Learning Near Optimal Policies with Low Inherent Bellman Error

Andrea Zanette [1] Alessandro Lazaric [2] Mykel Kochenderfer [1] Emma Brunskill [1]

Abstract

We study the exploration problem with approximate linear action-value functions in episodic reinforcement learning under the notion of low inherent Bellman error, a condition normally employed to show convergence of approximate value iteration. First we relate this condition to other common frameworks and show that it is strictly more general than the low rank (or linear) MDP assumption of prior work. Second we provide an algorithm with a high probability regret bound $\widetilde{O}(\sum_{t=1}^{H} d_t\sqrt{K} + \sum_{t=1}^{H} \sqrt{d_t}\mathcal{I}K)$ where H is the horizon, K is the number of episodes, $\mathcal{I}$ is the value if the inherent Bellman error and d_t is the feature dimension at timestep t. In addition, we show that the result is unimprovable beyond constants and logs by showing a matching lower bound. This has two important consequences: 1) it shows that exploration is possible using only *batch assumptions* with an algorithm that achieves the optimal statistical rate for the setting we consider, which is more general than prior work on low-rank MDPs 2) the lack of closedness (measured by the inherent Bellman error) is only amplified by $\sqrt{d_t}$ despite working in the online setting. Finally, the algorithm reduces to the celebrated LinUCB when $H = 1$ but with a different choice of the exploration parameter that allows handling misspecified contextual linear bandits. While computational tractability questions remain open for the MDP setting, this enriches the class of MDPs with a linear representation for the action-value function where statistically efficient reinforcement learning is possible.

1. Introduction

Improving the sample efficiency of reinforcement learning (RL) algorithms through effective exploration-exploitation strategies is a major focus of the recent theoretical literature. Strong results are available with a generative model (Azar et al., 2012; Sidford et al., 2018; Agarwal et al., 2019; Zanette et al., 2019a) as well as in the *online* setting when the learning performance is measured by the cumulative regret, i.e., the difference between the performance of the optimal policy and the reward accumulated by the learner. For finite horizon problems, UCBVI (Azar et al., 2017) achieves worst-case optimal regret, while algorithms with domain adaptive bounds have been introduced by (Zanette & Brunskill, 2019) and (Simchowitz & Jamieson, 2019). Randomized (Russo, 2019) and model-free (Jin et al., 2018) variants have also been proposed, together with methods with other beneficial properties (Dann et al., 2019; Efroni et al., 2019). Similar results are also available in the infinite horizon setting (Jaksch et al., 2010; Maillard et al., 2014; Fruit et al., 2018; Zhang & Ji, 2019; Tossou et al., 2019).

Approximate dynamic programming. While the results for tabular settings are encouraging, function approximation is normally required to tackle problems where the state or action spaces may be intractably large. In this case, even when the Bellman operator can be applied exactly, simple dynamic programming algorithms coupled with linear architectures may diverge (Baird, 1995; Tsitsiklis & Van Roy, 1996), thus suggesting that effective approximate RL may not be feasible in the general case.

Convergence guarantees (Lagoudakis & Parr, 2003) and finite-sample analyses (Lazaric et al., 2012) are available for the least-squares policy improvement (LSPI) algorithm under the assumption that the value function of *all policies can be well approximated* within the chosen function class (*LSPI conditions*, for short). For concreteness, let ϵ be the worst-case misspecification error of a d-dimensional linear function approximator over the policy action-value functions (i.e., for any policy π, there exists an approximation $\widehat{Q}^{\pi}$ such that $\|\widehat{Q}^{\pi} - Q^{\pi}\| \leqslant \epsilon$). Recently, (Du et al., 2019) showed that when using highly misspecified approximators $\epsilon \gtrapprox 1/\sqrt{d}$ the worst-case sample complexity may be exponential in d. At the same time, when $\epsilon \lessapprox 1/\sqrt{d}$, (Van Roy & Dong, 2019) and (Lattimore & Szepesvari,

[1] Stanford University [2] Facebook Artificial Intelligence Research. Correspondence to: Andrea Zanette <zanette@stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

2020) showed algorithms with $\sqrt{d}$ loss times the misspecification level ϵ. In particular, (Lattimore & Szepesvari, 2020) showed that LSPI attains polynomial sample complexity using G-optimal design with a $\approx \sqrt{d}\epsilon$ additive error using a *generative model*.

Similarly, for the least-squares value iteration algorithm (LSVI) convergence guarantees (Munos, 2005) and finite sample analysis (Munos & Szepesvári, 2008) are also available under the assumption of *low inherent Bellman error (IBE)*, (*LSVI conditions*, for short). Given a function class $\mathcal{F}$, the IBE measures the error in approximating the image of any function in $\mathcal{F}$ through the Bellman operator. Whenever the IBE is not small, it is easy to show that approximation errors may be amplified by a constant factor at each application of the Bellman operator, leading to divergence. Although methods exist to limit this amplification of errors (Zanette et al., 2019b; Kolter, 2011), the question of when sample-efficient value-based RL is possible remains open even in the absence of misspecification.

In this paper we focus on the problem of exploration-exploitation using LSVI approaches in settings with low IBE. We make several contributions.

Exploration with low inherent Bellman error. We first show that the notion of inherent Bellman error is distinct from the LSPI condition, and more general than the low-rank assumption on the dynamics used in a series of recent works on exploration with linear function approximation (Yang & Wang, 2020; Jin et al., 2020; Zanette et al., 2020). For a finite horizon MDP, when the LSVI conditions are satisfied either exactly or approximately (i.e., the inherent Bellman error is either zero or small) we propose *Efficient Linear Exploration of Actions by Nonlinear Optimization of the Residuals* (ELEANOR), an optimistic generalization of the popular LSVI algorithm. We analyze ELEANOR and derive the first regret bound for this setting and show it is unimprovable in terms of statistical rates, though we leave its computational tractability open.

Our analysis shows that the performance of ELEANOR degrades gracefully in the case of positive inherent Bellman error. Interestingly, we recover a similar $\sqrt{d}$ amplification of the misspecification error (the IBE in our case) as for LSPI (Lattimore & Szepesvari, 2020) , despite the fact that we consider the more challenging online setting as opposed to the generative model by Lattimore & Szepesvari (2020).

Low-rank MDPs and contextual misspecified linear bandits. Our result applies to low-rank MDPs and improves upon the best-known regret bound for that setting (Jin et al., 2020) by a $\sqrt{d}$ factor. When applied to contextual linear bandits, our algorithm reduces to the celebrated LINUCB (or OFUL) algorithm of (Abbasi-Yadkori et al., 2011). In addition, however, it *can handle contextual misspecified linear bandits while retaining computationally tractability*, making this the first algorithm and analysis for this setting, although we require knowledge of the misspecification level. A similar result was recently derived for a different algorithm based on G-experimental design (Lattimore & Szepesvari, 2020) for the more restrictive setting of non-contextual (i.e., with features not depending on the state and fixed action space) misspecified linear bandits; however, their approach is agnostic to the misspecification level.

Core ideas. LSVI-based algorithms have been successfully analyzed for low-rank MDPs (Jin et al., 2020) by adding exploration bonuses at every experienced state, thereby ensuring optimism by backward induction. In contrast, our more general setting demands that the value function stays linear, ruling out approaches based on exploration bonuses. In fact, if the value function used for backup is not linear, low inherent Bellman error does not provide any guarantee about how errors may propagate, which can be exponential in the general case (Zanette et al., 2019b).

