Value Iteration-based Provably Efficient Exploration

Viet Nguyen
McGill University, Mila
Montreal, QC
viet.nguyen@mila.quebec

Eric Hu
McGill University
Montreal, QC
eric.hu@mail.mcgill.ca

1 Introduction

1.1 A brief history

Reinforcement learning has long piqued the interest of many due to its vast applicability and its highly promising potential. With foundational ideas stemming from the traditional theory of artificial intelligence [18], it's been undergoing continuous development since the later decades of the 20th century [24]. When the deep learning revolution occurred in the early 2010s with cornerstone works by Krizhevsky et. al. [14], Srivastava et. al. [23], and later on Vaswani et. al. [26], many employed deep neural networks in reinforcement learning to achieve superhuman level game-playing [22, 16], as well as a multitude of applications in other fields [3, 29, 6]. Thus was born the sub-field of deep reinforcement learning, and since then a non-small number crossed the event horizon and spiraled down a dark abyssal alley...

Theorists found themselves with more questions than answers. Now that deep reinforcement learning has constantly been proving itself with solid empirical performances on tasks that even challenge humans, the general theory of learning began rapid development, in particular, the development of frameworks for *exploration* in reinforcement learning that guarantees with high probability that deep RL agents explore efficiently. Recently, by imposing various assumptions, Wang et. al. [28] proposed a provably efficient algorithm in the general function approximation (GFA) setting, a setting that covers a restriction of the space of affine compositions of neural networks. In this work, as we take a deeper look into several important components of Wang et. al.'s work, we will later argue that although they spearhead recent efforts towards a theory for deep reinforcement learning, the strong assumptions that make up the heart of their proofs hint at the need for a different/alternate fundamental understanding of the reinforcement learning problem.

1.2 From *tabula rasa* to linear MDPs

From here onwards, we understand $\mathcal{S}$ to be a state space, $\mathcal{A}$ to be an action space, H to be the horizon of an episode, and $T = KH$ where K is the number of episodes played, unless otherwise stated. Furthermore, P is a transition operator, and r is a reward function. Jaksch et. al. [10] proposed in 2010 the UCLR2 algorithm, achieving a regret of $\tilde{O}(H\mathcal{S}\sqrt{\mathcal{A}T})$ with the finite state-action space assumption. After every episode, UCLR2 updates its empirical MDP, computes confidence sets for its transition models and reward models, and selects an optimistic MDP as well as an optimistic policy to follow. In 2017, Azar et. al. [5] introduces an optimistic modification of least-squares value iteration (LSVI), incurring an improved regret of $\tilde{O}(H\sqrt{\mathcal{SA}T} + H^2\mathcal{S}^2\mathcal{A})$, nearing the best known bound in the *tabula rasa* setting of $\Omega(\sqrt{H\mathcal{SA}T})$. This quickly becomes a problem since many problems in reinforcement learning are set in a continuous state space, or even a continuous action space, examples of such are robot arm control and online navigation for self-driving vehicles. A popular approach is binning, i.e. grouping the continuum into bins and treating such bins as individual states, however this hands-on approach presumes knowledge of the underlying environment, thus unfit for many learning tasks where we do not have this luxury. A classical approach would be to use linear function approximation [24, 25].

In 2019, Jin et. al. published their foundational work [12] on a function approximation OFU approach when the underlying MDP is linear (or approximately linear), that is, they assume the existence of a feature map $\phi : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbf{R}^d$ such that for any $h \in H$, there exists d unknown signed measures $\mu_h = \left(\mu_h^{(1)}, \ldots, \mu_h^{(d)}\right)$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbf{R}^d$ such that $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, we have that

$$P_h(\cdot|s, a) = \langle \phi(s, a), \mu_h(\cdot) \rangle , \qquad r_h(s, a) = \langle \phi(s, a), \theta_h \rangle$$

and w.l.o.g. $\|\phi(s, a)\| \leq 1$ for all s, a, as well as $\|\mu_h(\mathcal{S})\| , \|\theta_h\| \leq \sqrt{d}$ for all $h \in [H]$. They achieved regret bounds of $\tilde{O}(d^{3/2}H^{3/2}\sqrt{T})$. We remark that here, we let P and r vary according to the per-episode timestep h, that is, a non-stationary model. However, these same regret bounds apply in the stationary case as well, as the latter is a trivial modification of the former.

Subsequent works by Cai et. al. [7] and Ishfaq et. al. [9] present exponential gradient updates-inspired approaches to a slightly modified linear kernel MDP setting, achieving similar bounds. The common denominator between these works is that when these linear assumptions are in place, the value function and the quality function (state-action value function) is also linear in the feature map ϕ, this readily follows by definition. Thus, one can interpret the algorithm as a search in the space of linear functions $\mathcal{F} = \left\{f : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbf{R}, f(s, a) = \langle \phi(s, a), w \rangle , w \in \mathbf{R}^d\right\}$ to find the quality function that best correspond to the underlying MDP.

