Model-Based Reinforcement Learning with Multinomial Logistic Function Approximation

Taehyun Hwang, Min-hwan Oh*

Seoul National University, Seoul, Republic of Korea
th.hwang@snu.ac.kr, minoh@snu.ac.kr

Abstract

We study model-based reinforcement learning (RL) for episodic Markov decision processes (MDP) whose transition probability is parametrized by an unknown transition core with features of state and action. Despite much recent progress in analyzing algorithms in the linear MDP setting, the understanding of more general transition models is very restrictive. In this paper, we propose a provably efficient RL algorithm for the MDP whose state transition is given by a multinomial logistic model. We show that our proposed algorithm based on the upper confidence bounds achieves $\widetilde{\mathcal{O}}(d\sqrt{H^3T})$ regret bound where d is the dimension of the transition core, H is the horizon, and T is the total number of steps. To the best of our knowledge, this is the first model-based RL algorithm with multinomial logistic function approximation with provable guarantees. We also comprehensively evaluate our proposed algorithm numerically and show that it consistently outperforms the existing methods, hence achieving both provable efficiency and practical superior performance.

Introduction

Reinforcement learning (RL) with function approximation has made significant advances in empirical studies (Mnih et al. 2015; Silver et al. 2017, 2018). However, the theoretical understanding of these methods is still limited. Recently, function approximation with provable efficiency has been gaining significant attention in the research community, trying to close the gap between theory and empirical findings. Most of the existing theoretical works in RL with function approximation consider linear function approximation (Jiang et al. 2017; Yang and Wang 2019, 2020; Jin et al. 2020; Zanette et al. 2020; Modi et al. 2020; Du et al. 2020; Cai et al. 2020; Ayoub et al. 2020; Wang, Salakhutdinov, and Yang 2020; Weisz, Amortila, and Szepesvári 2021; He, Zhou, and Gu 2021; Zhou, Gu, and Szepesvari 2021; Zhou, He, and Gu 2021; Ishfaq et al. 2021). Many of these linear model-based methods and their analyses rely on the classical upper confidence bound (UCB) or randomized exploration methods such as Thompson sampling extending the analysis of linear contextual bandits (Chu et al. 2011; Abbasi-Yadkori, Pál, and Szepesvári 2011; Agrawal and Goyal 2013; Abeille and Lazaric 2017; Kveton et al. 2020a).

While new methods are still being proposed under the linearity assumption and performance guarantees have been improved, the linear model assumption on the transition model of Markov decision processes (MDPs) faces a simple yet fundamental challenge. A transition model in MDPs is a *probability distribution* over states. A linear function approximating the transition model needs to satisfy that the function output is within $[0, 1]$ and, furthermore, the probabilities over all possible next states sum to 1 exactly. Note that such requirements are not just imposed approximately, but rather exactly, since almost all existing works in linear function approximation assume realizability, i.e., the true transition model is assumed to be linear (Yang and Wang 2020; Jin et al. 2020; Zanette et al. 2020; Zhou, Gu, and Szepesvari 2021; Ishfaq et al. 2021).

The linear model assumption also limits the set of feature representations of states or state-action pairs that are admissible for the transition model. In function approximation settings, the transition models are typically functions of feature representations. However, for a given linear transition model, an arbitrary feature may not induce a proper probability distribution. Put differently, nature can reveal a set of feature representations such that no linear model can properly construct a probability distribution over states. Hence, the fundamental condition required for the true transition model can easily be violated for the linear model. This issue becomes even more challenging for a *estimated* model.[1] Furthermore, when there is model misspecification, sublinear guarantees on the regret performances that the existing methods enjoy become not valid, hence potentially leading to serious deterioration of the performances.

In supervised learning paradigm, a distribution over multiple possible outcomes is rarely learned using a separate linear model for each outcome. For example, consider a learning problem with a binary outcome. One of the most obvious choices of a model to use in such a learning problem is a *logistic* model. Acknowledging that the state transition model of MDPs with a finite number of next states (the total num-

*Corresponding Author

[1] That is, even if the true model is truly linear and satisfies the basic conditions for a probability distribution, the estimated model can still violate them.

ber of states can be infinite) is essentially a categorical distribution, the *multinomial logistic* (MNL) model is certainly one of the first choices to consider. In statistics and machine learning research, the generalization of the linear model to a function class suitable for particular problem settings has been an important milestone both in terms of applicability and theoretical perspectives. In parametric bandit research, a closely related field of RL, the extension of linear bandits to generalized linear bandits, including logistic bandits for binary feedback and multinomial logistic bandits for multi-class feedback, has been an active area of research (Filippi et al. 2010; Li, Lu, and Zhou 2017; Jun et al. 2017; Kveton et al. 2020b; Oh and Iyengar 2019, 2021). Surprisingly, there has been no prior work on RL with *multinomial logistic* function approximation (or even logistic function approximation), in spite of a vast amount of literature on linear function approximation and despite the fact that the multinomial logistic model can naturally capture the state transition probabilities.

To the best of our knowledge, our work is the first to study a provably efficient RL under the multinomial logistic function approximation. The generalization of the transition probability model beyond the simple linear model to the multinomial logistic model allows for broader applicability, overcoming the crucial limitations of the linear transition model. On theoretical perspectives, going beyond the linear model to the MNL model requires more involved analysis without closed-form solutions for the estimation and accounting for non-linearity. Note that the linear model assumption for the transition model induces a linear value function which enables the utilization of the least-square estimation and the linear bandit techniques for regret analysis (Jin et al. 2020; Zanette et al. 2020; Ishfaq et al. 2021). However, in the MNL function approximation, we no longer have a linearly parametrized value function nor do we have any closed form expression for the value function. It appears that these aspects pose greater technical challenges in RL with MNL function approximation. Therefore, the following research question arises:

Can we design a provably efficient RL algorithm for the multinomial logistic transition model?