Our proposal extends the LSVI algorithm to return an optimistic solution at the initial state through *global* optimization over the value function parameters, while still enforcing linearity of the representation. This has two advantages: 1) (*handling of the bias*) it enables us to use the concept of inherent Bellman error, requiring that the Bellman operator be applied to *linear* action-value functions and avoiding a $\sqrt{d}$ amplification of the value function error at every step (Zanette et al., 2019b); 2) (*handling of the variance*) it keeps the complexity of the action-value functional space small (linear), enabling the use of confidence intervals that are as tight as those used in the bandit literature, yielding the optimal finite-sample statistical rate.

2. Notation

We consider an undiscounted finite-horizon MDP (Puterman, 1994) $M = (\mathcal{S}, \mathcal{A}, p, r, H)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, and horizon length $H \in \mathbb{N}^+$. For every $t \in [H] \overset{def}{=} \{1, \ldots, H\}$, every state-action pair is characterized by an expected reward $r_t(s,a)$ with an associated reward random variable $R_t(s,a)$ and a transition kernel $p_t(\cdot \mid s,a)$ over next state. We assume $\mathcal{S}$ to be a measurable, possibly infinite, space and $\mathcal{A}$ can be any (compact) time and state dependent set (we omit this dependency for brevity). For any $t \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, the state-action value function of a non-stationary policy $\pi = (\pi_1, \ldots, \pi_H)$ is defined as $Q_t^\pi(s,a) = r_t(s,a) + \mathbb{E}\left[\sum_{l=t+1}^{H} r_l(s_l, \pi_l(s_l)) \mid s, a\right]$ and the value function is $V_t^\pi(s) = Q_t^\pi(s, \pi_t(s))$. Since the horizon is finite, under some regularity conditions, e.g., (Shreve & Bertsekas, 1978), there always exists an optimal policy $\pi^\star$ whose value and action-value functions are defined as $V_t^\star(s) \overset{def}{=} V_t^{\pi^\star}(s) = \sup_\pi V_t^\pi(s)$ and

$Q_t^\star(s,a) \overset{def}{=} Q_t^{\pi^\star}(s,a) = \sup_\pi Q_t^\pi(s,a)$.

The value iteration (or backward induction) algorithm (Sutton & Barto, 2018) computes $\pi^\star$ and $V^\star$ as follows: it starts from $V_{H+1}^\star(s) = 0$ for all $s \in \mathcal{S}$ and it computes $Q_t^\star$ using the Bellman equation in each state-action pair recursively from $t = H$ down to 1 and it returns the optimal policy $\pi_t^\star(s) = \arg\max_a Q_t^\star(s,a)$. In particular, the Bellman operator $\mathcal{T}_t$ applied to Q_{t+1} is defined as

$$\mathcal{T}_t(Q_{t+1})(s,a) = r_t(s,a) + \mathbb{E}_{s'\sim p_t(s,a)} \max_{a'} Q_{t+1}(s',a').$$

3. Linear Value Function Frameworks

In this section we introduce basic notation and assumptions for linear function approximation, we define the concept of inherent Bellman error, and we investigate connections with alternative settings.

Whenever the state space $\mathcal{S}$ is too large or continuous, value functions cannot be represented by enumerating their values at each state or state-action pair. A common approach is to define a feature map $\phi_t : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{d_t}$, possibly different at any $t \in [H]$, embedding each state-action pair (s,a) into a d_t-dimensional vector $\phi_t(s,a)$. The action-value functions are then represented as a linear combination between the features ϕ_t and a vector parameter $\theta_t \in \mathbb{R}^{d_t}$, such that $Q_t(s,a) = \phi_t(s,a)^\top\theta_t$. This effectively reduces the complexity of the problem from $|\mathcal{S}\times\mathcal{A}|$ down to d_t.

We define the space of parameters θ inducing uniformly bounded action-value functions

$$\mathcal{B}_t \overset{def}{=} \{\theta_t \in \mathbb{R}^{d_t} \mid |\phi_t(s,a)^\top\theta_t| \leqslant D, \forall(s,a)\}. \tag{1}$$

We will later require the constant $D \in \mathbb{R}$ to be chosen to satisfy Asm. 1. For instance, $D = 1$ requires the value function to be in $[-1,+1]$ and complies with the assumption.

Each parameter θ identifies an (action) value function

$$Q_t(\theta_t)(s,a) = \phi_t(s,a)^\top\theta_t, \quad V_t(\theta_t) = \max_a \phi_t(s,a)^\top\theta_t$$

and the associated functional spaces

$$\mathcal{Q}_t \overset{def}{=} \{Q_t(\theta_t) \mid \theta_t \in \mathcal{B}_t\},\ \mathcal{V}_t \overset{def}{=} \{V_t(\theta_t) \mid \theta_t \in \mathcal{B}_t\}. \tag{2}$$

Inherent Bellman error. The value iteration algorithm can be used to compute an optimal policy (Sutton & Barto, 2018) and it smoothly extends to linear approximators. The procedure repeatedly applies the Bellman operator $\mathcal{T}_t$ to an action-value function[1] $Q_t \in \mathcal{Q}_t$ and projects the computed point $\mathcal{T}_tQ_t$ back to $\mathcal{Q}_{t+1}$ using a (e.g., least-squares) projection operator Π_t. The projection error is precisely the inherent Bellman error, which can be thought of as how *close* the space $\mathcal{Q}_t$ is w.r.t. the Bellman operator $\mathcal{T}_t$.

Definition 1. *The inherent Bellman error*[2] *of an MDP with a linear feature representation ϕ is denoted with $\mathcal{I}$ and is the maximum over the timesteps $t \in [H]$ of*

$$\sup_{\theta_{t+1}\in\mathcal{B}_{t+1}} \inf_{\theta_t\in\mathcal{B}_t} \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\phi_t(s,a)^\top\theta_t - (\mathcal{T}_tQ_{t+1}(\theta_{t+1}))(s,a)|.$$

Our definition of inherent Bellman error is *natural* in the sense that it is defined with respect to the linear action-value function class without additional clipping if the value function exceeds a prescribed threshold and is not enlarged to incorporate exploration bonuses (see e.g., (Wang et al., 2019)). Alternative definitions may enlarge the underlying functional space in an artificial, non linear, possibly algorithm-dependent way, and result in a much more restrictive definition of inherent Bellman error. We notice that while our definition is less restrictive, it rules out traditional forms of exploration based on *adding exploration bonuses*, making it harder to design effective exploration strategies.

Properties. We discuss the properties of MDPs with $\mathcal{I} = 0$. An immediate consequence of def. 1 is that when $\mathcal{I} = 0$ the reward function is linear, and so is the transition kernel *when applied to elements of $\mathcal{V}_{t+1}$*.