However, this setting is still quite restrictive. Even if there are methods that incorporate a misspeficiation error (i.e. imposing the linear MDP assumption on MDPs that are not actually linear), the regret bound depends linearly on the amount of misspecification (formally defined in [11]), and thus would prove to be quite problematic in a lot of situations where the linearity assumption is poor.

Wang et. al.'s work tackles this challenge by proposing a working approach in a much broader setting. They leverage current understandings of the eluder dimension [20] to control the complexity of the function class

1.3 Notation, general function approximation (GFA)

This subsection establishes the notation and the assumptions that will be used throughout this paper. As previously stated, let $(\mathcal{S}, \mathcal{A}, P, r, H)$ be an MDP such that $|\mathcal{A}| < \infty$, $P : \mathcal{S} \times \mathcal{A} \longrightarrow \Delta(\mathcal{S})$ taking a state-action pair to a distribution over states, and $r : \mathcal{S} \times \mathcal{A} \longrightarrow [0, 1]$. Assume also that there is some underlying initial distribution $\mu \in \Delta(\mathcal{S})$.

A (deterministic) policy π chooses an action a based on the current state and current timestep, that is, $\pi = \{\pi_h : h \in [H]\}$ such that $\pi_h : \mathcal{S} \longrightarrow \mathcal{A}$.

Given a policy π, a timestep $h \in [H]$, and a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define the quality of the policy, or the Q-function, denoted by $Q^\pi = \{Q_h^\pi : h \in [H]\}$ as

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'} | s_h = s, a_h = a, \pi\right]$$

that is, the expected rewards from playing the policy until the end of the episode given that we are currently in state s and play action a. The value of a policy, V, is defined for any $s \in S$ and $h \in [H]$ as

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'} | s_h = s, \pi\right]$$

A policy maximixing the expected reward of an entire episode $\mathbb{E}\left[\sum_{h=1}^{H} r_h | \pi\right]$ is an optimal policy, and we denote it by π^*. The optimal policy has quality Q^* and value V^*. At the beginning of the k-th episode, the agent chooses a policy $\pi^k = \left\{\pi_h^k : h \in [H]\right\}$ to play, and uses the history of all trajectories induced by past policies to choose the next policy π^{k+1} until K episodes are played. The (frequentist) regret after K episodes from playing policy π^k instead of the optimal pi^* is defined as

$$\mathrm{Regret}(K) = \sum_{k=1}^{K} \left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right)$$

To establish the GFA setting and perform subsequent analyses, we employ the following notation. For $f : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbf{R}$ and $v : \mathcal{S} \longrightarrow \mathbf{R}$, define

$$\|f\|_\infty = \max_{(s,a)\in S\times A} |f(s,a)|, \qquad \|v\|_\infty = \max_{s\in S} |v(s)|$$

Given a dataset $\mathcal{D} = (s_i, a_i, q_i)_{i=1}^{|\mathcal{D}|} \subset \mathcal{S} \times \mathcal{A} \times \mathbf{R}$, define

$$\|f\|_{\mathcal{D}} = \left(\sum_{t=1}^{|\mathcal{D}|} (f(s_t, a_t) - q_t)^2 \right)^{1/2}$$

and for some $\mathcal{Z} \subset \mathcal{S} \times \mathcal{A}$, define

$$\|f\|_{\mathcal{Z}} = \left(\sum_{(s,a)\in\mathcal{Z}} f(s,a)^2 \right)^{1/2}$$

For a set of functions $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbf{R}\}$, given some $(s,a) \in \mathcal{S} \times \mathcal{A}$, the *width* of (s,a) is

$$w(\mathcal{F}, s, a) = \max_{f,f'\in\mathcal{F}} \{f(s,a) - f'(s,a)\}$$

Intuitively, the functional $\|\cdot\|_{\mathcal{Z}}$ can be used as a metric that determines the "difference" between two functions on a set of state-action pairs, and is at the center of the defintion of the eluder dimension.

Recall a remark made previously, linear MDP algorithms essentially search for a quality function in the space of linear functions $\mathcal{F} = \left\{f : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbf{R}, f(s,a) = \langle \phi(s,a), w \rangle, w \in \mathbf{R}^d\right\}$ that best matches that of the underlying MDP (once this quality function is acceptably approximated, a greedy policy will be near-optimal). The same logic applies here, except that we can now make the search space much richer than the space of linear functions. This is the main theme of the following assumption about search space $\mathcal{F}$.