In this paper, we address the above question affirmatively. We study a finite-horizon RL problem where the transition probability is assumed to be a MNL model parametrized by a transition core. We propose a provably efficient model-based RL algorithm that balances the exploration-exploitation trade-off, establishing the first results for the MNL transition model approximation of MDPs. Our main contributions are summarized as follows:

- We formally discuss the shortcomings of the linear function approximation (e.g., Proposition 1). To the best of our knowledge, the rigorous discussion on the limitation of the linear transition model provides meaningful insights and may be of independent interest. Our finding is that the linear transition model is restricted not just in the functional form, but also the set of features that satisfy the requirement imposed by the linear MDP is very limited.
- The MNL function approximation that we study in this paper is a more flexible and practical function approximation than the linear function approximation which had been studied extensively in the recent literature. To our best knowledge, our paper is the first work to consider multinomial logistic function approximation (that provides provable guarantees) and hence, we believe, serves as an important milestone. We believe such a modeling assumption not only naturally captures the essence of the state transition probabilities, overcoming the drawbacks of the linear function approximation, but also induces an efficient algorithm that utilizes the structure.
- We propose a provably efficient algorithm for model-based RL in feature space, Upper Confidence RL with MNL transition model (`UCRL-MNL`). To the best of our knowledge, this is the first *model-based* RL algorithm with multinomial logistic function approximation.
- We establish that `UCRL-MNL` is statistically efficient achieving $\widetilde{\mathcal{O}}(d\sqrt{H^3T})$ regret, where d is the dimension of the transition core, H is the planning horizon, and T is the total number of steps. Noting that d is the total dimension of the unknown parameter, the dependence on dimensionality as well as dependence on total steps matches the corresponding dependence of the regret bound in linear MDPs (Zhou, Gu, and Szepesvari 2021).
- We evaluate our algorithm on numerical experiments and show that it consistently outperforms the existing provably efficient RL methods by significant margins. We performed experiments on tabular MDPs. Hence, no modeling assumption on the true functional form is imposed for the transition model, which does not favor any particular model of approximation. The experiments provide the evidences that our proposed algorithm is both provably and practically efficient.

The MNL function approximation that we study in this paper is a much more flexible and practical generalization of the tabular RL than linearly parametrized MDPs, which have been widely studied in the recent literature. As the first work to study RL with MNL transition model, we believe that both our proposed transition model and the proposed algorithm provide sound contributions in terms of theory and practicality.

Related Work

For tabular MDPs with a finite H-horizon, there are a large number of works both on model-based methods (Jaksch, Ortner, and Auer 2010; Osband and Roy 2014; Azar, Osband, and Munos 2017; Dann, Lattimore, and Brunskill 2017; Agrawal and Jia 2017; Ouyang et al. 2017) and on model-free methods (Jin et al. 2018; Osband et al. 2019; Russo 2019; Zhang, Zhou, and Ji 2020, 2021). Both model-based and model-free methods are known to achieve $\widetilde{\mathcal{O}}(H\sqrt{SAT})$ regret, where S is the number of states, and A is the number of actions. This bound is proven to be optimal up to logarithmic factors (Jin et al. 2018; Zhang, Zhou, and Ji 2020).

Extending beyond tabular MDPs, there have been an increasing number of works on function approximation with

provable guarantees (Jiang et al. 2017; Yang and Wang 2019, 2020; Jin et al. 2020; Zanette et al. 2020; Modi et al. 2020; Du et al. 2020; Cai et al. 2020; Ayoub et al. 2020; Wang, Salakhutdinov, and Yang 2020; Weisz, Amortila, and Szepesvári 2021; He, Zhou, and Gu 2021; Zhou, Gu, and Szepesvari 2021; Zhou, He, and Gu 2021; Ishfaq et al. 2021). For regret minimization in RL with linear function approximation, (Jin et al. 2020) assume that the transition model and the reward function of the MDPs are linear functions of a d-dimensional feature mapping and propose an optimistic variant of the Least-squares Value Iteration (LSVI) algorithm (Bradtke and Barto 1996; Osband, Van Roy, and Wen 2016) with $\widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T})$ regret. (Zanette et al. 2020) propose a randomized LSVI algorithm where exploration is induced by perturbing the least-squares approximation of the action-value function and provide $\widetilde{\mathcal{O}}(d^2H^2\sqrt{T})$ regret. For model-based methods with function approximation, (Yang and Wang 2020) assume the transition probability kernel to be a bilinear model parametrized by a matrix and propose a model-based algorithm with $\widetilde{\mathcal{O}}(d^{3/2}H^2\sqrt{T})$ regret. (Jia et al. 2020) consider a special class of MDPs called linear mixture MDPs where the transition probability kernel is a linear mixture of a number of basis kernels, which covers various classes of MDPs studied in previous works (Modi et al. 2020; Yang and Wang 2020). For this model, (Jia et al. 2020) propose a UCB-based RL algorithm with value-targeted model parameter estimation with $\widetilde{\mathcal{O}}(dH^{3/2}\sqrt{T})$ regret. The same linear mixture MDPs has been also used by (Ayoub et al. 2020; Zhou, Gu, and Szepesvari 2021; Zhou, He, and Gu 2021). In particular, (Zhou, Gu, and Szepesvari 2021) propose a variant of the method proposed by (Jia et al. 2020) and prove $\widetilde{\mathcal{O}}(dH\sqrt{T})$ regret with a matching lower bound $\Omega(dH\sqrt{T})$ for linear mixture MDPs. Extending function approximation beyond linear models, (Ayoub et al. 2020; Wang, Salakhutdinov, and Yang 2020; Ishfaq et al. 2021) also prove regret bounds depending on Eluder dimension (Russo and Van Roy 2013). There has been also some literature that aim to propose sample-efficient methods with more "general" function approximation. Yet, such claims may have been hindered by computational intractability (Krishnamurthy, Agarwal, and Langford 2016; Jiang et al. 2017; Dann et al. 2018) or having to rely on other stronger assumptions (Du et al. 2019), such that the resulting methods may turn out to be not as general or practical.

