Provable Efficiency: Finding Regret Bounds in Reinforcement Learning

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

Viet Nguyen
McGill University
Montreal, QC
baviet.nguyen@mail.mcgill.ca

Abstract

Reinforcement learning algorithms have been making great progress in many domains, from superhuman levels of play in abstract strategy games [12] to robot optimal control [8]. However, while these algorithms often show impressive empirical results, they also tend to lack rigorous theory proving their efficiency.

In this work we shall examine the elements underlying the current research regarding provably efficient reinforcement learning algorithms and consider their effectiveness and practicality in context. In particular we shall focus on the emergent use of eluder dimension in proving efficiency, the use of importance sampling in provably efficient algorithms, and touch upon the methods of encouraging exploration.

1 Incentivizing Exploration

Balancing exploration and exploitation remains one of the main goals guiding reinforcement learning research.

The two most explored methods of encouraging exploration are Thompson sampling, which is essentially sampling from posterior distributions, and optimism-based algorithms, which explicitly rewards uncertain actions.

Here we consider such methods used and the role they play in algorithms with provable efficiency.

1.1 Thompson Sampling

Thompson sampling long been a known strategy to the multi-armed bandit problem, but it has recently enjoyed a surge in popularity due to its empirical success in machine learning applications.

Potential advantages of Thompson sampling algorithms over optimistic algorithms include the ability to incorporate the structure of the problem into the prior distribution and computational advantages depending on the problem setting.

In the context of learning an unknown finite Markov Decision Process, Ouyang et al. [9] present a Thompson sampling algorithm that achieves a regret bound of

$$R(T) \leq (H+1)\sqrt{2|\mathcal{S}||\mathcal{A}|T\log(T)} + 49H|\mathcal{S}|\sqrt{|\mathcal{A}|T\log(|\mathcal{A}|T)}$$

Where H is a special constant, $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, and T is the number of time steps on which the algorithm is run.

We note that there are particular assumptions in this algorithm, namely that it is weakly communicating (every state is either transient or is reachable from every other non-transient state under every stationary policy), is finite, and satisfies extra restrictions on the transition probabilities.

The main benefit derived from using Thompson sampling instead of UCB seems to be empirical, since the algorithm claims a computational benefit of having to solve fewer MDPs (only the one sampled during policy iteration) and showing lower regret in a couple selected examples.

However, this does show that Thompson sampling is comparable both empirically and theoretically in certain cases, meaning that this method of encouraging exploration could culminate in another promising branch of theory.

1.2 Optimism in the Face of Uncertainty

The idea of using optimistic parameters is among the most popular methods of balancing exploration against exploitation, using the principle of optimism in the face of uncertainty.

Algorithms that use this principle construct confidence sets of the system parameters and use optimistic parameters to select actions during each time step.

1.2.1 Upper Confidence Bound

One of the most popular methods utilizing this principle is the upper confidence bound, which can be interpreted as a bonus function that boosts the expected rewards of relatively unexplored states and actions.

In the main work which we are focusing on, the general outline for the UCB variant is given thus:

Giving the confidence set for fixed $\beta > 0$ as

$$\hat{\mathcal{F}} = \left\{ f \in \mathcal{F} | \left\| f - \hat{f} \right\|^2 \leq \beta \right\}$$

then the bonus function b is implemented as

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

then the Q-function used to select the optimal policy in the k'th time step is given by

$$Q_k^h = \min\{f_h^k + b_h^k, H\}$$

Of course, the true value of b cannot be efficiently measured, so the main innovation of the paper is finding an estimate for b that (with probability $1-\delta$) satisfies enough desired properties to demonstrate both upper and lower bounds on b.

1.3 Maximum Entropy Reinforcement Learning

It may be worth considering the other functions used to incentivize exploration. To this end the idea behind maximum entropy reinforcement learning is to encourage the agent to behave stochastically. Intuitively, stochastic policies allow the agent to learn not only the solution to a given problem but all possible solutions to the problem, which allows for fine tuning and generalization.

However, while there are many results indicating the practical value of this version of the bonus function, there is a regrettable lack of work proving showing the theoretical efficiency of these algorithms.