Assumption 1.1. For any $V : \mathcal{S} \longrightarrow [0, H]$, there exists a $f_V \in \mathcal{F}$ such that $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$,

$$f_V(s,a) = r(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V(s')\right]$$

This intuitively means that for function $V : \mathcal{S} \longrightarrow [0, H]$, applying the Bellman backup operator results in a function which lies in $\mathcal{F}$. It can be shown that when $\mathcal{S}$ and $\mathcal{A}$ are finite, $\mathcal{F} = \{f : \mathcal{S} \times \mathcal{A} \longrightarrow [0, H+1]\}$ satisfies this property, while for linear MDPs, the space $\mathcal{F} = \left\{f : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbf{R}, f(s,a) = \langle \phi(s,a), w \rangle, w \in \mathbf{R}^d\right\}$ will also satisfy the assumption.

Indeed, this assumption does approximately hold for the case of affine compositions. That is, when the Q-function is approximated by a neural network, under several more constraints the assumption holds, and thus the complexity the RL problem, being searching over the space $\mathcal{F}$ that best represents the underlying Q-function, is dependent on the complexity of this search space. To characterize this complexity, the authors used the eluder dimension [20], a tool invented in the stochastic bandits setting, and later on found various applications in RL. For a more in-depth treatment of the eluder dimension, we refer the reader to [8], and will henceforth assume knowledge of it.

The next assumption relies on the important notion of covering numbers. When (X, d) is a metric space, an ϵ-net of X is a subset $E \subset X$ such that $\forall x \in X, \exists y \in E$ such that $d(x,y) \leq \epsilon$. The ϵ-covering number of X is

$$\mathcal{N}(X, \epsilon) = \inf\{|V| : V \text{ is an } \epsilon\text{-net}\}$$

Assumption 1.2. $\forall \epsilon > 0$ there exists an ϵ-cover $\mathcal{C}(\mathcal{F}, \epsilon) \subseteq \mathcal{F}$ with $|\mathcal{C}(\mathcal{F}, \epsilon)| \leq \mathcal{N}(\mathcal{F}, \epsilon)$ such that $\forall f \in \mathcal{F}, \exists f' \in \mathcal{C}(\mathcal{F}, \epsilon)$ with $\|f - f'\|_\infty \leq \epsilon$.

$\forall \epsilon > 0$ there exists an ϵ-cover $\mathcal{C}(\mathcal{S} \times \mathcal{A}, \epsilon) \subseteq \mathcal{S} \times \mathcal{A}$ with $|\mathcal{C}(\mathcal{S} \times \mathcal{A}, \epsilon)| \leq \mathcal{N}(\mathcal{S} \times \mathcal{A}, \epsilon)$ such that $\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \exists (s', a') \in \mathcal{C}(\mathcal{S} \times \mathcal{A}, \epsilon)$ with $\max_{f\in\mathcal{F}} |f(s,a) - f(s',a')| \leq \epsilon$.

This assumption essentially requires spaces $\mathcal{F}$ and the product space $\mathcal{S} \times \mathcal{A}$ to have bounded covering numbers. One does indeed remark here that these assumptions are slightly restrictive. For example, take $\mathcal{F}$ to be the space of all affine compositions with intermediate activation. For $\mathcal{F}$ to satisfy the second assumption would require applying certain bounds either the norms of the weights of the affine connections or the type of activation function. Otherwise, it can be readily seen that there exists no finite covers of $\mathcal{F}$.

Taking a step back, there are indeed quite restrictive assumptions, in particular, on the search space $\mathcal{F}$, being some control of its eluder dimension as well as its covering number. However, as we shall see in a later section, these are indeed the key pieces that make the algorithm provably efficient.

2 Algorithm, bonus stability

2.1 Design

The algorithm follows the upper-confidence bound (UCB) method, an implementation of the OFU paradigm. At the end of each episode, it computes a candidate for a Q-value function, and adds to it a bonus function. The latter scales in a way to incite exploration in directions yet unknown but deem interesting by the algorithm. The algorithm proceeds thus: at the beginning of episode $k \in [K]$, let the history until the $k-1$-th episode be $\{(s_h^\tau, a_h^\tau) : h \in [H], \tau \in [k-1]\}$, the algorithm sets $Q_{H+1}^k = 0$, and proceeds backwards as follows: for $h = H, \ldots, 1$,

$$f_h^k(\cdot,\cdot) \leftarrow \arg\min_{f \in \mathcal{F}} \sum_{\tau=1}^{k-1} \sum_{h'=1}^{H} \left(f(s_{h'}^\tau, a_{h'}^\tau) - \left(r_{h'}^\tau + \max_{a \in \mathcal{A}} Q_{h+1}^k(s_{h'+1}^\tau, a) \right) \right)^2$$