Despite the fact that there are a vast number of the existing works on RL with linear function approximation, there is very little work that extend beyond the linear model to other parametric models. To our best knowledge, (Wang et al. 2021) is the only existing work with generalized linear function approximation where the Bellman backup of any value function is assumed to be generalized linear function of feature mapping. (Wang et al. 2021) proposes a model-free algorithm under this assumption with $\widetilde{\mathcal{O}}(d^{3/2}H\sqrt{T})$ regret. In addition to the fact that their proposed method is model-free, the significant difference between the problem setting of our work and that of (Wang et al. 2021) is that the transition probability in (Wang et al. 2021) is *not* a generalized linear model, but rather generalized linear approximation is imposed directly on the value function update. Thus, the question of whether it is possible to design a provably efficient RL algorithm for an MDP with transition probability approximated by any generalized linear model including multinomial logistic model has still remained open.

Preliminaries

Notations

We denote by $[n]$ the set $\{1, 2, \ldots, n\}$ for a positive integer n. For a d-dimensional vector $x \in \mathbb{R}^d$, we use $\|x\|_2$ to denote the Euclidean norm of x. The weighted ℓ_2-norm associated with a positive definite matrix A is denoted by $\|x\|_A := \sqrt{x^\top Ax}$. The minimum and maximum eigenvalues of a symmetric matrix A are written as $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ respectively. The trace of a matrix A is $\mathrm{tr}(A)$. For two symmetric matrices A and B of the same dimensions, $A \succcurlyeq B$ means that $A - B$ is positive semi-definite.

Problem Formulation

We consider episodic Markov decision processes (MDPs) denoted by $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, P, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the length of horizon, P is the collection of transition probability distributions, and r is a reward function. Every episode starts at some initial state s_1 and ends after H steps. Then for every step $h \in [H]$ in an episode, the learning agent interacts with an environment defined by $\mathcal{M}$, where the agent observes state $s_h \in \mathcal{S}$, selects an action $a_h \in \mathcal{A}$, and receives an immediate reward $r(s_h, a_h) \in [0, 1]$. And then, the next state s_{h+1} is drawn from the transition probability distribution $P(\cdot \mid s_h, a_h)$ and repeats its interactions until the end of the episode, followed by a newly started episode. A policy $\pi : \mathcal{S} \times [H] \to \mathcal{A}$ is a function that determines which action the agent takes in state s_h at each step $h \in [H]$, $a_h \sim \pi(s_h, h)$. Then, we define the value function of policy π, $V_h^\pi : \mathcal{S} \to \mathbb{R}$ as the expected sum of rewards under the policy π until the end of the episode when starting from $s_h = s$, i.e.,

$$V_h^\pi(s) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r\left(s_{h'}, \pi(s_{h'}, h')\right) \mid s_h = s \right].$$

We define the action-value function of policy π, $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as the expected sum of rewards when following π starting from step h until the end of the episode after taking action a in state s,

$$Q_h^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r\left(s_{h'}, \pi(s_{h'}, h')\right) \mid s_h = s, a_h = a \right].$$

We define an optimal policy π^* to be a policy that achieves the highest possible value at every state-step pair $(s, h) \in \mathcal{S} \times [H]$. We denote by $V_h^*(s) = V_h^{\pi^*}(s)$ and $Q_h^*(s, a) = Q_h^{\pi^*}(s, a)$ as the the optimal value function and the optimal action-value function, respectively. To make notation simpler, we denote $P_h V_{h+1}(s, a) := \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V_{h+1}(s')]$.

Recall that both Q^π and Q^* can be written as the result of the Bellman equations as

$$Q_h^\pi(s,a) = (r + P_h V_{h+1}^\pi)(s,a)\,,$$
$$Q_h^*(s,a) = (r + P_h V_{h+1}^*)(s,a)$$

where $V_{H+1}^\pi(s) = V_{H+1}^*(s) = 0$ and $V_h^*(s) = \max_{a\in\mathcal{A}} Q_h^*(s,a)$ for all $s \in \mathcal{S}$.

The goal of the agent is to maximize the sum of future rewards, i.e., to find an optimal policy, through the repeated interactions with the environment for K episodes. Let policy $\pi = \{\pi_k\}_{k=1}^K$ be a collection of policies over K episodes, where π_k is the policy of the agent at k-th episode. Then, the cumulative regret of π over K episodes is defined as

$$\textbf{Regret}_\pi(K) := \sum_{k=1}^{K} (V_1^* - V_1^{\pi_k})(s_{k,1})$$

where $s_{k,1}$ is the initial state in the k-th episode. Therefore, maximizing the cumulative rewards of policy π over K episodes is equivalent to minimizing the cumulative regret $\textbf{Regret}_\pi(K)$.

In this paper, we make a structural assumption for the MDP $\mathcal{M}(\mathcal{S},\mathcal{A},H,P,r)$ where the transition probability kernel is given by an MNL model. Before we formally introduce the MNL function approximation, we first introduce the following definition.

Definition 1 (Reachable States)**.** *For each $(s,a) \in \mathcal{S}\times\mathcal{A}$, we define the "reachable states" of (s,a) to be the set of all states which can be reached by taking action a in state s within a single transition, $\mathcal{S}_{s,a} := \{s' \in \mathcal{S} : P(s' \mid s,a) \neq 0\}$. Also, we denote $\mathcal{U} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\mathcal{S}_{s,a}|$ to be the maximum size of reachable states.*

It is possible that even when the size of the state space $|\mathcal{S}|$ is very large, $\mathcal{U}$ is still small. For example, consider the RiverSwim problem (shown in Figure 2) with exponentially large state space. However, $\mathcal{U}$ would be still 3, regardless of the state space size.