This being said, Hazan et al. [4] propose an algorithm that has provable sample and computational complexity bounds.

In this work, the authors present an algorithm that makes use of an oracle that calculates a near-optimal policy (with a parameterized sub-optimality gap) and an estimator of the state distribution (with parameterized maximum distance from the true state distribution) in order to produce a policy that is within ϵ of the optimal policy according to any β-smooth entropy measure R for any β.

With these functions, the authors develop an algorithm that gives the following bounds:

Theorem 1.1. (Maximum Entropy Reinforcement Learning Bounds)

We are given $\epsilon > 0$, an oracle function A that given inputs r, ϵ_1 outputs a policy such that $V_\pi \geq \max_{\pi \in \Pi} V_\pi - \epsilon_1$, an oracle function E that given inputs π, ϵ_0 outputs a state distribution d such that $\|d_\pi - d\| \leq \epsilon_0$, and a reward functional R that is β-smooth, B-bounded, and satisfies the following inequalities:

$$\|\nabla R(X) - \nabla R(Y)\|_\infty \leq \beta \|X - Y\| \ \forall X, Y$$
$$\|\nabla R(X)\| \leq B \ \forall X$$

Then for all $\epsilon > 0$, then there is an algorithm that when run for T iterations with

$$T \geq 10\beta\epsilon^{-1} \ln 10B\epsilon^{-1}$$

then

$$R(d_{\pi,T}) \geq \max_{\pi \in \Pi} R(d_\pi) - \epsilon$$

Of course, the issue with this is the number of conditions required to give the bound as well as the presence of functions that may be computationally infeasible to calculate.

In the specific lower dimensional cases for which empirical trials were done, the authors were able to get past the computational demand of the oracle functions by reducing the dimensions of the spaces in which they implement the algorithm using heuristic methods such as projection onto lower dimensional spaces and kernel density estimation.

In more complex settings, it is likely that these two oracle functions will be computationally infeasible to calculate without approximations that allow some probability of failure (the reward bounds are guaranteed under the current algorithm).

2 Eluder Dimension

A central element to many of the recently developed provably efficient reinforcement learning algorithms is the eluder dimension, formally described by Russo and Van Roy [11], which can be interpreted in the online learning context to represent the ability of an adversary to avoid giving data that will accurately represent the entire space.

First, we establish preliminary definitions before presenting the full concept:

Definition 2.1. ($\mathcal{F}$-ϵ dependence/independence) Given a function class $\mathcal{F}$ over a space $\mathcal{A}$, a fixed $\epsilon > 0$ and a set $A = \{a_1, \cdots, a_n\} \subseteq \mathcal{A}$, a value a is ϵ-dependent on A if for all of pairs functions $f, \hat{f} \in \mathcal{F}$ such that

$$\sqrt{\sum_{i=1}^{n} (\hat{f}(a_i) - f(a_i))^2} \leq \epsilon$$

Then $|f(a) - \hat{f}(a)| \leq \epsilon$.

We then say that a is ϵ-independent of A if a is not ϵ-dependent on A.

Definition 2.2. (Eluder Dimension) Given an $\epsilon > 0$, a function class $\mathcal{F}$ over a space $\mathcal{A}$, the eluder dimension, denoted as $\dim_E(\mathcal{F}, \epsilon)$ is the longest sequence $(a_1, \cdots, a_n)$ such that for all $1 < i \leq n$, there is an $\epsilon' \geq \epsilon$ where a_i is ϵ'-independent of $\{a_1, \cdots, a_{i-1}\}$.

The intuition behind this definition goes thus: In a space $\mathcal{A}$, an online learner will use previous information to attempt inference of future inputs, and so it might reasonably expect that if two functions $f, \hat{f}$ have similar values according to previously known information $\{a_1, \cdots, a_n\}$, then they would have similar values when given a as well.

The eluder dimension then describes how much trouble this learner would have under this expectation by considering the worst-case scenario in which its information about $f, \hat{f}$ on the existing set will not help it in learning how $f, \hat{f}$ are related for a.