$$Q_h^k(\cdot,\cdot) \leftarrow \min \left\{ f_h^k(\cdot,\cdot) + b_h^k(\cdot,\cdot), H \right\}$$

where b_h^k is a bonus function for that specific episode and timestep. Then, a greedy policy w.r.t. Q_h^k is played for the next episode. This is equivalent to the following formulation. If we make the dataset $\mathcal{D}_h^k = \left\{ \left(s_{h'}^\tau, a_{h'}^\tau, r_{h'}^\tau + V_{h+1}^k(s_{h'+1}^\tau) \right) : (\tau, h') \in [k-1] \times [H] \right\}$, we can express the optimization objective as:

$$f_h^k \leftarrow \arg\min_{f \in \mathcal{F}} \|f\|_{\mathcal{D}_h^k}^2$$

In the linear MDP approach, generally UCB methods perform least-squares on the history to compute a solid guess at the linear parameter of the underlying Q-function. If after every episode, we define the per-timestep design matrix:

$$\Sigma_h^k = \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^T + \lambda I$$

then, the bonus scales depending on the inverse of the design matrix. For optimistic sampling approaches (topic of a later discussion), one would sample normal perturbations that scale with the inverse of the design matrix as well. This incites exploration in directions yet unknown as columns of the design matrix scale inversely with the number of times that state has been visited.

Recall that in the UCB paradigm, as a step-up from the classical extended value iteration introduced by Jaksch et. al. [10], Azar et. al. [5] introduced the bonus directly to the final computation of the Q-function, very much like the above. This bonus represents the upper-confidence bound for the estimate of the Q-function, that is, w.p.a.l. $1-\delta$ for small δ, the chosen Q-function is greater than the true Q-function for inputs $(s, a) \in \mathcal{S} \times \mathcal{A}$, and is derived using Chernoff-Hoeffding concentration techniques on the empirical MDP. Here, however, with the assumption that $\mathcal{S}$ is continuous, empirical MDPs would not work. The authors of [28] recycles a familiar construction from the literature to overcome this challenge.

Define the value function $V_{h+1}^k(\cdot) = \max_{a \in \mathcal{A}} Q_{h+1}^k(\cdot, a)$. Letting $\mathcal{Z}^k = \{(s_h^\tau, a_h^\tau) : h \in [H], \tau \in [k-1]\}$ denote the state-action history up until the episode $k-1$, define

$$\mathcal{F}_h^k = \left\{ f \in \mathcal{F} : \left\| f - f_h^k \right\|_{\mathcal{Z}^k}^2 \leq \beta \right\}$$

with a choice of β such that applying the Bellman backup operator to the value function V_{h+1}^k results in a Q-function remaining within $\mathcal{F}_h^k$ with high probability, that is,

$$r(\cdot,\cdot) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V_{h+1}^k(s') \right] \in \mathcal{F}_h^k$$

Now, $b_h^k(\cdot,\cdot) = w(\mathcal{F}_h^k,\cdot,\cdot)$ would make a good choice for bonus function, as the width function maximizes the difference between all pairs of Q-functions within the confidence region, and thus gives an upper bound on the confidence interval of the estimate of the Q-function.

As the authors remark, the complexity of this bonus function could be extremely high as the state-action history $\mathcal{Z}^k$ scales linearly with every episode. Therefore, instead of directly computing the width of $\mathcal{F}_h^k$, they designed their ingenious process of thinning the state-action history $\mathcal{Z}^k$ into $\hat{\mathcal{Z}}^k$ such that the width function w.r.t. $\hat{\mathcal{F}}_h^k = \left\{ f \in \mathcal{F} : \left\| f - f_h^k \right\|_{\hat{\mathcal{Z}}^k}^2 \leq \beta \right\}$ is approximately equal to the width function w.r.t. $\mathcal{F}_h^k$. In this way, the confidence region is approximately preserved, thus an application of the Bellman backup operator to the value function V_{h+1}^k would still lie within $\mathcal{F}_h^k$ with high probability.

The process of sub-sampling $\hat{\mathcal{Z}}^k$ from $\mathcal{Z}^k$ relies on a measure of sensitivity that reveals how much each state-action in the history contribute to $\|f - f'\|_{\mathcal{Z}^k}^2$, and trimming accordingly. The authors successfully bound $|\hat{\mathcal{Z}}^k|$ in terms of the eluder dimension of $\mathcal{F}$ times the log-covering nubmer of $\mathcal{F}$. However, this sensitivity-sampling process is not the topic of this paper's discussion, and we refer the reader to the original work for more related discussions.