Assumption 1 (MNL Transition Models)**.** *For each $(s,a) \in \mathcal{S}\times\mathcal{A}$, $s' \in \mathcal{S}$, let feature vector $\varphi(s,a,s') \in \mathbb{R}^d$ be given. Then we assume that the probability of state transition to $s' \in \mathcal{S}_{s,a}$ when an action a is taken at a state s is given by,*

$$P(s' \mid s,a) = \frac{\exp(\varphi(s,a,s')^\top \theta^*)}{\sum_{\widetilde{s}\in\mathcal{S}_{s,a}} \exp(\varphi(s,a,\widetilde{s})^\top \theta^*)} \qquad (1)$$

where $\theta^* \in \mathbb{R}^d$ is an unknown transition core parameter.

In order to focus on the main challenge of model-based RL, we assume, without loss of generality, that the reward function r is known for the sake of simplicity.[2] This assumption on r is standard in the model-based RL literature (Yang and Wang 2019, 2020; Zhou, Gu, and Szepesvari 2021).

[2]Note that this is assumed without loss of generality since learning r is much easier than learning P.

Linear Transition Model vs. Multinomial Logistic Transition Model

In this section, we show how the linear model assumption is restrictive for the transition model of MDPs. To our knowledge, this is the first rigorous discussion on the limitation of the linear transition model. We first show that for an arbitrary set of features, a linear transition model (including bilinear, linear MDPs, and linear mixture MDPs) cannot induce a proper probability distribution over next states.

Proposition 1. *For an arbitrary set of features of state and actions of an MDP, there exist no linear transition model that can induce a proper probability distribution over next states.*

Therefore, the linear model cannot be a proper choice of transition model in general. The restrictive linearity assumption on the transition model also affects the regret analysis of algorithms that are proposed under that assumption. As an example, we show that one of the recently proposed model-based algorithms using the linear function approximation cannot avoid the dependence on the size of the state space $|\mathcal{S}|$. (Yang and Wang 2020) assumes the transition probability kernel is given by the bilinear interaction of the state-action feature ϕ, the next state feature ψ, and the unknown transition core matrix M^*, i.e., $P(s'|s,a) = \phi(s,a)^\top M^* \psi(s')$. Before we show the suboptimal dependence, one can see that it is difficult to ensure that the estimated transition probability satisfies one of the fundamental probability properties, $\sum_{s'} \hat{P}(s' \mid s,a) = 1$ based on Proposition 1. In the following proposition, we show that the regret of the proposed algorithm in (Yang and Wang 2020) actually depends linearly on the size of the state space despite the use of function approximation.

Proposition 2. *The MatrixRL algorithm proposed in (Yang and Wang 2020) based on the linear model has the regret of $\widetilde{\mathcal{O}}(|\mathcal{S}|H^2 d^{3/2}\sqrt{T})$ where d is the dimension of the underlying parameter.*

Hence, the bilinear model-based method cannot scale well with the large state space. On the other hand, the MNL model defined in (1) can naturally capture the categorical distribution for any feature representation of states and actions and for any parameter choice. This is because, due to the normalization term of the MNL model, i.e., $\sum_{\widetilde{s}} \exp(\varphi(s,a,\widetilde{s})^\top\theta)$ — even if any estimated parameter for θ^* is used to estimate the transition probability, the sum of the transition probabilities is always 1. This holds not only for the true transition model but also for the estimated model. Hence, the MNL function approximation offers a more sensible model of the transition probability.

Algorithms and Main Results

Algorithm: `UCRL-MNL`

Estimation of Transition Core. Each transition sampled from the transition model provides information to the agent that updates the estimate for the transition core based on observed samples. For the sake of simple exposition, we assume discrete state space so that for all $k \in [K], h \in [H]$, we

Algorithm 1: Upper Confidence Model-based RL for MNL Transition Model (UCRL-MNL)

Inputs: An episodic MDP $\mathcal{M}$, Feature map $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, Total number of episodes K, Regularization parameter λ, Confidence radius β_k.
Initialize: $A_1 = \lambda I_d$, $\hat{\theta}_1 = \mathbf{0} \in \mathbb{R}^d$
for episode $k = 1, 2, \cdots, K$ **do**
  Set $\{\hat{Q}_{k,h}\}_{h=1}^H$ as described in (5) using $\hat{\theta}_k$, β_k
  for horizon $h = 1, 2, \cdots, H$ **do**
    Select an action $a_{k,h} = \operatorname{argmax}_{a \in \mathcal{A}} \hat{Q}_{k,h}(s_{k,h}, a)$ and observe $s_{k,h+1}$, $y_{k,h}$
  end for
  Update $A_{k+1} = A_k + \sum_{h \leqslant H} \sum_{s' \in \mathcal{S}_{k,h}} \varphi_{k,h,s'} \varphi_{k,h,s'}^\top$
  Compute $\hat{\theta}_{k+1} = \operatorname{argmax}_\theta \ell_{k+1}(\theta) - \frac{\lambda}{2}\|\theta\|_2^2$
end for

define the transition response variable $y_{k,h} = (y_{k,h}^{s'})_{s' \in \mathcal{S}_{k,h}}$ as $y_{k,h}^{s'} = \mathbb{1}(s_{k,h+1} = s')$ for $s' \in \mathcal{S}_{s_{k,h},a_{k,h}} =: \mathcal{S}_{k,h}$. Then the transition response variable $y_{k,h}$ is a sample from the following multinomial distribution:

$$y_{k,h} \sim \text{multinomial}\big(1, [p_{k,h}(s_{i_1}, \theta^*), \ldots, p_{k,h}(s_{i_{|\mathcal{S}_{k,h}|}}, \theta^*)]\big)$$

where the parameter 1 indicates that $y_{k,h}$ is a single-trial sample, and each probability is defined as

$$p_{k,h}(s', \theta^*) := \frac{\exp(\varphi(s_{k,h}, a_{k,h}, s')^\top \theta^*)}{\sum_{\widetilde{s} \in \mathcal{S}_{k,h}} \exp(\varphi(s_{k,h}, a_{k,h}, \widetilde{s})^\top \theta^*)} .$$

Also, we define noise $\epsilon_{k,h}^{s'} := y_{k,h}^{s'} - p_{k,h}(s', \theta^*)$. Since $\epsilon_{k,h}^{s'}$ is bounded in $[-1, 1]$, $\epsilon_{k,h}^{s'}$ is σ^2-sub-Gaussian with $\sigma^2 = 1$.