The authors justify the naming by relating this value to other dimensions, where the dimension of a vector space can be described as the size of the longest sequence of linearly independent vectors and the VC dimension can be described as the size of the largest set of points that can be shattered by the function class.

It is also worth noting that this definition is not equivalent to the longest sequence (a_i) such that a_i is ϵ-independent of all the a_j before it, because for $0 < \epsilon' < \epsilon$ it is possible for elements to be ϵ-independent but not ϵ'-independent.

However, such a sequence can still serve as a lower bound for eluder dimension since we can pick $\epsilon' = \epsilon$ to satisfy the criteria given for such sequences.

2.1 Eluder Dimension in Reinforcement Learning

One of the most difficult and valuable branches of reinforcement learning focuses on algorithms that use general value function approximation. These algorithms have the most potential applications but also require efficiency that holds over generic classes of functions. These algorithms are already widespread in practice, though, since it is usually impossible to exactly describe reward functions and the promise of neural networks as universal approximators has given significant empirical results.

We shall be focusing on the work by Ruosong et al. [13], since its claimed regret bounds are among the tightest of such provable bounds seen so far and it touches upon several novel concepts which are used in other works showing regret or reward bounds.

In this case, the authors describe an algorithm that has the following claimed regret bound:

Claim 2.3. (Regret Bound in the Realizable Case) We consider an MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$, and select estimates of the value function from a function class $\mathcal{F}$ with a time horizon of H.

By realizable we mean that, defining the bounded reward function $r : \mathcal{S} \times \mathcal{A} \longrightarrow [0, 1]$, the value function $V : \mathcal{S} \longrightarrow [0, H + 1]$, and the Q-function $Q : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbf{R}$ in the usual way, there is a function $f_V \in \mathcal{F}$ such that

$$f_V(s, a) = r(s, a) + \sum_{s' \in S} P(s'|s, a)V(s') \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}$$

Then, assuming that the optimal reward function is in $\mathcal{F}$, there is an algorithm that after interacting with the environment for $T = KH$ steps (K episodes of length H), with probability at least $1 - \delta$, achieves a regret bound of

$$\mathrm{R}(K) \leq \sqrt{\iota \cdot H^2 \cdot T}$$

Where for a fixed $C > 0$,

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

For the non-realizable case, there is a similar claim involving the misspecification error, which is described using a value ζ such that given a set of functions $\mathcal{F} \subseteq \{f : S \times A \longrightarrow [0, H + 1]\}$, for all $V : \mathcal{S} \longrightarrow [0, H]$ there is some $f_V \in \mathcal{F}$ with

$$\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| f_V(s, a) - r(s, a) + \sum_{s' \in S} P(s'|s, a)V(s') \right| \leq \zeta$$

Claim 2.4. (Regret Bound with Misspecification) With all terms defined identically to as above, under the above definition of ζ, there is an algorithm that achieves a regret bound of

$$\mathrm{R}(K) \leq \sqrt{\iota \cdot H^2 \cdot T} + \sqrt{\dim_E(\mathcal{F}, 1/T) \cdot H \cdot \zeta} \cdot T$$

There are many terms which must be examined here, each of which has nontrivial contribution to the bound. Along with the eluder dimension, we note the presence of covering numbers for both the function space and the combined state and action space. These factors emerge due to the use of sensitivity sampling, which we shall discuss later on.

Unfortunately, although the regret bound on general function models depends on the eluder dimension, for general functions no such upper bound exists and we can easily construct spaces with infinite eluder dimensions (e.g. $\ell^2 \subseteq \mathcal{F}(\mathbf{N})$, with any sequence of distinct natural numbers being ϵ-independent for any $\epsilon > 0$).

This is not a shortcoming of this work in and of itself, but it means that the bounds proven require that the function class is generally well behaved according to this newly defined value. For example, if it turns out that certain desirable classes of functions have very large or infinite eluder dimensions, then these bounds will be useless in theory.

Indeed, the authors acknowledge the nascent state of the theory on eluder dimension by mentioning that one of the steps in the algorithm dependent on this value can be tunable if the eluder dimension is unknown. In practice (for spaces of finite eluder dimension) this should suffice for most practical purposes.