2.2 Regret

The authors prove that under the assumptions 1.1. and 1.2., w.p.a.l., $1 - \delta$,

$$\operatorname{Regret}(K) \leq \sqrt{\iota \cdot H^2 \cdot T}$$

where

$$\iota \leq C \cdot \log^2 \frac{T}{\delta} \cdot \dim_E^2 \left(\mathcal{F}, \frac{\delta}{T^3} \right) \cdot \ln \left(\mathcal{N} \left(\mathcal{F}, \frac{\delta}{T^2} \right) / \delta \right) \cdot \log \left(\mathcal{N} \left(\mathcal{S} \times \mathcal{A}, \frac{\delta}{T} \right) \cdot T/\delta \right)$$

with $C > 0$. We notice here that the ι term depends on the eluder dimension of the search class, the log-covering number of the search class, as well as the log-covering number of the state-action space. This is indeed due to the sensitivity-sampling process, namely its upper-control on the size of the surrogate history $\hat{\mathcal{Z}}^k$. We also stress here that while we could explicitly write out eluder dimensions of the space of linear functions as well as their log-covering numbers and obtain a worse regret bound than those found in the analyses of [11, 7, 9], the regret bound is in its most general form and encompasses the tabular setting, the linear and linear kernel MDP settings, as well as the generalized linear function approximation settings.

3 A deeper look into the proof

This section is dedicated to the proof and analysis of key lemmas that make up the proof of the regret bound theorem of the algorithm proposed by Wang et. al.. In what follows, let $V_h^k(s) = \max_{a \in \mathcal{A}} Q_h^k(s,a)$ where $Q_h^k = f_h^k + b_h^k$ given by the algorithm, while V^{π^k} and Q^{π^k} are the value function and the Q-function on the underlying MDP of the greedy policy $\pi^k = \left\{ \pi_h^k : h \in [H] \right\}$, induced by $\left\{ Q_h^k : h \in [H] \right\}$ from the algorithm. One remarks that V^k and V^{π^k} are certainly not the same, as V^k serves as the algorithmic approximate for the true V^{π^k} on the underlying MDP. We begin with a regret decomposition and upper control.

Furthermore, to denote the integral $\mathbb{E}_{s' \sim P(\cdot|s,a)}[V]$, we sometimes use the notation $\mathbb{E}\left[V|s,a\right]$.

Lemma 3.1. (**Lemma 8** from Wang et. al.) W.p.a.l. $1 - \delta/2$,

$$\operatorname{Regret}(K) \leq 2 \sum_{k=1}^{K} \sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) + 4H\sqrt{KH \cdot \log(8/\delta)}$$

This lemma upper bounds the regret by the total sum of bonuses plus an additional trailing term. It remains to control the bonus sum. To this end, we use the fact that $\mathcal{F}$ has bounded eluder dimension, the focus of the next lemma:

Lemma 3.2. (**Lemma 9** from Wang et. al.) W.p.a.l. $1-\delta/4$, for any $\epsilon > 0$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{1}\left\{b_h^k(s_h^k, a_h^k) > \epsilon\right\} \le \left(\frac{c\beta(\mathcal{F},\delta)}{\epsilon^2}\right) \cdot \dim_E(\mathcal{F},\epsilon)$$

where $\beta = \beta(\mathcal{F},\delta)$ is chosen so that an application of the Bellman backup operator to Q_h^k stay in $\mathcal{F}_h^k$ with high probability.

Finally, applying this lemma gives us the following.

Lemma 3.3. (**Lemma 10** from Wang et. al.) W.p.a.l. $1-\delta/4$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \le 1 + 4H^2 \dim_E(\mathcal{F}, 1/T) + \sqrt{c \cdot \dim_E(\mathcal{F}, 1/T) \cdot T \cdot \beta(\mathcal{F},\delta)}$$

for some $c > 0$.

One remarks that applying this control on the regret bound in **Lemma 3.1.** would result in the the main regret bound of the algorithm.

Lemma 3.1. hinges on the following lemma which we are not going to prove, but refer the reader once again to the original authors.

Lemma 3.4. (**Lemma 7** from Wang et. al.) W.p.a.l. $1-\delta/4$, for all $(k,h) \in [K] \times [H]$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$Q_h^*(s,a) \le Q_h^k((s,a) \le r(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V_{h+1}^k(s')\right] + 2b_h^k(s,a)$$

This lemma essentially states that with high probability, the algorithm over-estimates the Q-function, and the amount of overestimation is controlled by the bonus function.