Proposition 2 (Linearity of Rewards and Restricted Linearity of Transitions)**.** *Given an MDP and a linear feature representation with $\mathcal{B}_t = \mathbb{R}^{d_t}$ and inherent Bellman error $\mathcal{I} = 0$ we have that the rewards are linear in the sense that:*

$$\inf_{\theta_t^R\in\mathcal{B}_t} \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |r_t(s,a) - \phi_t(s,a)^\top\theta_t^R| = 0$$

and the transition have a linear effect on members of $\mathcal{V}_{t+1}$

$$\sup_{\theta_{t+1}\in\mathcal{B}_{t+1}} \inf_{\theta_t^P\in\mathcal{B}_t} \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\,\mathbb{E}_{s'\sim p_t(s,a)} V_{t+1}(\theta_{t+1})(s') - \phi_t(s,a)^\top\theta_t^P| = 0.$$

If $\mathcal{I} = 0$, the application of the Bellman operator $\mathcal{T}_t$ to members of $\mathcal{Q}_{t+1}$ always produces a member of $\mathcal{Q}_t$, i.e., $\mathcal{T}_t\mathcal{Q}_{t+1} \subseteq \mathcal{Q}_t$. From here, we can immediately see that the zero inherent Bellman error assumption is more general than low-rank MDPs (Yang & Wang, 2020; Jin et al., 2020; Zanette et al., 2020). Indeed, in low-rank MDPs the Bellman operator returns a function in the range of the features (i.e., in $\mathcal{Q}_t$) *regardless of value function* Q_{t+1}, while problems with zero inherent Bellman error are only required to map elements of $\mathcal{Q}_{t+1}$ to $\mathcal{Q}_t$, and are thus more general approximators.

Proposition 3 (Low Rank $\subseteq$ LSVI Conditions)**.** *Let $\mathcal{B}_t = \mathbb{R}^{d_t}$, and consider an MDP with associated linear feature*

[1] One can reason with either the value function V or the action-value function Q.

[2] A different definition, more suitable for generative models with stationary policies using a p-norm induced by the sampling distribution is provided by (Munos & Szepesvári, 2008).

representation ϕ. If the MDP is a low rank (or linear) MDP, i.e., for a parameter $\theta_t^R \in \mathbb{R}^{d_t}$ and a measure function[3] *$\psi_t(\cdot)$:*

$$\begin{aligned} \forall (s,a,t,s'), \quad r_t(s,a) &= \phi_t(s,a)^\top \theta_t^R \\ p_t(s' \mid s,a) &= \phi_t(s,a)^\top \psi_t(s') \end{aligned} \tag{16}$$

then $\mathcal{I} = 0$. However, the converse does not hold, i.e., there exists an MDP and a linear feature extractor ϕ with $\mathcal{I} = 0$ which is not a linear MDP in the sense of eq. (16).

Another assumption often made on the approximation space is that the action-value functions for *all policies* do belong to $\mathcal{Q}_t$ (LSPI condition), a condition normally employed to show convergence of LSPI (Lagoudakis & Parr, 2003). This assumption is also strictly less restrictive than low-rank (see also (Jin et al., 2020) for a claim in one direction).

Proposition 4 (Low Rank $\subseteq$ LSPI Conditions)**.** *If a given MDP is low rank in the sense of eq. (16) then the value function of all policies admit a linear parameterization:*

$$\forall \pi, \; \forall t \in [H], \; \exists \theta_t^\pi \quad \textit{such that} \quad Q_t^\pi(s,a) = \phi_t(s,a)^\top \theta_t^\pi.$$

However, there exists an MPD and a linear approximator with feature extractor ϕ which satisfies the above display but there exists no ψ_t such that eq. (16) holds.

One may wonder what is the relation between MDPs with no inherent Bellman error and MDPs where all action-value function for all policies are linear, i.e., the LSVI and LSPI conditions. These are two very distinct assumptions: the former deals with policies *that are optimal with respect to a parameter*, while the latter deals with arbitrary policies. Conversely, the latter deals with the Q values that actually corresponds to Q values of policies, while the former measures the error with respect to any function in the class.

Proposition 5 (LSVI Conditions $\neq$ LSPI Conditions)**.** *There exists an MDP and a linear representation with feature extractor ϕ with $\mathcal{I} = 0$ and yet the policies are not linearly parameterizable in the sense that:*

$$\exists \pi, \exists t \in [H], \nexists \theta_t^\pi \in \mathbb{R}^{d_t} \quad \textit{s.t.} \quad Q_t^\pi = \phi_t(s,a)^\top \theta_t.$$

Vice-versa, there exists an MDP and a feature representation such that all action-value functions of all policies admit a linear parameterization:

$$\forall \pi, \forall t \in [H], \; \exists \theta_t^\pi \quad \textit{that satisfies} \quad Q_t^\pi(s,a) = \phi_t(s,a)^\top \theta_t^\pi$$

and yet the inherent Bellman is non-zero: $\mathcal{I} > 0$.

The final connection we make is with settings with *low Bellman rank*, see (Jiang et al., 2017). It is possible to show that if the LSVI conditions are satisfied, the Bellman rank is at most d, where d is the dimensionality of the features. However, no statistically efficient algorithm exists for this setting, because OLIVE from (Jiang et al., 2017) has an explicit dependence on the size of the action space, which can be very large or infinite in the setting we consider here.

4. Algorithm

We consider the standard online learning protocol in finite-horizon problems, where at each episode k, the learner executes a policy π_k, records the samples in the trajectory, updates the policy and reiterates over the next episode. We first recall the standard LSVI. At the beginning of episode k, consider timestep t and assume the next-step parameter is fixed and equal to θ_{t+1}. The objective function of the regularized least-square is

$$\sum_{i=1}^{k-1} \left(\phi_{ti}^\top \theta - r_{ti} - V_{t+1}(\theta_{t+1})(s_{t+1,i}) \right)^2 + \lambda \|\theta\|_2^2 \tag{3}$$

where $\{\phi_{ti}\}_{i=1,\dots,k-1}$ are the features observed at timestep t in state s_{ti} and r_{ti} are the corresponding rewards. For any $\lambda > 0$ the prior display has a closed-form solution

$$\widehat{\theta}_t = \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \Big[r_{ti} + V_{t+1}(\theta_{t+1})(s_{t+1,i}) \Big] \tag{4}$$

with $\Sigma_{tk} \overset{def}{=} \sum_{i=1}^{k-1} \phi_{ti}\phi_{ti}^\top + \lambda I$ as the empirical covariance.

We introduce an optimistic variant of LSVI, where the optimistic parameters are chosen by solving a global optimization problem across the whole horizon H. At each episode, ELEANOR (in Alg. 1) solves the following problem.

Definition 2 (Planning Optimization Program)**.**

$$\max_{\substack{(\overline{\xi}_1,\dots,\overline{\xi}_H) \\ (\widehat{\theta}_1,\dots,\widehat{\theta}_H) \\ (\overline{\theta}_1,\dots,\overline{\theta}_H)}} \quad \max_a \phi_1(s_{1k},a)^\top \overline{\theta}_1 \quad \textit{subject to}$$

$$\widehat{\theta}_t = \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}^\top \left(r_{ti} + V_{t+1}(\overline{\theta}_{t+1})(s_{t+1,i}) \right)$$

$$\overline{\theta}_t = \widehat{\theta}_t + \overline{\xi}_t; \quad \|\overline{\xi}_t\|_{\Sigma_{tk}} \leqslant \sqrt{\alpha_{tk}}; \quad \overline{\theta}_t \in \mathcal{B}_t$$

As we will show in the technical analysis, a feasible solution $(\theta_1^\star, \dots, \theta_H^\star)$, corresponding to the best approximator (in eq. (9)) always exists and so the program is well posed.