We estimate the unknown transition core θ^* using the regularized maximum likelihood estimation (MLE) for the MNL model. Based on the transition response variable $y_{k,h}$, the log-likelihood function under the parameter θ is then given by

$$\ell_k(\theta) = \sum_{\substack{k' < k \\ h \leqslant H}} \sum_{s' \in \mathcal{S}_{k',h}} y_{k',h}^{s'} \log p_{k',h}(s', \theta)$$

Then, the ridge penalized maximum likelihood estimation for the MNL model is given by the following optimization problem with the regularization parameter $\lambda \geqslant 0$:

$$\hat{\theta}_k = \operatorname*{argmax}_\theta \left[\ell_k(\theta) - \frac{\lambda}{2}\|\theta\|_2^2 \right] . \tag{2}$$

Model-Based Upper Confidence. To balance the exploration-exploitation trade-off, we construct an upper confidence action-value function which is greater than the optimal action-value function with high probability. The upper confidence bounds (UCB) approaches are widely used due to their effectiveness in balancing the exploration and exploitation trade-off not only in bandit problems (Auer 2002; Auer, Cesa-Bianchi, and Fischer 2002; Dani, Hayes, and Kakade 2008; Filippi et al. 2010; Abbasi-Yadkori, Pál, and Szepesvári 2011; Chu et al. 2011; Li, Lu, and Zhou 2017; Zhou, Li, and Gu 2020) but also in RL with function approximation (Wang et al. 2021; Jin et al. 2020; Ayoub et al. 2020; Jia et al. 2020).

At the k-th episode, the confidence set $\mathcal{C}_k$ for θ^* is constructed based on the feature vectors that have been collected so far. For previous episode $k' < k$ and the horizon step $h \leqslant H$, we denote the associated features by $\varphi(s_{k',h}, a_{k',h}, s') =: \varphi_{k',h,s'}$ for $s' \in \mathcal{S}_{s_{k',h},a_{k',h}}$ $(=: \mathcal{S}_{k',h})$. Then from all previous episodes, we have $\{\varphi_{k',h,s'} : s' \in \mathcal{S}_{k',h}, k' < k, h \leqslant H\}$ and the observed transition responses of $\{y_{k',h}\}_{k' < k, h \leqslant H}$. Let $\hat{\theta}_k$ be the estimate of the unknown transition core θ^* at the beginning of k-th episode, and suppose that we are guaranteed that θ^* lies within the confidence set $\mathcal{C}_k$ centered at $\hat{\theta}_k$ with radius $\beta_k > 0$ with high probability. Then for $(s, a) \in \mathcal{S} \times \mathcal{A}$ and for all $h \in [H]$, we construct the optimistic value function as follows:

$$\begin{aligned} &\hat{Q}_{k,H+1}(s, a) := 0 , \\ &\hat{Q}_{k,h}(s, a) := r(s, a) + \max_{\theta \in \mathcal{C}_k} \sum_{s' \in \mathcal{S}_{s,a}} p_{s,a}(s', \theta) \hat{V}_{k,h+1}(s') , \end{aligned} \tag{3}$$

where $p_{s,a}(s', \theta) = \frac{\exp(\varphi(s,a,s')^\top \theta)}{\sum_{\widetilde{s} \in \mathcal{S}_{k,h}} \exp(\varphi(s,a,\widetilde{s})^\top \theta)}$ and $\hat{V}_{k,h}(s) := \min\big\{ \max_a \hat{Q}_{k,h}(s, a), H \big\}$. Also, the confidence set $\mathcal{C}_k$ for θ^* is constructed as

$$\mathcal{C}_k := \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_k\|_{A_k} \leqslant \beta_k\}$$

where radius β_k is specified later, and the gram matrix A_k is given for some $\lambda > 0$ by

$$A_k = \lambda I_d + \sum_{\substack{k' < k \\ h \leqslant H}} \sum_{s' \in \mathcal{S}_{k',h}} \varphi_{k',h,s'} \varphi_{k',h,s'}^\top . \tag{4}$$

As long as the true transition core $\theta^* \in \mathcal{C}_k$ with high probability, the action-value estimates defined as (3) are optimistic estimates of the actual Q values. Based on these action-values $\{\hat{Q}_{k,h}\}_{h=1}^H$, in each $h \in [H]$ time step of the k-th episode, the agent selects the optimistic action $a_{k,h} = \operatorname{argmax}_{a \in \mathcal{A}} \hat{Q}_{k,h}(s_{k,h}, a)$. The full algorithm is summarized in Algorithm 1.

Closed-Form UCB. When we construct the optimistic value function as described in Eq.(3), it is required to solve a maximization problem over a confidence set. However, an explicit solution of this maximization problem is not necessary. The algorithm only requires the estimated action-value function to be optimistic. We can use a closed-form confidence bound instead of computing the maximal θ over the confidence set. Also, we can verify that the regret bound in Theorem 1 still holds even when we replace Eq.(3) with the following equation: for all $h \in [H]$,

$$\begin{aligned} \hat{Q}_{k,h}(s, a) = r(s, a) &+ \sum_{s' \in \mathcal{S}_{s,a}} p_{s,a}(s', \hat{\theta}) \hat{V}_{k,h+1}(s') \\ &+ 2H\beta_k \max_{s' \in \mathcal{S}_{s,a}} \|\varphi(s,a,s')\|_{A_k^{-1}} . \end{aligned} \tag{5}$$

Regret Bound for UCRL-MNL Algorithm

In this section, we present the regret upper-bound of UCRL-MNL. We first start by introducing the standard regularity assumptions.

Assumption 2 (Feature and Parameter)**.** *For some positive constants $L_\varphi, L_\theta > 0$, we assume that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}_{s,a}$, $\|\varphi(s,a,s')\|_2 \leqslant L_\varphi$. Also, $\|\theta^*\|_2 \leqslant L_\theta$.*

Note that this assumption is used to make the regret bounds scale-free for convenience and is in fact standard in the literature of RL with function approximation (Jin et al. 2020; Yang and Wang 2020; Zanette et al. 2020).