3.1 Proof of Lemma 3.1.

Proof. Let $\delta_h^k := V_h^k - V_h^{\pi^k}$, define $\xi_h^k = \mathbb{E}\left[\delta_{h+1}^k | s_h^k, a_h^k\right] - \delta_{h+1}^k := \mathbb{E}_{s'\sim P(\cdot|s_h^k,a_h^k)}\left[\delta_{h+1}^k\right] - \delta_{h+1}^k$. Since the computation of V_h^k is independent of the observation s_h^k at episode k, as it only depends on the previous $k-1$ episodes. Therefore, if we let the filtration induced by the history up until episode k, timestep h to be $\mathbb{F}_h^k$, we have that $\mathbb{E}\left[\xi_h^k | \mathbb{F}_h^k\right] = 0$, and thus ξ_h^k is a martingale difference sequence (MDS) satisfying $|\xi_h^k| \le 2H$ since $|V^k - V^{\pi^k}| \le H$ as $V^k, V^{\pi^k} \in [0,H]$. A straightforward application of Azuma-Hoeffding yields that w.p.a.l. $1-\delta/4$:

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k \le 4H\sqrt{KH\log(8/\delta)}$$

We remark here that terms ξ_H^k for $k \in [K]$ are all zero. as $V_{H+1}^k = V_{H+1}^{\pi^k} = 0$ by definition. We union bound this event with the one in **Lemma 3.4** (optimistic with high probability) from now onwards, thus what remains of the proof occurs w.p.a.l. $1-\delta/2$.

From **Lemma 3.4.**, as the algorithm is optimistic, we can upper bound V_h^* by V_h^k for any timestep h to yield:

$$\text{Regret}(K) \le \sum_{k=1}^{K}\left(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\right)$$

A one-step look-ahead gives the general form of the summands:

$$r(s_1^k, a_1^k) + \mathbb{E}\left[V_2^k | s_1^k, a_1^k\right] + 2b_1^k(s_1^k, a_1^k) - r(s_1^k, a_1^k) - \mathbb{E}\left[V_2^{\pi^k} | s_1^k, a_1^k\right]$$

reducing to

$$2b_1^k(s_1^k, a_1^k) + \mathbb{E}\left[V_2^k - V_2^{\pi^k} | s_1^k, a_1^k\right]$$

Therefore, we have the following regret decomposition:

$$
\begin{aligned}
\mathrm{Regret}(K) &\leq \sum_{k=1}^{K} \left(2b_1^k(s_1^k, a_1^k) + \mathbb{E}\left[V_2^k - V_2^{\pi^k} | s_1^k, a_1^k\right]\right) \\
&= \sum_{k=1}^{K} \left(V_2^k(s_2^k) - V_2^{\pi^k}(s_2^k) + \xi_1^k + 2b_1^k(s_1^k, a_1^k)\right) \\
&\leq \sum_{k=1}^{K} \left(V_3^k(s_3^k) - V_3^{\pi^k}(s_3^k) + \xi_1^k + \xi_2^k + 2b_1^k(s_1^k, a_1^k) + 2b_2^k(s_2^k, a_2^k)\right) \\
&\leq \dots \\
&\leq \sum_{k=1}^{K} \sum_{h=1}^{H-1} \xi_h^k + \sum_{k=1}^{K} \sum_{h=1}^{H} 2b_h^k(s_h^k, a_h^k)
\end{aligned}
$$

The statement of the lemma immediately follows, as we already obtained the upper control to the term in ξ_h^k.

🚲

3.2 Proof of Lemma 3.3.

We assume now we have **Lemma 3.2.** to show **Lemma 3.3.**.

Remark 3.5. We highlight here that **Lemma 3.2.** is an extremely technical lemma that relies on smart constructions of ϵ-independent sequences in order to use the ϵ-eluder dimension of $\mathcal{F}$ as an upper bound to the quantity of interest. We again refer the curious reader to the original work, as we would like to stay focused on the main branch of thought for the proof of the overall regret bound claim.

Proof. We use $b_h^k := b_h^k(s_h^k, a_h^k)$ as a shorthand, and arrange them into $b_1 \geq b_2 \geq \cdots \geq b_T$. Conditioning on the events of **Lemma 3.2.**, we have that when $b_t \geq 1/T$, so will $b_1, \dots, b_{t-1}$, and thus

$$
t \leq \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{1}\left\{b_h^k > b_t\right\} \leq \left(\frac{c\beta(\mathcal{F}, \delta)}{b_t^2}\right) \cdot \dim_E(\mathcal{F}, b_t)
$$

but $\dim_E(\mathcal{F}, b_t) \leq \dim_E(\mathcal{F}, 1/T)$. Call the longest sequence which has length equal to the eluder dimension *eluder-defining*, i.e. it is the longest sequence which bears the property that defines the eluder dimension. We readily see that an ϵ-eluder-defining sequence can also be an ϵ'-eluder-defining sequence for $\epsilon' \leq \epsilon$, as it would also satisfy the properties of the ϵ'-eluder dimension.