The least-square solution $\widehat{\theta}_t$ is used as a constraint and perturbed by adding a vector $\overline{\xi}_t$ as optimization variable,[4]

[3] a positive function such that $\|\Psi_t\|_{TV} = 1$

[4] We add the subscript k later to indicate the actual variable chosen by the optimization procedure in episode k.

subject to

$$\|\overline{\xi}_t\|_{\Sigma_{tk}} \leqslant \sqrt{\alpha_{tk}} := \underbrace{\sqrt{\beta_{tk}}}_{\text{noise}} + \underbrace{\sqrt{\lambda}\mathcal{R}_t}_{\text{regularization}} + \underbrace{\sqrt{k}\mathcal{I}}_{\text{misspec.}}, \qquad (5)$$

where α_{tk} is designed to account for the noise, misspecification, and regularization bias. The actual bound is a function of the allowable radius $\mathcal{R} \leqslant \sqrt{d_t}$ for the parameter (as in assumption 1) and the noise parameter $\sqrt{\beta_{tk}} = \widetilde{O}(\sqrt{d_t})$ stems from self-normalizing concentration inequalities as described in the technical analysis later, while $\mathcal{I}$ is the inherent Bellman error. The resulting parameter $\overline{\theta}_t = \widehat{\theta}_t + \overline{\xi}_t$ must satisfy the constraint $\overline{\theta}_t \in \mathcal{B}_t$. This is equivalent to clipping the value function to avoid out-of-range values, with the difference that such clipping occurs directly in the parameter space as opposed to state by state, and thus preserves linearity.

We emphasize that the optimization over the $\overline{\xi}_t$'s is *global*, in stark contrast to the tabular setting and even the setting of linear MDPs considered by (Yang & Wang, 2020; Jin et al., 2020), where any perturbation (clipping, exploration bonus, etc) can be done state by state. For example, (Jin et al., 2020) define $\overline{Q}_t(s,a) \overset{\text{redefined}}{=} \min\{1, \phi_t(s,a)^\top \overline{\theta}_t + \text{BONUS}\}$ where the bonus is the result of maximizing $\overline{\xi}_t$ state by state. This trick works in the low-rank setting of (Jin et al., 2020), since any non-linear component is filtered out by the low-rank projector. ELEANOR instead pushes that maximization over the $\overline{\xi}_t$'s "outside" of local states, i.e., it performs a *global maximization* to ensure linearity of the value function representation, a mandatory condition in our setting to avoid an exponential propagation of the errors.

When linear representations are enforced, however, the algorithm cannot choose a value function everywhere optimistic due to values in different states possibly being negatively correlated. ELEANOR shoots for being optimistic at the initial state, but in general the algorithm does not play optimistic actions in the encountered states at later timesteps. Fortunately, this is enough to attain a rate-optimal efficiency.

Algorithm 1 ELEANOR

```
Input: failure probability $\delta$, regularization $\lambda = 1$, feature extractor $\phi$, inherent Bellman residual $\mathcal{I}$
Initialize $\Sigma_{t1} = \lambda I$, for $t = 1, 2, \ldots, H$.
for $k = 1, 2, \ldots$ do
    Receive starting state $s_{1k}$
    Set $\overline{\theta}_{H+1,k} = \widehat{\theta}_{H+1,k} = \overline{\xi}_{H+1,k} = 0$
    Solve program of definition 2.
    Execute $\pi_k : (s,t) \mapsto \arg\max_a \phi_t(s,a)^\top \overline{\theta}_{tk}$ and collect $(s_{tk}, a_{tk}, r_{tk})$ for $t \in [H]$.
end for
```

Although ELEANOR is proved to be near optimal, it is difficult to implement the algorithm efficiently. This should not be seen as a fundamental barrier, however. The issue of computational tractability arises even for tabular problems (Bartlett & Tewari, 2009; Zhang & Ji, 2019), but of course the problem is more pronounced when function approximators are implemented (Krishnamurthy et al., 2016; Jiang et al., 2017; Sun et al., 2018; Osband & Van Roy, 2014), and even for low-rank MDPs the first regret result has been obtained at the expense of a practical algorithm (Yang & Wang, 2020). Fortunately, later work has made progress on the computational aspects for many of these settings (Tossou et al., 2019; Fruit et al., 2018; Dann et al., 2018; Jin et al., 2020). For now, we leave this to future work.

Relaxations. With an eye towards a possible relaxation, we notice that the constraint $\overline{\theta}_t \in \mathcal{B}_t$ can be expensive to evaluate because it would require checking that every product $\phi_t(\cdot,\cdot)^\top \overline{\theta}_t$ is bounded. However, one can use simpler, more restrictive geometries and assume $\mathcal{B}_t$ is a unit ball, bypassing this problem. The algorithm regret bound for this case is the same as that of theorem 1.

Finally, it is possible to avoid the regularization in the least square objective of eq. (3) and relax the requirement $\|\overline{\theta}_t\|_2 \leqslant \sqrt{d_t}$ as presented later in assumption 1. In fact, the constraint on $\mathcal{B}_t$ suffices to avoid ill-conditioned solutions, but then one would need to resort to pseudo-inverse computations (Auer, 2002), making the algorithm / analysis more complicated.

5. Main Result: Regret Upper Bound

Assumption 1 (Main Assumption)**.** *We assume:*

- $|Q_t^\pi(s,a)| \leqslant 1, \quad \forall \pi, \forall (s,a,t)$
- $\|\phi_t(s,a)\|_2 \leqslant L_\phi \leqslant 1, \quad \forall (s,a,t)$
- *For any* $Q_t \in \mathcal{Q}_t$ *and any* $(s,a,t) \in \mathcal{S} \times \mathcal{A} \times [H]$ *define the random variable*[5] $X = R_t(s,a) + \max_{a'} Q_{t+1}(s',a')$. *Then the noise* $\eta = X - \mathbb{E}X$ *is* 1*-subgaussian*
- $\forall t \in [H], \forall \theta_t \in \mathcal{B}_t$, *it holds that* $\|\theta_t\| \leqslant \mathcal{R}_t \leqslant \sqrt{d_t}$, *and* $\mathcal{B}_t$ *is compact*

The first condition is a condition on the scaling of the problem and the bound on the feature norm is without loss of generality. The sub-Gaussianity is standard already for linear bandits (Abbasi-Yadkori et al., 2011; Lattimore & Szepesvári, 2020). In particular, if the reward are in $[0,1]$ and $D = 1$ in eq. (1), which gives $\overline{V}(\cdot) \in [-1,1]$, then this condition is automatically satisfied. Finally, the bound

[5] Here, $R_t(s,a)$ is the reward random variable, and $s' \sim p_t(s,a)$ is the successor state random variable under the distribution $p_t(s,a)$.

on the parameter limits the bias introduced by regularization which scales with the norm of the parameter, but a psedoinverse computation would relax this requirement.

After rescaling, however, our assumptions are much weaker the the usual setting that requires $r_t(\cdot,\cdot) \in [0,1]$ and $V^\pi_t(\cdot) \in [0,H]$ since we allow the reward to be of the same order as the value function after rescaling and even be negative. This is a harder setting (Jiang & Agarwal, 2018; Zanette & Brunskill, 2019).

Theorem 1 (Main Result)**.** *Under assumption 1 with $\lambda = 1$, with probability at least $1-\delta$ jointly over all episodes it holds that the regret of* ELEANOR *is bounded by:*

$$\textsc{Regret}(T) = \widetilde{O}\Big(\underbrace{\sum_{t=1}^{H} d_t\sqrt{K}}_{\textit{variance term}} + \underbrace{\sum_{t=1}^{H} \sqrt{d_t}\mathcal{I}K}_{\textit{approximation term}}\ \Big).$$

There are no additional "lower order" terms in the above display, although the $\widetilde{O}(\cdot)$ notation hides, as usual, logarithms of $d_t, H, K, 1/\delta$.