Assumption 3. *There exists $0 < \kappa < 1$ such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $s', s'' \in \mathcal{S}_{s,a}$ and for all $k \in [K], h \in [H]$, $\inf_{\theta \in \mathbb{R}^d} p_{k,h}(s',\theta) p_{k,h}(s'',\theta) \geqslant \kappa$.*

Assumption 3 is equivalent to a standard assumption in generalized linear contextual bandit literature (Filippi et al. 2010; Désir, Goyal, and Zhang 2014; Li, Lu, and Zhou 2017; Oh and Iyengar 2019; Kveton et al. 2020b; Russac, Cappé, and Garivier 2020; Oh and Iyengar 2021) to guarantee the Fisher information matrix is non-singular and is modified to suit our setting.

Under these regularity conditions, we state the regret bound for Algorithm 1.

Theorem 1 (Regret bound of UCRL-MNL)**.** *Suppose that Assumptions 1-3 hold. For $\delta \in (0,1)$, if $\lambda \geqslant L_\varphi^2$ and $\beta_k(\delta) = \frac{1}{\kappa}\sqrt{d \log\left(1 + \frac{kH\mathcal{U}}{d\lambda}\right) + 2\log\frac{1}{\delta}} + \frac{1}{\kappa}\sqrt{\lambda} L_\theta$, then with probability at least $1-\delta$, the cumulative regret of the UCRL-MNL policy π is upper-bounded by*

$$\textbf{\textit{Regret}}_\pi(K) = \widetilde{\mathcal{O}}\left(\kappa^{-1} d\sqrt{H^3 T}\right) .$$

Discussion of Theorem 1. In terms of the key problem primitives, Theorem 1 states that UCRL-MNL achieves $\widetilde{\mathcal{O}}(d\sqrt{H^3T})$ regret. To our best knowledge, this is the first result to guarantee a regret bound for the MNL model of the transition probability kernel. Among the existing model-based methods with function approximation, the most related method to ours is a bilinear matrix-based algorithm in (Yang and Wang 2020). (Yang and Wang 2020) shows $\widetilde{\mathcal{O}}(d^{3/2}H^2\sqrt{T})$ regret under the assumption that the transition probability can be a linear model parametrized with an unknown transition core matrix. Hence, the regret bound in Theorem 1 is sharper in terms of dimensionality and the episode length. Furthermore, as mentioned in Proposition 1, the regret bound in (Yang and Wang 2020) contains additional $|\mathcal{S}|$ dependence. Therefore, our regret bound shows an improved scalability over the method developed under a similar model. On the other hand, for linear mixture MDPs (Jia et al. 2020; Ayoub et al. 2020; Zhou, Gu, and Szepesvari 2021), the lower bound of $\Omega(dH\sqrt{T})$ has been proven in (Zhou, Gu, and Szepesvari 2021). Hence, noting the total dimension of the unknown parameter, the dependence on dimensionality as well as dependence on total steps matches the corresponding dependence in the regret bound for linear MDP (Zhou, Gu, and Szepesvari 2021). This provides a conjecture that the dependence on d and T in Theorem 1 is best possible although a precise lower bound in our problem setting has not yet been shown.

Proof Sketch and Key Lemmas

In this section, we provide the proof sketch of the regret bound in Theorem 1 and the key lemmas for the regret analysis. In the following lemma, we show that the estimated transition core $\hat{\theta}_k$ concentrates around the unknown transition core θ^* with high probability.

Lemma 1 (Concentration of the transition core)**.** *Suppose that Assumptions 1-3 hold. For given $\delta \in (0,1)$, let radius $\beta_k(\delta) = \frac{1}{\kappa}\sqrt{d\log\left(1+\frac{kH\mathcal{U}}{d\lambda}\right) + 2\log\frac{1}{\delta}} + \frac{1}{\kappa}\sqrt{\lambda} L_\theta$. Suppose $\hat{\theta}_k$ is the solution to the regularized MLE in Eq.(2) at the k-th episode. Then with probability at least $1-\delta$, the true transition core θ^* lies in the confidence set*

$$\mathcal{C}_k = \left\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_k\|_{A_k} \leqslant \beta_k(\delta)\right\} .$$

Then, we show that when our estimated parameter $\hat{\theta}_k$ is concentrated around the transition core θ^*, the estimated upper confidence action-value function is deterministically greater than the true optimal action-value function. That is, the estimated $\hat{Q}_{k,h}(s,a)$ is optimistic.

Lemma 2 (Optimism)**.** *Suppose that Lemma 1 holds for all $k \in [K]$. Then for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$, we have*

$$Q_h^*(s,a) \leqslant \hat{Q}_{k,h}(s,a) .$$

This optimism guarantee is crucial because it allows us to work with the estimated value function which is under our control rather than working with the unknown optimal value function. Next, we show that the value iteration per step is bounded.

Lemma 3 (Concentration of the value function)**.** *Suppose that Lemma 1 holds for all $k \in [K]$. For $\delta \in (0,1)$ and for any $h \in [H]$, we have*

$$\hat{Q}_{k,h}(s_{k,h}, a_{k,h}) - \left[r(s_{k,h}, a_{k,h}) + P_h \hat{V}_{k,h+1}(s_{k,h}, a_{k,h})\right] \leqslant 2H\beta_k \max_{s' \in \mathcal{S}_{k,h}} \|\varphi_{k,h,s'}\|_{A_k^{-1}} .$$

With these lemmas at hand, by summing per-episode regrets over all episodes and by the optimism of the estimated value function, the regret can be bounded by the sum of confidence bounds on the sample paths. All of the detailed proofs are included in the appendix.

Numerical Experiments

In this section, we evaluate the performances of our proposed algorithm, UCRL-MNL in numerical experiments.