Isolating for b_t, we have that

$$
b_t \leq \left(\frac{t}{\dim_E(\mathcal{F}, 1/T)} - H\right)^{-1/2} \sqrt{c\beta(\mathcal{F}, \delta)}
$$

To proceed with computing $\sum_{t=1}^{T} b_t$, we split the sum in two parts. When $t \leq \dim_E(\mathcal{F}, 1/T) \cdot H$, we upper-bound b_t with $4H$ (Q-functions are bounded), whereas when $t > H \dim_E(\mathcal{F}, 1/T)$, the authors breezed through some complicated combinatorial arguments to show that

$$
\sum_{t=1}^{T} b_t \leq 1 + 4H^2 \dim_E(\mathcal{F}, 1/T) + 2\sqrt{c \cdot \dim_E(\mathcal{F}, 1/T) \cdot T \cdot \beta(\mathcal{F}, \delta)}
$$

We recognize the factor of $4H$ stated previously in the final expression.

🚲

Thus, we are now ready to piece **Lemmas 3.1** and **3.3.** together for the overall regret bound.

3.3 Regret bound proof

Proof. We have that the algorithm enjoys a regret of

$$\operatorname{Regret}(K) \leq \min \left\{ KH, 2\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k + 4H\sqrt{KH\log(8/\delta)} \right\}$$

By collecting all constants onto a term $\kappa > 0$, we obtain

$$\operatorname{Regret}(K) \leq \kappa \min \left\{ KH, H^2 \dim_E(\mathcal{F}, 1/T) + \sqrt{\dim_E(\mathcal{F}, 1/T) \cdot T \cdot \beta(\mathcal{F}, \delta)} + H\sqrt{KH\log(1/\delta)} \right\}$$

w.p.a.l. $1 - \delta$. 🚲

3.4 Comments

The term $\beta(\mathcal{F}, \delta)$ is chosen so that the Bellman backup operator updates fall within $\mathcal{F}_h^k$ (defined previously) with high probability. In particular, this scalar is decided via the sensitivity sampling process that we decided to omit. Precisely, this coefficient is given in Algorithm 3 of Wang et. al. and scales as follows:

$$\beta(\mathcal{F}, \delta) = \tilde{O}\left(H^2 \cdot \log^2(T/\delta) \cdot \dim_E(\mathcal{F}, \delta/T^3) \cdot \ln(\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \log(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T) \cdot T/\delta)\right)$$

We notice here the H^2 dependency, as well as the dependency on the eluder dimension of $\mathcal{F}$ and the log-covering numbers of $\mathcal{F}$ and $\mathcal{S} \times \mathcal{A}$. Substituting this value into our final expression in the regret bound proof (3.3) yields the general theorem made in [28].

In general, proofs of UCB-based algorithms such as those of [11, 7] employ very similar patterns. They start off with a regret decomposition into martingale difference sequences involving $\delta_h^k := V_h^k - V_h^{\pi^k}$ plus additional terms for the bonuses as well as the policy optimization step. In our case, the policy optimization (greedy w.r.t. Q_h^k) step contributes to the overall regret via δ_h^k. Then, we look to apply the high probability guarantee that our estimates are optimistic to upper-bound the regret into something we can work with, in our case, that would be rewriting it in terms of the sum of bonuses. We ended up with only having to control the latter, as the martingale difference sequence was handled by Azuma-Hoeffding.

The usage of eluder dimensions is indeed more apparent in the sensitvity sampling portion of the paper, as it is used to control the size of the surrogate dataset $\hat{\mathcal{Z}}^k$. However, it does have its place in the later portion of the proof: it is via the crucial **Lemma 3.2.** that we recognize the importance of the boundedness of $\dim_E \mathcal{F}, \epsilon$, the entire proof of **Lemma 3.3.** would not make sense as the upper bound to b_t is essentially ∞.

4 A Thompson sampling approach?

Inspired by recent developments of optimistic sampling techniques by Zanette et. al. [30] and Ishfaq et. al. [9], one could modify Wang et. al.'s algorithm to instead sample bonuses from a normal distribution, and then taking the max Q-function. This approach stems from the posterior sampling literature, where one is theoretically sampling from the posterior over the space of all MDPs. However, as Abeille et. al. pointed out in [2], it suffices to sample from any distribution with certain concentration and anti-concentration properties to approximate this intangible and computationally intractable posterior. In practice, posterior sampling methods are preferred over UCB-based methods, primarily due to the approximation shortcut via sampling perturbations from normal distributions instead of computing large numbers of UCB bonuses.