Care must be taken when comparing across settings with different scaling. In particular, *rescaling the problem* (i.e., the reward function) *by H* increases the sub-Gaussian norm of the rewards and transitions, and the value of the inherent Bellman error alike, yielding *an extra H factor in the regret bound*. For example, in the setting that the rewards are bounded in $[0,1]$ and the value function is in $[0,H]$ with $d_1 = \dots = d_H \overset{def}{=} d$ and $\mathcal{I} = 0$ for simplicity, the above regret bound reduces (with $T = KH$) to $\widetilde{O}(dH^{\frac{3}{2}}\sqrt{T})$.

Low-rank MDPs As explained in proposition 3, our result applies to low-rank MDPs; surprisingly, this shows that at least $\sqrt{d}$ improvement is possible in the main rate compared to the best-known $\widetilde{O}((dH)^{3/2}\sqrt{T})$ of (Jin et al., 2020) upper bound despite ELEANOR is not specifically tailored to handle low-rank MDPs. This is possible because ELEANOR looks for optimistic solutions directly in the θ parameter space instead of perturbing the value function by an exploration bonus as in (Jin et al., 2020). When the value function is perturbed by a bonus, it grows in complexity as it departs from the linear space; this requires an additional union bound over a more complicated value function class and ultimately loses a $\sqrt{d}$ factor. Finally, the inherent Bellman error covers the notion of approximate low-rank MDPs (Jin et al., 2020), and on the misspecification regret term we save a $\sqrt{d}$ factor as well thanks to a more careful projection argument in lemma 8.

6. Contextual Misspecified Linear Bandits

Our framework reduces to bandits with linear approximators when $H = 1$ (we drop the time subscript t in this case): ELEANOR can handle *contextual misspecified linear bandits*, where contextual refers to allowing the action set to change as the feature extractor can be a function of the context. It follows from the definition that the inherent Bellman error is the reward function misspecification in this case.

Corollary 1 (LINUCB Regret on Contextual Misspecified Linear Bandits)**.** *Consider a misspecified contextual linear bandit problem with reward response*

$$r(s,a) = \phi(s,a)^\top\theta^\star + \eta + f(s,a)$$

with $|\phi(s,a)^\top\theta^\star| \leqslant 1$, $\|\theta^\star\|_2 \leqslant \sqrt{d}$, $\|\phi(s,a)\|_2 \leqslant 1$, misspecification $|f(s,a)| \leqslant \mathcal{I}$ and 1 sub-Gaussian noise η. If ELEANOR *is informed that $H = 1$ then the algorithm reduces to the* LINUCB *(aka* OFUL*) algorithm of (Abbasi-Yadkori et al., 2011) with arm selection strategy $\arg\max_{a\in\mathcal{A}, \|\overline{\xi}\|_{\Sigma_k} \leqslant \sqrt{\alpha_k}} \phi(s_k,a)^\top\left(\widehat{\theta}_k + \overline{\xi}_k\right)$ but a different confidence interval: $\|\overline{\theta}_k - \widehat{\theta}_k\|_{\Sigma_k} = \|\overline{\xi}_k\|_{\Sigma_k} \leqslant \sqrt{\alpha_k}$. The arm selection strategy admits the closed-form solution $\arg\max_{a\in\mathcal{A}}\left[\phi(s_k,a)^\top\widehat{\theta}_k + \|\phi(s_k,a)\|_{\Sigma_k^{-1}}\sqrt{\alpha_k}\right]$ and the algorithm has a high probability regret bound*

$$\widetilde{O}\left(d\sqrt{K} + \sqrt{d}\mathcal{I}K\right).$$

The corollary above is immediate upon substituting $H = 1$ in theorem 1 and verifying that our assumptions match the setting described in the corollary, which is the standard linear bandit setting[6] (Lattimore & Szepesvári, 2020) with the addition of misspecification (few more details in appendix E).

Due to the equivalence to LINUCB the algorithm is computationally tractable when applied to bandits; the *key* difference with vanilla LINUCB resides in the width of the confidence intervals, parameter α_k. In the absence of misspecification ($\mathcal{I} = 0$), $\sqrt{\alpha_k} = \sqrt{\beta_k} + \sqrt{\lambda}\mathcal{R} = \widetilde{O}(\sqrt{d})$, as in the work of (Abbasi-Yadkori et al., 2011). When misspecification is present, however, there is a correction factor $\sqrt{k}\mathcal{I}$ in the definition of $\sqrt{\alpha_k}$, see equation eq. (5). In other words, this is the factor one should add to the exploration bonus for an LinUCB-like algorithm in case of (potentially adversarial) misspecification.

The recent result by (Du et al., 2019) applies here (see also the work of (Van Roy & Dong, 2019)). They show that for large misspecification $\mathcal{I} \gtrapprox 1/\sqrt{d}$ an exponential sample complexity is unavoidable to identify an arm with positive return. This does not contradict our result, because our regret is $\widetilde{O}(K)$ under such large misspecification, which is vacuous as the maximum loss up to episode K is exactly K.

Notice that the equivalence is established by informing ELEANOR of the setting (through the horizon $H = 1$) unlike (Zanette & Brunskill, 2018). Finally, if the corruption

[6] We drop the constraint $\theta \in \mathcal{B}$ for simplicity

$f(\cdot)$ is only a function of the context then it is possible to do much better (Krishnamurthy et al., 2018).

This surprising connection with the popular LINUCB makes ELEANOR (or LINUCB with a correction on the exploration bonus) the first algorithm capable of handling misspecified *contextual* linear bandits, although we are not the first to consider misspecification in linear bandits per se: (Ghosh et al., 2017) propose an algorithm that switches to tabular if misspecification is detected and (Gopalan et al., 2016) consider the case that the misspecification is less than roughly the action gap; (Van Roy & Dong, 2019) comment on the lower bound by (Du et al., 2019) using the Eluder dimension. Finally, (Lattimore & Szepesvari, 2020) have recently obtained a result similar to ours, but for a different setting. Their algorithm can leverage having finitely many actions (where a $\sqrt{d}$ factor can be saved; otherwise their regret is the same as ours) but relies heavily on G-experimental design: the algorithm will not work without a stationary action set, ruling out the important case of contextual linear bandits where the action is allowed to depend on the context. However, our correction to vanilla LINUCB relies on having knowledge of the misspecification, while the approach of (Lattimore & Szepesvari, 2020) is agnostic. Furthermore, concurrently to our work (Lattimore & Szepesvari, 2020) also consider the same modification to LINUCB as we do here, and provide proof that the algorithm can fail if no modification is implemented. However, these definitions of misspecification are adversarial in nature, and for less pathological problems the algorithm is expected to perform well.

7. Lower Bounds

In terms of statistical rate, ELEANOR is unimprovable due to a lower bound directly borrowed from the bandit literature.