With this being said, there are results giving upper bounds on the eluder dimensions of specific classes of functions, such as the following bound for generalized linear spaces due to Russo and Van Roy [10].

Theorem 2.5. (Eluder Dimension of Generalized Linear Spaces)

Consider $\Theta \subseteq \mathbf{R}^d$ and a given feature mapping ϕ so that $f_\theta(a) = g(\theta^T \phi(a))$ with g differentiable and strictly increasing. Then, assume the existence of constants $\underline{h}, \overline{h}, \gamma, S$ where for all $a \in \mathcal{A}, \rho \in \Theta$, $0 < \underline{h} \leq g'(\rho^T \phi(a)) \leq \overline{h}, \|\rho\|_2 \leq S, \|\phi(a)\|_2 \leq \gamma$.

Then, $\dim_E(\mathcal{F}, \epsilon) \leq 3dr^2 \frac{e}{e-1} \ln\left(3r^2 + 3r^2 \cdot \left(\frac{2S\overline{h}}{\epsilon}\right)^2\right) + 1$

where r is given as

$$r = \frac{\sup_{\rho \in \Theta, a \in \mathcal{A}} g'(\langle \phi(a), \rho \rangle)}{\inf_{\rho \in \Theta, a \in \mathcal{A}} g'(\langle \phi(a), \rho \rangle)}$$

Proof. We first define for any sequence $a_1, \cdots, a_k$ the values

$$w_k := \sup\left\{(f_{\rho_1} - f_{\rho_2})(a_k) | \sqrt{\sum_{i=1}^{k-1}((f_{\rho_1} - f_{\rho_2})(a_k))^2} \leq \epsilon', \rho_1, \rho_2 \in \Theta\right\}$$

Then, if $w_k \geq \epsilon'$ then letting $\Phi_k = \sum_{i=1}^{k-1} \phi\phi^T$, $V_k = \Phi_k + \lambda I$, $\lambda = \left(\frac{\epsilon'}{2S\underline{h}}\right)^2$,

$$w_k \leq \max\{\rho^T \phi_k | \sum_{i=1}^{k-1} g(\rho^T \phi(a_i))^2 \leq (\epsilon')^2, \rho^T I \rho \leq (2S)^2\}$$

and by the bound on g' we have

$$\begin{aligned} w_k &\leq \max\{\rho^T \phi_k | \sum_{i=1}^{k-1} g(\rho^T \phi(a_i))^2 \leq (\epsilon')^2, \rho^T I \rho \leq (2S)^2\} \\ &\leq \max\{\overline{h}\rho^T \phi_k | \underline{h}^2 \rho^T \Phi_k \rho \leq (\epsilon')^2, \rho^T I \rho \leq (2S)^2\} \\ &\leq \max\{\overline{h}\rho^T \phi_k | \underline{h}^2 \rho^T V_k \rho \leq 2(\epsilon')^2\} \\ &= \sqrt{2(\epsilon')^2/r^2} \cdot \|\phi_k\|_{V_k^{-1}} \end{aligned}$$

Where since $\rho^T I \rho \leq (2S)^2$, then $\rho^t V_k \rho = \rho^T \Phi_k \rho + \rho\rho^T \leq (\epsilon')^2 + \lambda \underline{h}^2 (2S)^2 \leq (\epsilon')^2 + (\epsilon')^2 = 2(\epsilon')^2$.

This means that $\phi_k^T V_k^{-1} \phi_k \geq \frac{1}{2r^2}$.

Then, if $w_i \geq \epsilon'$ for each $i < k$ then $\det V_k \geq \lambda^d \left(\frac{3}{2}\right)^{k-1}$

In particular, using the Matrix Determinant Lemma,

$$\det V_k = \det V_{k-1}(1 + \phi_t^T V_k^{-1} \phi_t) \geq \det V_{k-1} \left(\frac{3}{2}\right) \geq \cdot (\det \lambda I) \left(\frac{3}{2}\right)^{k-1} = \lambda^d \cdot \left(\frac{3}{2}\right)^{k-1}$$