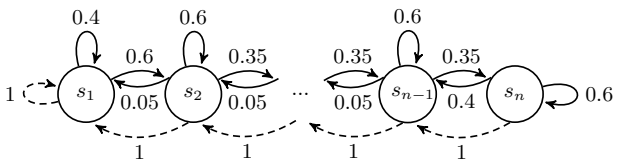

Figure 2: The "RiverSwim" environment with n states

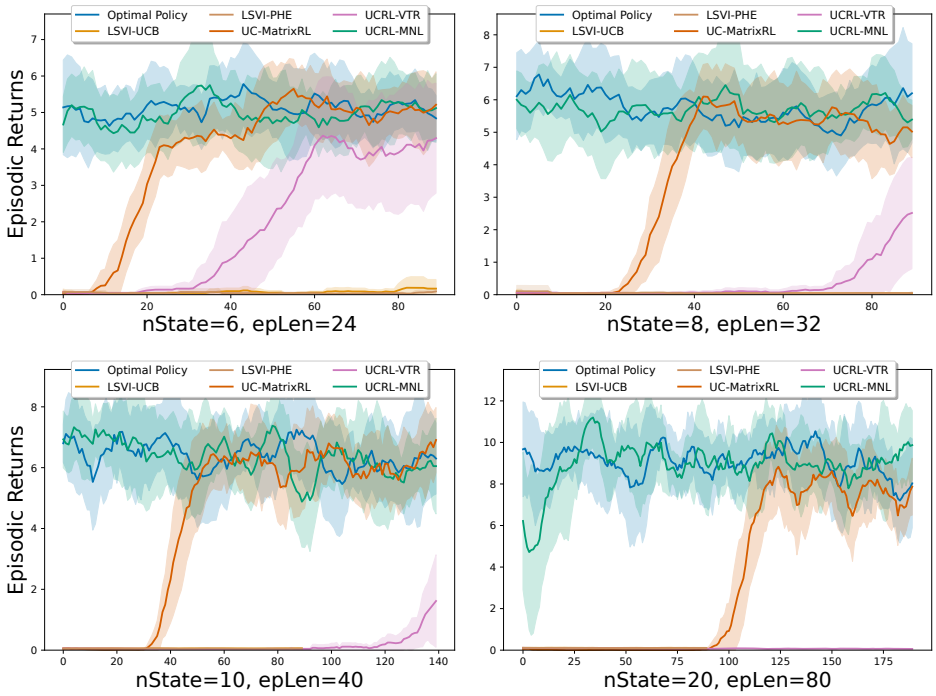

Figure 1: Episodic returns over 10 independent runs under the different RiverSwim environments

The RiverSwim environment (Osband, Russo, and Van Roy 2013) is considered to be a challenging problem setting where naïve dithering approaches, such as the ϵ-greedy policy, are known to have poor performance and require efficient exploration. The RiverSwim environment consists of n states (i.e., $|\mathcal{S}| = n$) lined up in a chain, with the number on each of the edges representing the transition probability. Starting in the leftmost state s_1, the agent can choose to swim to the left — whose outcomes are represented by the dashed lines — and collect a small reward. i.e., $r(s_1, \texttt{left}) = 0.05$. Or, the agent can choose to swim to the right — whose outcomes are represented by the *solid* lines — in each succeeding state. The agent's goal is to maximize its return by attempting to reach the rightmost state s_n where a large reward $r(s_n, \texttt{right}) = 1$ can be obtained by swimming right.

Since the objective of this experiment is to see how efficiently our algorithm explores compared to other provably efficient RL algorithms with function approximation, we choose both model-based algorithms, `UC-MatrixRL` (Yang and Wang 2020) and `UCRL-VTR` (Ayoub et al. 2020) and model-free algorithms, `LSVI-UCB` (Jin et al. 2020) and `LSVI-PHE` (Ishfaq et al. 2021) for comparisons.

We perform a total of four experiments while increasing the number of states for RiverSwim. To set the hyperparameters for each algorithm, we performed a grid search over certain ranges. In each experiment, we evaluated the algorithms on 10 independent instances to report the average performance. First, Figure 1 shows the episodic return of each algorithm over 10 independent runs. When the number of states is small (e.g., $|\mathcal{S}| = 6$), it can be seen that not only our algorithm but also other model-based algorithms learn the optimal policy relatively well. However, our algorithm `UCRL-MNL` clearly outperforms the existing algorithms. As the number of states increases (e.g., $|\mathcal{S}| = 20$), we observe that our algorithm reaches the optimal values remarkably quickly compared to the other algorithms, outperforming the existing algorithms by significant margins.

Methods	$\lvert\mathcal{S}\rvert = 6, H = 24$	$\lvert\mathcal{S}\rvert = 20, H = 80$
Optimal Policy	513.70 ± 39.67	1816.50 ± 79.89
`UCRL-MNL`	$\mathbf{492.64 \pm 32.23}$	$\mathbf{1777.11 \pm 99.23}$
`UC-MatrixRL`	372.47 ± 18.04	682.15 ± 64.41
`UCRL-VTR`	191.72 ± 68.47	15.76 ± 0.848
`LSVI-UCB`	7.52 ± 5.284	10.97 ± 1.813
`LSVI-PHE`	5.15 ± 1.506	10.17 ± 0.139

Table 1: Average returns over 10 independent runs in RiverSwim environments

Table 1 shows the average cumulative reward over the episodes of each algorithm for 10 independent runs. The proposed algorithm has an average cumulative reward similar to that of the optimal policy across all problem settings. The results of these experiments provide evidence for the practicality of our proposed model and proposed algorithm.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1C1C1006859) and by Creative-

Pioneering Researchers Program through Seoul National University and by Naver.

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24: 2312–2320.

Abeille, M.; and Lazaric, A. 2017. Linear Thompson Sampling Revisited. In Singh, A.; and Zhu, J., eds., *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, 176–184. PMLR.

Agrawal, S.; and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 127–135. PMLR.

Agrawal, S.; and Jia, R. 2017. Posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, 1184–1194.