Proposition 1 (Lower Bound Without Misspecification)**.** *Let $\widetilde{d} \overset{def}{=} \sum_{t=1}^{H} d_t$. There exist a class of H-horizon MDPs that satisfy asm. 1 and H feature maps $\phi_t(\cdot,\cdot) \in \mathbb{R}^{d_t}$, with $\widetilde{d} \geqslant 2H$ such that for $K = \Omega(\widetilde{d}^2)$ the expected regret of any algorithm is $\Omega(\widetilde{d}\sqrt{K})$.*

The fact that our result matches the lower bound can appear surprising, because our work relies on a sub-Gaussian conditions and disregards the variance in the process. It does not use a "law of total variance" argument (Azar et al., 2012; 2017), which was necessary in the past to obtain rate-optimal algorithms for tabular settings. One may wonder whether a $\sqrt{H}$ factor can be saved by that argument for MDPs parameterized by linear action-value function. Due to the bandit lower bound, no such improvement is possible with linear function approximations, unless the structure is restricted further. The reason is that our setting is a superset of tabular RL (Azar et al., 2017) and contains harder instances than the lower bound for tabular RL (in particular, a linear bandit problem at a single timestep) but the law of total variance would bring no benefit to those structures.

Approximation error Our positive result regarding misspecification matches the LSPI analysis of (Lattimore & Szepesvari, 2020) but for the harder *online* setting. Although the two respective frameworks (i.e., LSPI vs LSVI conditions) are incompatible as explained in proposition 5, we notice a similar effect: a square-root factor of the problem dimensionality multiplies the "misspecification" error. While the LSPI analysis of (Lattimore & Szepesvari, 2020) relies on having features from G-optimal design to query the system, *in the online setting we're not free to choose arbitrary features anywhere in the state-action space*. As a result, the agent can learn on an ill-conditioned basis, and the prediction error on features much different from those experienced can be very large. Our analysis shows that while this can indeed be the case, the situation of high prediction error cannot persist for too long and the $\sqrt{d}$ loss in prediction accuracy is, *on average*, recovered. Using the recent result by (Du et al., 2019), we can augment proposition 1 by including a sequence of misspecified linear bandits, obtaining the following result (see also appendix D):

Theorem 2 (Lower Bound for Inherent Bellman Error Setting)**.** *There exist feature maps $\phi_1, \ldots, \phi_H$ that define an MDP class $\mathcal{M}$ such that every MDP in that class satisfies assumption 1 with inherent Bellman error $\mathcal{I}$ and such that the expected regret of any algorithm on at least a member of the class (for $A \geqslant 3, d_t \geqslant 3, K = \Omega((\sum_{t=1}^{H} d_t)^2)$) is $\Omega(\sum_{t=1}^{H} d_t\sqrt{K} + \sum_{t=1}^{H} \sqrt{d_t}\mathcal{I}K)$, that is:*

$$\min_{\mathscr{A}} \max_{M \in \mathcal{M}} \sum_{k=1}^{K} \left(V_1^\star - V_1^{\pi_k}\right)(s_{1k}) \\ = \Omega(\sum_{t=1}^{H} d_t\sqrt{K} + \sum_{t=1}^{H} \sqrt{d_t}\mathcal{I}K).$$

8. Proof Overview

We now give a quick proof sketch and highlighting how working in the parameter space allows us to 1) avoid an exponential propagation of the errors by leveraging the notion of inherent Bellman error (handling of the bias) and 2) preserve confidence intervals that are as tight as in a bandit problem (handling of the variance). Our objective is to bound the regret: $\textsc{Regret}(K) \overset{def}{=} \sum_{k=1}^{K} \left(V_1^\star - V_1^{\pi_k}\right)(s_{1k})$ for the chosen policies π_k, but first we need to discuss how the errors propagate and how to ensure optimism.

8.1. Propagation of errors

The inherent Bellman error condition ensures that there exists a parameter $\mathring{\theta}_t$ and a Bellman residual function $\mathring{\Delta}_t$, both depending on $\overline{Q}_{t+1}$, such that $\mathring{\Delta}_t(\overline{Q}_{t+1})(s,a) =$

$$= \phi_t(s,a)^\top \mathring{\theta}_t(\overline{Q}_{t+1}) - \left(\mathcal{T}_t\overline{Q}_{t+1}\right)(s,a) \tag{6}$$

with $\|\mathring{\Delta}_t(\overline{Q}_{t+1}))\|_\infty \leqslant \mathcal{I}$ *provided that* $\overline{Q}_{t+1} \in \mathcal{Q}_{t+1}$. In other words, we can successfully represent $\mathcal{T}_t\overline{Q}_{t+1}$ up to an additive error $\mathcal{I}$ *if the next-step* $\overline{Q}_{t+1}$ *function is linear*.

This representational constraint unfortunately rules out adding exploration bonuses as in prior low-rank work (Yang & Wang, 2020; Jin et al., 2020) as well as in tabular MDPs; their addition can have the backup $\mathcal{T}_t\overline{Q}_{t+1}$ leave the linear space (which is equivalent to having large $\mathcal{I}$) and can lead to divergence of the repeated least-square procedure (Baird, 1995; Sutton & Barto, 2018; Zanette et al., 2019b).

Error decomposition We aim to compute the error encountered in minimizing eq. (3) with $V_{t+1} = \overline{V}_{t+1}$ fixed and no regularization. Denote with s_{ti} the i-th state encountered at timestep t of episode i, and let $a_{ti} = \pi_{ti}(s_{ti})$. Define the i-th sample noise $\eta_{ti}(\overline{V}_{t+1}) \overset{def}{=} r_{ti} - r_t(s_{ti},a_{ti}) + \overline{V}_{t+1}(s_{t+1,i}) - \mathbb{E}_{s'\sim p_t(s_{ti},a_{ti})}\overline{V}_{t+1}(s')$ and the misspecification $\mathring{\Delta}_{ti}(\overline{Q}_{t+1}) \overset{def}{=} \mathring{\Delta}_t(\overline{Q}_{t+1})(s_{ti},a_{ti})$. Premultiply $\widehat{\theta}_{tk}$ (which minimizes eq. (3)) by $\phi_t(s,a)^\top$ and use the definitions just introduced: $\phi_t(s,a)^\top\widehat{\theta}_{tk} =$

$$\begin{aligned}
&\phi_t(s,a)^\top\Sigma_{tk}^{-1}\sum_{i=1}^{k-1}\phi_{ti}\left(\mathcal{T}_t\overline{Q}_{t+1}(s_{ti},a_{ti}) + \eta_{ti}(\overline{V}_{t+1})\right)\\
&= \phi_t(s,a)^\top\Big[\mathring{\theta}_t(\overline{Q}_{t+1}) +\\
&+ \Sigma_{tk}^{-1}\sum_{i=1}^{k-1}\phi_{ti}\left(\mathring{\Delta}_{ti} + \eta_{ti}\right)(\overline{Q}_{t+1})\Big]\\
&\overset{eq.\ (6)}{=} \mathcal{T}_t(\overline{Q}_{t+1})(s,a) + \mathring{\Delta}_t(\overline{Q}_{t+1})(s,a) +\\
&+ \phi_t(s,a)^\top\Sigma_{tk}^{-1}\sum_{i=1}^{k-1}\phi_{ti}\left(\mathring{\Delta}_{ti} + \eta_{ti}\right)(\overline{Q}_{t+1}).
\end{aligned} \tag{7}$$

We discuss the main error terms below.