Back to our proposed modification, let $M \in \mathbf{N}$ be the number of samples, we sample random functions $\xi_h^{k,m}$ for $m = 1, \ldots, M$ such that for inputs (s, a), $\xi_h^{k,m}(s, a) \propto N(0, \kappa \cdot w(\mathcal{F}, s, a)$ with $\kappa > 0$. In this way, we obtain m Q-functions $\left\{Q_h^{k,m} = f_h^k + \xi_h^{k,m} : m \in [M]\right\}$ and update according to $Q_h^k = \min\left\{\max_{m \in [M]}\left\{Q_h^{k,m}\right\}, H\right\}$. Here, κ and M are chosen so that the function Q_h^k is optimal with constant probability, that is, it upper bounds the true Q-function with constant probability, much in the same way Wang et. al.'s work does. One usually argues that optimistic sampling bears strong resemblance to UCB and is essentially no different from UCB, however the method is inherently grounded in posterior sampling as perturbations are generated randomly, and in a sense, one is still sampling from an approximate posterior. However, despite being more application-friendly, this approach is yet to be proven an efficient exploration algorithm, as there are currently no known bounds. However, we speculate that modifying works of Ishfaq et. al. [9] as well as very recent works currently under blind review at ICLR 2021 and thus cannot be cited here, one might be able to control regrets of the above algorithm with eluder dimensions and log-covering numbers in the same way that Wang et. al.'s proofs.

5 Conclusion and future work

We have seen a few key pieces to the proof of the general regret bound of the work by Wang et. al. [28], in particular, the regret decomposition and bounding each component individually. We also saw the importance of the boundedness assumption of the eluder dimension of the search space $\mathcal{F}$, as it is key to making various lemmas work as well as appears in the bound to the overall regret of the algorithm. We pointed out that UCB proofs follow a very similar pattern, and suggest that one can potentially look into laxing the various assumptions of this paper and apply the same patterns for more general provably efficient algorithms, as long as one can make a statement about the search space.

The assumptions are indeed restrictive, the general class of neural networks (without any further conditions) enjoy infinite eluder dimensions for any ϵ, as well as infinite log-covering numbers. Furthermore, the log-covering numbers of the state-action space is in general either unknown or computationally intractable. Neural networks are significant due to their universal approximation properties, and if one has a theory for general neural networks, one has the holy grail to provable efficiency in reinforcement learning. Unfortunately, the marginal return from relaxing the assumptions bit by bit to still employ eluder dimension and log-covering numbers type proofs might be diminishing, as these notions are inherently combinatorial and might not be rich enough to formulate a more general theory for exploration in reinforcement learning, as we can already see from the above. This calls for the development of alternate frameworks of analysis of reinforcement learning algorithms, or a richer way of characterizing some notion of "width" to a search space.

It could also be interesting to employ these techniques in the development and analysis of information-directed sampling methods, methods that incite exploration based on some notion of information gain, for the linear MDP setting as well as the GFA setting. It's been shown in [13, 21] that they are a promising paradigm that, when carefully designed, could help tackle RL settings with heteroscedastic rewards. Not having as much attention as the popular UCB and Thompson sampling paradigms of value/policy optimization-based algorithms, there is definitely room for development.

Acknowledgements

I would like to thank Eric and Haque for insightful discussions and sharing their enthusiasm for this subject. I would like to thank Prakash for pretending not to notice that I never came to class. I would like to thank my guitar for keeping me company. Finally I would like to thank Gabriela for her advanced memery pre-deadline.

References

[1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320. Curran Associates, Inc., 2011.

[2] Marc Abeille, Alessandro Lazaric, et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

[3] Smruti Amarjyoti. Deep reinforcement learning for robotic manipulation-the state of the art, 2017.

[4] Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin F Yang. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

[5] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*, 2017.

[6] Prasanna Balaprakash, Romain Egele, Misha Salim, Stefan Wild, Venkatram Vishwanath, Fangfang Xia, Tom Brettin, and Rick Stevens. Scalable reinforcement-learning-based neural architecture search for cancer deep learning research. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, Nov 2019.

[7] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[8] Eric Hu and Viet Nguyen. Provable efficiency: Finding regret bounds in reinforcement learning. 2020.

[9] Haque Ishfaq, Zhuoran Yang, Andrei Lupu, Viet Nguyen, Lewis Liu, Riashat Islam, Zhaoran Wang, and Doina Precup. Provably efficient policy optimization with thompson sampling. 2020.

[10] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[11] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient?, 2018.

[12] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020.

[13] Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. *arXiv preprint arXiv:1801.09667*, 2018.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc., 2012.

[15] Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. *arXiv preprint arXiv:1902.10089*, 2019.

[16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

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

[18] Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. 2002.

[19] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling, 2017.

[20] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26:2256–2264, 2013.

[21] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27:1583–1591, 2014.

[22] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[24] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* A Bradford Book, Cambridge, MA, USA, 2018.

[25] Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

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

[28] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33, 2020.

[29] Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: A survey, 2020.

[30] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.