponential family of MDPs. Exponential families are studied widely in RL theory, from bandits to MDPs [LMT21, KKM13, FCGS10, KH06], as an expressive parametric family to design theoretically-grounded model-based algorithms. [CGM21] 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, linear MDPs, and linearly controlled dynamical systems, the BEF-family of MDPs has received increasing attention [LLS$^+$21]. [LLS$^+$21] estimates the model parameters based on score matching that enables them to replace regularity assumption on the log-partition function with Fisher-information and assumption on the parameters. Both [CGM21, LLS$^+$21] achieve a worst-case regret of order $\tilde{O}(\sqrt{d^2H^4K})$ for known reward. On a different note, [DKL$^+$21, FKQR21] 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.*

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, [DKWY19] 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 [SS20, LSW20, VRD19] propose to leverage different low-dimensional representations of value functions or transitions to perform efficient planning. Here, we take note from [RZSD21] 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 [RR07] 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, [RZSD21, NY21] use representations given by random Fourier features [RR07] to approximate the transition dynamics and provide experiments validating the benefits of this approach for high-dimensional Atari-games.

7 Conclusion and future work

We propose the `BEF-RLSVI` algorithm for the bilinear exponential family of MDPs in the setting of episodic-RL. `BEF-RLSVI` explores using a Gaussian perturbation of rewards, and plans tractably (complexity of $\mathcal{O}(pH^3K\log(HK))$) thanks to properties of the RBF kernel. Our proof shows that clipping can be forwent for similar `RLSVI`-type algorithms. Moreover, we prove a $\sqrt{d^3H^3K}$ frequentist regret bound, which improves over existing work, accommodates unknown rewards, and matches the lower bound in terms of H and K. Regarding future work, we believe that our proof approach can be extended to rewards with bounded variance. We also believe that the extra $\sqrt{d}$ in our bound is an artefact of the proof, and specifically, the anti-concentration. We will investigate it further. Finally, we plan 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

[AL17] Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017. (Cited on 6, 21, 27, 28)

[AOM17] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017. (Cited on 1, 2)

[AW01] Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001. (Cited on 8)

[AYS11] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26. JMLR Workshop and Conference Proceedings, 2011. (Cited on 2)

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013. (Cited on 29)

[CG19] 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. PMLR, 2019. (Cited on 7, 23)

[CGM21] Sayak Ray Chowdhury, Aditya Gopalan, and Odalric-Ambrym Maillard. Reinforcement learning in parametric mdps with exponential families. In *International Conference on Artificial Intelligence and Statistics*, pages 1855–1863. PMLR, 2021. (Cited on 2, 3, 5, 6, 9, 23)

[DKL+21] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021. (Cited on 2, 9)

[DKWY19] Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019. (Cited on 2, 9)

[DLB17] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017. (Cited on 1)

[DMKV21] Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021. (Cited on 6)

[DSL+18] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134. PMLR, 2018. (Cited on 9)

[FCGS10] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010. (Cited on 9)

[FKQR21] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021. (Cited on 2, 9)

[ICN+21] Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin F Yang. Randomized exploration for reinforcement learning with general value function approximation. *arXiv preprint arXiv:2106.07841*, 2021. (Cited on 3, 7)

[JBNW17] Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. *arXiv preprint arXiv:1706.00136*, 2017. (Cited on 5)

[JYWJ20] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020. (Cited on 2)

[KH06] Branislav Kveton and Milos Hauskrecht. Solving factored mdps with exponential-family transition models. In *ICAPS*, pages 114–120, 2006. (Cited on 9)

[KK99] Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pages 740–747, 1999. (Cited on 2)

[KKM13] Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. *Advances in neural information processing systems*, 26, 2013. (Cited on 9)

[LLS+21] Gene Li, Junbo Li, Nathan Srebro, Zhaoran Wang, and Zhuoran Yang. Exponential family model-based reinforcement learning via score matching. *arXiv preprint arXiv:2112.14195*, 2021. (Cited on 2, 9)

[LMT21] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Low-rank generalized linear bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pages 460–468. PMLR, 2021. (Cited on 5, 9)

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. (Cited on 33)

[LSW20] Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020. (Cited on 2, 9)

[NY21] Ofir Nachum and Mengjiao Yang. Provable representation learning for imitation with contrastive fourier features. *Advances in Neural Information Processing Systems*, 34, 2021. (Cited on 9)

[OMP21] Reda Ouhamma, Odalric-Ambrym Maillard, and Vianney Perchet. Stochastic online linear regression: the forward algorithm to replace ridge. *Advances in Neural Information Processing Systems*, 34:24430–24441, 2021. (Cited on 8)

[ORVR13] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013. (Cited on 1, 2)

[OVR14] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014. (Cited on 2)

[OVRW16] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016. (Cited on 2, 3, 4, 6)

[RR07] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. (Cited on 6, 7, 9)

[RZSD21] Tongzheng Ren, Tianjun Zhang, Csaba Szepesvári, and Bo Dai. A free lunch from the noise: Provable and practical exploration for representation learning. *arXiv preprint arXiv:2111.11485*, 2021. (Cited on 3, 4, 6, 9, 16)

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. (Cited on 1)

[SS20] Roshan Shariff and Csaba Szepesvári. Efficient planning in large mdps with weak linear function approximation. *arXiv preprint arXiv:2007.06184*, 2020. (Cited on 2, 9)

[Vov01] Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001. (Cited on 8)

[VRD19] Benjamin Van Roy and Shi Dong. Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019. (Cited on 2, 9)

[WSY20] Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *arXiv preprint arXiv:2005.10804*, 2020. (Cited on 9)

[ZBB+20] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020. (Cited on 3, 6, 7, 8)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a purely theoretical contribution
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See the appendices
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite creator of the bilinear exponential family model.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]