Inherent Bellman error Cauchy-Schwartz and a projection argument (lemma 8) gives:

$$|\phi_t(s,a)^\top\Sigma_{tk}^{-1}\sum_{i=1}^{k-1}\phi_{ti}\mathring{\Delta}_{ti}(\overline{Q}_{t+1})| \leqslant \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}\sqrt{k}\mathcal{I}.$$

The inability to correctly represent the application of the Bellman operator could be exploited adversarially to introduce an error that grows with $\sqrt{k}$ (where k is the number of episodes). On average, however, the Σ_{tk}^{-1}-norm of those features that are selected shrinks as $\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \approx \sqrt{d_t/k}$. While the agent can select a (s,a) pair where the product $\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}\sqrt{k}\mathcal{I}$ can be large, this cannot happen for too long. Intuitively, a large prediction error is made only on features that are significantly different from those seen in the past, but trying those features reveals the correct prediction, which decreases the prediction error for that direction in the future.

Noise error and covering argument Cauchy-Schwartz again gives

$$\begin{aligned}
&|\phi_t(s,a)^\top\Sigma_{tk}^{-1}\sum_{i=1}^{k-1}\phi_{ti}\eta_{ti}(\overline{V}_{t+1})|\\
&\leqslant \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}\|\sum_{i=1}^{k-1}\phi_{ti}\eta_{ti}(\overline{V}_{t+1})\|_{\Sigma_{tk}^{-1}}\\
&\overset{def}{\leqslant} \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}\sqrt{\beta_{tk}}
\end{aligned}$$

where β_{tk} follows from the self normalizing bound of (Abbasi-Yadkori et al., 2011) modified to cover the functional space $\mathcal{V}_t$. The covering argument is necessary since the noise depends on $\overline{V}_{t+1}$ which is itself random. More precisely, we can write $\sqrt{\beta_{tk}} \lessapprox \sqrt{\ln\det(\Sigma_{tk})^{\frac{1}{2}} + \ln\mathcal{N}}$, where $\mathcal{N}$ is the covering number to ϵ accuracy of $\mathcal{V}_{t+1}$. The determinant-trace inequality (see lemma 10 of (Abbasi-Yadkori et al., 2011)) bounds the volume of the covariance matrix $\ln\det(\Sigma_{tk})^{\frac{1}{2}} = \widetilde{O}(d_t)$; fortunately the metric entropy $\ln\mathcal{N}$ is of the same order. To see this, remember that to cover $\mathcal{V}_t$ it is sufficient to cover $\mathcal{B}_t$, which is a d_t dimensional object ($\subset \mathbb{R}^{d_t}$), and hence $\ln\mathcal{N} = \widetilde{O}(d_t)$. Therefore, despite having an additional union bound compared to (Abbasi-Yadkori et al., 2011) because of the moving target $\overline{V}_{t+1}$, our confidence intervals are of the same order of magnitude.

This is the place where a $\sqrt{d_t}$ can be saved compared to for example (Jin et al., 2020; Wang et al., 2019), which need to do a union bound over a more complicated function class because of the exploration bonuses.

Final expression Adding $\phi_t(s,a)^\top\overline{\xi}_t$ to both sides of eq. (7) and using the bounds just derived gives $|\left(\overline{Q}_t - \mathcal{T}_t\overline{Q}_{t+1}\right)(s,a)| =$

$$\begin{aligned}
&\leqslant \underbrace{\mathcal{I}}_{\text{misspecification}} + \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}\times\\
&\Big(\underbrace{\sqrt{k}\mathcal{I}}_{\text{misspecification}} + \underbrace{\sqrt{\alpha_{tk}}}_{\text{exploration}} + \underbrace{\sqrt{\beta_{tk}}}_{\text{noise}}\Big).
\end{aligned} \tag{8}$$

It remains to define α_{tk}, which controls the size of optimization parameters, justifying eq. (5).

8.2. Feasibility, best approximator and optimism

A key point of optimistic approaches for exploration is to overestimate the value of policies by assigning them a statistically plausible return, and play the policy with the highest such value.

Since the optimal value function is an upper bound to the value of all policies, technically an optimistic learner is only required to identify a policy with value at least as high as $V_1^\star$ while satisfying some confidence intervals. To show it possible to achieve this with our formulation, we will find a feasible solution to the program of definition 2 that is "close" to $V^\star$. In general $V_t^\star \notin \mathcal{V}_t$, and so we need to define the "best" approximator in $\mathcal{V}_t$ for $V_t^\star$. We denote its parameter with $\theta_t^\star \in \mathcal{B}_t$, inductively defined (see def. 4 in appendix) as the parameter one obtains by applying the *exact* Bellman operator and then by minimizing the ∞ norm of the Bellman residual: $\theta_t^\star \overset{def}{=}$

$$\operatorname*{arg\,min}_{\theta \in \mathcal{B}_t} \sup_{(s,a)} \left| \phi_t(s,a)^\top \theta - \left(\mathcal{T}_t Q_{t+1}(\theta_{t+1}^\star)\right)(s,a) \right| \quad (9)$$

If $\mathcal{I} = 0$ then $\phi_t(s,a)^\top \theta_t^\star = Q_t^\star(s,a)$ inductively follows.

Computation of α_{tk} Under an inductive argument, assume the program of definition 2 admits a partial solution $\overline{\xi}_{t+1}, \dots, \overline{\xi}_H$ that satisfies $\overline{\theta}_{t+1} = \theta_{t+1}^\star, \dots, \overline{\theta}_H = \theta_H^\star$ (the parameters for timesteps less than $t+1$ have not been decided yet).

Now setting:

$$\overline{\xi}_t = -\Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \left(\mathring{\Delta}_{ti} + \eta_{ti} \right) (Q_{t+1}(\theta_{t+1}^\star)) \quad (10)$$

and adding $\phi_t(s,a)^\top \overline{\xi}_t$ back to eq. (7) evaluated with $\overline{Q}_{t+1} = Q_{t+1}(\theta_{t+1}^\star)$ can "undo" the effect of noise and approximation error at timestep t, producing (recall $\overline{\theta}_t = \widehat{\theta}_t + \overline{\xi}_t$)

$$\begin{aligned} &\phi_t(s,a)^\top \overline{\theta}_t \\ &= \mathcal{T}_t(Q_{t+1}(\theta_{t+1}^\star))(s,a) + \mathring{\Delta}_t(Q_{t+1}(\theta_{t+1}^\star))(s,a). \end{aligned}$$

Comparing with eq. (9) we can claim $\overline{\theta}_t = \theta_t^\star$, completing the induction. Thus, the best approximator defined through $\theta_t^\star$ is a feasible solution to the program of definition 2. The corresponding value function $V_t(\theta_t^\star)$ can make an error of size $\mathcal{I}$ in representing the Bellman backup, and this accumulates linearly, and hence ELEANOR is ultimately nearly-optimistic:

$$\overline{V}_1(s_{1k}) \geqslant V_1^\star(s_{1k}) - H\mathcal{I}. \quad (11)$$

As we'll see in a second, this near-optimism is enough to obtain a solid regret bound. Finally, eq. (10) gives:

$$\|\overline{\xi}_t\|_{\Sigma_{tk}} \leqslant \underbrace{\left\| \sum_{i=1}^{k-1} \phi_{ti} \Delta_{ti} \right\|_{\Sigma_{tk}^{-1}}}_{\leqslant \sqrt{k}\mathcal{I}} + \underbrace{\left\| \sum_{i=1}^{k-1} \phi_{ti} \eta_{ti} \right\|_{\Sigma_{tk}^{-1}}}_{\leqslant \sqrt{\beta_{tk}}} \quad (12)$$

which matches eq. (5) after adding the regularization term.