Then, since the determinant of V_k is maximized when all its eigenvalues are equal,

$$\det V_k \leq \left(\frac{1}{d}\text{Tr}(V_k)\right)^d \leq \left(\frac{\gamma^2(t-1)}{d} + \lambda\right)^d$$

Then, this inequality means that k must satisfy

$$\left(\frac{3}{2}\right)^{\frac{k-1}{d}} \leq 1 + \left(\frac{\gamma^2}{\lambda}\right) \cdot \frac{k-1}{d}$$

Then, fixing $0 \leq x \leq \frac{1}{2}$ and $\alpha > 0$, for all values p such that $(1+x)^p \leq \alpha p + 1$, $p \geq 1 \implies \ln(1+x) \cdot p \leq \ln(1+\alpha) + \ln p$, and using the fact that $\ln(1+x) \geq x/(1+x)$, we substitute $y = \frac{px}{1+x}$ to get

$$\begin{aligned} y &\leq \ln(1+\alpha) + \ln\frac{1+x}{x} + \ln y \\ &\leq \ln(1+\alpha) + \ln\frac{1+x}{x} + \frac{y}{e} \\ \frac{e}{e-1}y &\leq \ln(1+\alpha) + \ln\frac{1+x}{x} \\ y &\leq \frac{e}{e-1}\left(\ln(1+\alpha) + \ln\frac{1+x}{x}\right) \end{aligned}$$

So that for all $p \geq 1$, $p \leq \frac{1+x}{x} \cdot \frac{e}{e-1}(\ln(1+\alpha) + \ln\frac{x+1}{x})$

Then, using this inequality on the $\frac{k-1}{d}$ term with $x = 1/2, \alpha = \left(\frac{\gamma^2}{\lambda}\right)$ we obtain

$$k \leq d \cdot 3 \cdot \frac{e}{e-1} \cdot (\ln(3r^2 + 3r^2\frac{\gamma^2}{\lambda})) + 1 = 3dr^2\frac{e}{e-1}\ln\left(3r^2 + 3r^2 \cdot \left(\frac{2S\overline{h}}{\epsilon}\right)^2\right) + 1$$

Note that if $p < 1$ in the above equation then although the given inequality does not apply, we instead get $k < d + 1$, which is a far stronger result.

Since the eluder dimension is equal to the largest such k, the desired result follows. ∎

2.2 Uses in Regret Decomposition

Russo and Van Roy's primary use of the eluder dimension is to bound the regret of Thompson sampling and standard UCB algorithms in a contextual multi-armed bandit model [11].

In order to do so, the authors first define a width function, defined over subsets of function classes $\tilde{\mathcal{F}} \subseteq \mathcal{F}$ as:

$$w_{\tilde{\mathcal{F}}}(s,a) = \sup_{f,f' \in \tilde{\mathcal{F}}} (f(s,a) - f'(s,a))$$

Then, an intermediate proposition is established:

Proposition 2.6. We consider a set of functions $\mathcal{F} \subseteq \{f : A \longrightarrow [0, C]\}$ for $C > 0$. For a sequence $\{\mathcal{F}_t : t \in \mathbf{N}\}$ of measurable function classes under the restriction that for all $t \in \mathbf{N}$, $R_t - f_\theta(A_t)$ conditioned on (H_t, θ, A_t) is η-sub-Gaussian, then for all $f \in \mathcal{F}, a \in \mathcal{A}$,

for UCB-based algorithms,

$$R(T, \pi^{\mathcal{F}_{1:\infty}}) \leq \sum_{t=1}^{T} w_{\mathcal{F}_t}(A_t) + C\mathbf{1}(f_\theta \notin \mathcal{F}_t)$$

and for Thompson sampling based algorithms,

$$\mathbf{E}[R(T, \pi^{TS})] \leq \mathbf{E}\left[\sum_{t=1}^{T} w_{\mathcal{F}_t}(A_t) + C\mathbf{1}(f_\theta \notin \mathcal{F}_t)\right]$$

with probability 1.

Proof. We consider the true reward function f_θ with θ a random variable (so that there is some mean reward combined with noise).