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

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2): 235–256.

Ayoub, A.; Jia, Z.; Szepesvari, C.; Wang, M.; and Yang, L. 2020. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, 463–474. PMLR.

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

Bradtke, S. J.; and Barto, A. G. 1996. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1): 33–57.

Cai, Q.; Yang, Z.; Jin, C.; and Wang, Z. 2020. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, 1283–1294. PMLR.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 208–214. JMLR Workshop and Conference Proceedings.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic Linear Optimization under Bandit Feedback. In Servedio, R. A.; and Zhang, T., eds., *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, 355–366. Omnipress.

Dann, C.; Jiang, N.; Krishnamurthy, A.; Agarwal, A.; Langford, J.; and Schapire, R. E. 2018. On Oracle-Efficient PAC RL with Rich Observations. In *Advances in Neural Information Processing Systems*, volume 31.

Dann, C.; Lattimore, T.; and Brunskill, E. 2017. Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 30, 5713–5723.

Désir, A.; Goyal, V.; and Zhang, J. 2014. Near-optimal algorithms for capacity constrained assortment optimization. *Available at SSRN*, 2543309.

Du, S.; Krishnamurthy, A.; Jiang, N.; Agarwal, A.; Dudik, M.; and Langford, J. 2019. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, 1665–1674. PMLR.

Du, S. S.; Kakade, S. M.; Wang, R.; and Yang, L. F. 2020. Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning? In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Filippi, S.; Cappé, O.; Garivier, A.; and Szepesvári, C. 2010. Parametric Bandits: The Generalized Linear Case. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, 586–594. Red Hook, NY, USA: Curran Associates Inc.

He, J.; Zhou, D.; and Gu, Q. 2021. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, 4171–4180. PMLR.

Ishfaq, H.; Cui, Q.; Nguyen, V.; Ayoub, A.; Yang, Z.; Wang, Z.; Precup, D.; and Yang, L. 2021. Randomized Exploration in Reinforcement Learning with General Value Function Approximation. In *International Conference on Machine Learning*, volume 139, 4607–4616. PMLR.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(4).

Jia, Z.; Yang, L.; Szepesvari, C.; and Wang, M. 2020. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, 666–686. PMLR.

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

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018. Is Q-Learning Provably Efficient? In *Advances in Neural Information Processing Systems*, volume 31, 4868–4878.

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

Jun, K.-S.; Bhargava, A.; Nowak, R.; and Willett, R. 2017. Scalable Generalized Linear Bandits: Online Computation and Hashing. In *Advances in Neural Information Processing Systems*, volume 30.

Krishnamurthy, A.; Agarwal, A.; and Langford, J. 2016. PAC Reinforcement Learning with Rich Observations. *Advances in Neural Information Processing Systems*, 29: 1840–1848.

Kveton, B.; Szepesvári, C.; Ghavamzadeh, M.; and Boutilier, C. 2020a. Perturbed-History Exploration in Stochastic Linear Bandits. In *Uncertainty in Artificial Intelligence*, 530–540. PMLR.

Kveton, B.; Zaheer, M.; Szepesvari, C.; Li, L.; Ghavamzadeh, M.; and Boutilier, C. 2020b. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, 2066–2076. PMLR.

Li, L.; Lu, Y.; and Zhou, D. 2017. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, 2071–2080. PMLR.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.

Modi, A.; Jiang, N.; Tewari, A.; and Singh, S. 2020. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, 2010–2020. PMLR.

Oh, M.-h.; and Iyengar, G. 2019. Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, 32.

Oh, M.-h.; and Iyengar, G. 2021. Multinomial logit contextual bandits: Provable optimality and practicality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9205–9213.

Osband, I.; and Roy, B. V. 2014. Model-based Reinforcement Learning and the Eluder Dimension. In *Advances in Neural Information Processing Systems*, 1466–1474.

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

Osband, I.; Van Roy, B.; Russo, D. J.; Wen, Z.; et al. 2019. Deep Exploration via Randomized Value Functions. *Journal of Machine Learning Research*, 20(124): 1–62.

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

Ouyang, Y.; Gagrani, M.; Nayyar, A.; and Jain, R. 2017. Learning Unknown Markov Decision Processes: A Thompson Sampling Approach. In *Advances in Neural Information Processing Systems*, 1333–1342.

Russac, Y.; Cappé, O.; and Garivier, A. 2020. Algorithms for non-stationary generalized linear bandits. *arXiv preprint arXiv:2003.10113*.

Russo, D. 2019. Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32.

Russo, D.; and Van Roy, B. 2013. Eluder Dimension and the Sample Complexity of Optimistic Exploration. In *Advances in Neural Information Processing Systems*, 2256–2264.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419): 1140–1144.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676): 354–359.

Wang, R.; Salakhutdinov, R. R.; and Yang, L. 2020. Reinforcement Learning with General Value Function Approximation: Provably Efficient Approach via Bounded Eluder Dimension. *Advances in Neural Information Processing Systems*, 33.

Wang, Y.; Wang, R.; Du, S. S.; and Krishnamurthy, A. 2021. Optimism in Reinforcement Learning with Generalized Linear Function Approximation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Weisz, G.; Amortila, P.; and Szepesvári, C. 2021. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, 1237–1264. PMLR.

Yang, L.; and Wang, M. 2019. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, 6995–7004. PMLR.

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

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

Zhang, Z.; Zhou, Y.; and Ji, X. 2020. Almost Optimal Model-Free Reinforcement Learningvia Reference-Advantage Decomposition. In *Advances in Neural Information Processing Systems*, volume 33, 15198–15207.

Zhang, Z.; Zhou, Y.; and Ji, X. 2021. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, 12653–12662. PMLR.

Zhou, D.; Gu, Q.; and Szepesvari, C. 2021. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, 4532–4576. PMLR.

Zhou, D.; He, J.; and Gu, Q. 2021. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, 12793–12802. PMLR.

Zhou, D.; Li, L.; and Gu, Q. 2020. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, 11492–11502. PMLR.