8.3. Regret Bound

Finally, we can present the regret bound, which now follows similarly to prior analyses for model free algorithms (e.g., (Jin et al., 2018)). Consider the usual decomposition from the starting state s_{1k}:

$$\textsc{Regret}(K) \overset{def}{=} \sum_{k=1}^{K} \left(V_1^\star - \overline{V}_{1k} + \overline{V}_{1k} - V_1^{\pi_k} \right)(s_{1k}).$$

The first term inside the parenthesis can be bounded by eq. (11); we can expand the second term using eq. (8) where π_k is the agent's policy in episode k and $a_{tk} = \pi_{tk}(s_{tk})$ for short. For a generic timestep t we obtain

$$\begin{aligned} \left(\overline{V}_{tk} - V_t^{\pi_k}\right)(s_{tk}) \leqslant \Bigg[& \mathbb{E}_{s' \sim p_t(s_{tk}, a_{tk})} \left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s') \\ & + \mathcal{I} + \|\phi_t(s_{tk}, a_{tk})\|_{\Sigma_{tk}^{-1}} \underbrace{\left(\sqrt{k}\mathcal{I} + \sqrt{\alpha_{tk}} + \sqrt{\beta_{tk}}\right)}_{\approx \widetilde{O}(\sqrt{k}\mathcal{I} + \sqrt{d_t})} \Bigg]. \end{aligned}$$

Now write $\mathbb{E}_{s' \sim p_t(s_{tk}, a_{tk})} \left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s')$ as $\left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s_{t+1,k})$ plus a martingale term $\dot{\zeta}_{tk}$ which we ignore for brevity (details in appendix). Induction over $t \in [H]$ and summing over $k \in [K]$ gives $\sum_{k=1}^{K} \sum_{t=1}^{H} \left(\overline{V}_{1k} - V_1^{\pi_k}\right)(s_{1k})$

$$\leqslant \sum_{k=1}^{K} \sum_{t=1}^{H} \left[\mathcal{I} + \|\phi_t(s_{tk}, a_{tk})\|_{\Sigma_{tk}^{-1}} \times \widetilde{O}(\sqrt{k}\mathcal{I} + \sqrt{d_t}) \right].$$

Recall $\sum_{k=1}^{K} \|\phi_t(s_{tk}, a_{tk})\|_{\Sigma_{tk}^{-1}} = \widetilde{O}(\sqrt{d_t K})$ from (Abbasi-Yadkori et al., 2011); substituting this concludes.

9. Conclusion

We have introduced an algorithm for online exploration with linear approximators under the notion of low-inherent Bellman error with an optimal regret bound with regards to statistical rates and the lack of closedness of the Bellman operator. The construction reveals that a shift to global optimization might be unavoidable with more general linear approximators than prior low-rank work, making computational tractability harder to achieve. A core idea is that by working directly in the parameter space we enable a linear propagation of the errors (as opposed to exponential) and we limit the complexity of the value function class, which can serve as inspiration to improve the statistical efficiency for other algorithms as well. Finally, a noteworthy contribution is our analysis for misspecified contextual linear bandit, which explains that a simple modification of a mainstream algorithm is sufficient to handle such setting.

Acknowledgments

Andrea Zanette is partially supported by a Total Innovation Fellowship. We thank Alekh Agarwal for pointing our the connection with the low Bellman rank setting. The authors are grateful to the reviewers for their helpful comments.

References

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Agarwal, A., Kakade, S., and Yang, L. F. On the optimality of sparse model-based planning for markov decision processes. *arXiv preprint arXiv:1906.03804*, 2019.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Azar, M., Munos, R., and Kappen, H. J. On the sample complexity of reinforcement learning with a generative model. In *International Conference on Machine Learning (ICML)*, 2012.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.

Baird, L. Residual algorithms: Reinforcement learning with function approximation. In *International Conference on Machine Learning (ICML)*. 1995.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1429–1439, 2018.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516, 2019.

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

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, 2019.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. https://arxiv.org/abs/1802.04020, 2018.

Ghosh, A., Chowdhury, S. R., and Gopalan, A. Misspecified linear bandits. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Golub, G. H. and Van Loan, C. F. *Matrix Computations*. JHU Press, 2012.

Gopalan, A., Maillard, O.-A., and Zaki, M. Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*, 2016.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

Jiang, N. and Agarwal, A. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference on Learning Theory (COLT)*, pp. 3395–3398, 2018.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low Bellman rank are PAC-learnable. In Precup, D. and Teh, Y. W. (eds.), *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1704–1713, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/jiang17c.html.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

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

Kolter, J. Z. The fixed points of off-policy td. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2169–2177, 2011.

Krishnamurthy, A., Agarwal, A., and Langford, J. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1840–1848, 2016.

Krishnamurthy, A., Wu, S., and Syrgkanis, V. Semiparametric contextual bandits. In *35th International Conference on Machine Learning, ICML 2018*, pp. 4330–4349. International Machine Learning Society (IMLS), 2018.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *Journal of machine learning research*, 4(Dec): 1107–1149, 2003.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Lattimore, T. and Szepesvari, C. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning (ICML)*, 2020.

Lazaric, A., Ghavamzadeh, M., and Munos, R. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13(Oct):3041–3074, 2012.

Maillard, O.-A., Mann, T. A., and Mannor, S. “how hard is my MDP?” the distribution-norm to the rescue. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Munos, R. Error bounds for approximate value iteration. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2005.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (May):815–857, 2008.

Osband, I. and Van Roy, B. Near-optimal reinforcement learning in factored mdps. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994. ISBN 0471619779.

Russo, D. Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, 2019.

Shreve, S. E. and Bertsekas, D. P. Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control. *SIAM Journal on control and optimization*, 16(6):953–978, 1978.

Sidford, A., Wang, M., Wu, X., Yang, L. F., and Ye, Y. Near-optimal time and sample complexities for for solving discounted markov decision process with a generative model. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Simchowitz, M. and Jamieson, K. Non-asymptotic gap-dependent regret bounds for tabular mdps. *arXiv preprint arXiv:1905.03814*, 2019.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based reinforcement learning in contextual decision processes. *arXiv preprint arXiv:1811.08540*, 2018.

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

Tossou, A., Basu, D., and Dimitrakakis, C. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. *arXiv preprint arXiv:1905.12425*, 2019.

Tsitsiklis, J. N. and Van Roy, B. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1-3):59–94, 1996.

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

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Yang, L. F. and Wang, M. Sample-optimal parametric q-learning with linear transition models. In *International Conference on Machine Learning (ICML)*, 2019.

Yang, L. F. and Wang, M. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning (ICML)*, 2020.

Zanette, A. and Brunskill, E. Problem dependent reinforcement learning bounds which can identify bandit structure in mdps. In *International Conference on Machine Learning (ICML)*, 2018.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning (ICML)*, 2019. URL http://proceedings.mlr.press/v97/zanette19a.html.

Zanette, A., Brunskill, E., and J. Kochenderfer, M. Almost horizon-free structure-aware best policy identification with a generative model. In *Advances in Neural Information Processing Systems*, 2019a.

Zanette, A., Lazaric, A., J. Kochenderfer, M., and Brunskill, E. Limiting extrapolation in linear approximate value iteration. In *Advances in Neural Information Processing Systems*, 2019b.

Zanette, A., Brandfonbrener, D., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *AISTATS*, 2020.

Zhang, Z. and Ji, X. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, pp. 2827–2836, 2019.