Then, assume that $f_\theta \in \tilde{\mathcal{F}}$. In this case, then the regret must be at most the width of that subset, since for all $f \in \tilde{\mathcal{F}}$,

$$f_\theta(a^*) - (f_\theta(a) + w_{\tilde{\mathcal{F}}}(a)) \leq 0 \iff f_\theta(a^*) - f_\theta(a) \leq w_{\tilde{\mathcal{F}}}(a)$$

This inequality is true since it essentially states that the algorithm would only choose the action a if it believed that it had the optimal reward, and the optimistic estimate of the reward given by taking action a must be bounded above by the sum of $f_\theta(a)$ (the actual reward) and $w_{\tilde{\mathcal{F}}}(a)$ (the greatest possible uncertainty of the reward estimation).

On the other hand, if $f_\theta \notin \tilde{\mathcal{F}}$, then because $f(a) \in [0, C]$ for all $f \in \tilde{\mathcal{F}} \subseteq \mathcal{F}$, the worst regret would always be at most C since the difference between any two rewards is bounded from above by $C - 0 = C$.

Thus, the given formula follows by using the width function if the function is in the confidence set and C otherwise, and summing over the time steps. ∎

If the confidence sets contain f_θ with high probability for each θ, then this regret bound simplifies to the sum of width functions, and we note that we should expect w_{F_t} should gradually decrease over time as an agent learns more about its environment.

Then, the eluder dimension is then used to bound the width function for each step, so that summing these terms over all time steps will give a bound in terms of eluder dimension:

$$R(T, \pi^{\mathcal{F}_{1:\infty}}) = \tilde{O}\left(\sqrt{\dim_E(\mathcal{F}, T^{-2})} \cdot \log(\mathcal{N}(\mathcal{F}, T^{-2}, \|\cdot\|_\infty)\right)$$

More importantly, though, this regret decomposition becomes useful in later analyses, since bounding the width function during each time step will then lend itself to corresponding regret bounds.

The work by Ruosong et al. uses an estimate of the width function itself in order to calculate the bonus function in their variant of UCB, meaning that the bonus functions can instead be used to bound the regret.

3 Importance Sampling

Importance sampling is technique finding a resurgence in the machine learning field. In particular, the goal of this sampling strategy is to estimate certain properties of a probability distribution by sampling from a different distribution.

The logic behind importance sampling follows from the rearrangement

$$\int f \mathrm{d}P = \int f \cdot \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}Q$$

Where the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is renamed as the importance function. Then, it is possible to take advantage of this second integral if estimating it is easier or lower variance than for the first.

Since the core of machine learning in general is to infer a policy by sampling from a distribution, it naturally follows that importance sampling can be used in various steps of reinforcement learning algorithms.

3.1 Policy Optimization by Importance Sampling

One explored use for importance sampling in reinforcement learning is for policy optimization. In this case, Metelli et al. present a method that uses importance sampling in order to lower the variance of policy estimation and hence increase the rate of convergence [7].

In particular, importance sampling is used for off-policy evaluation, in which the goal is to estimate the expected value of a deterministic value of a bounded function over a probability distribution P using samples collected from another distribution Q.

In order to be able to infer information in P from $x = (x_1, \cdots, x_N)^T$ sampled from Q, the work introduces an importance sampling estimator, which introduces importance weights (the Radon-Nikodym derivative) $w_{P/Q}(x) = p(x)/q(x) = \frac{\mathrm{d}P}{\mathrm{d}Q}(x)$ to the sample, so that giving $\hat{\mu}$ as the importance sampling estimator,

$$\hat{\mu}_{P/Q} = \frac{1}{N} \sum_{i=1}^{N} w_{P/Q}(x_i) \cdot f(x_i)$$

Then, in order to reduce the potential of infinite variance in this estimator, a normalized version of this estimator is also introduced, namely

$$\widetilde{\mu}_{P/Q} = \frac{\sum_{i=1}^{N} w_{P/Q}(x_i) f(x_i)}{\sum_{i=1}^{N} w_{P/Q}(x_i)}$$

which we can interpret as the expected value of P under the approximation by N Dirac deltas centered at each x_i and weighted by $w_{P/Q} / \sum_{j=1}^{N} w_{P/Q}(x_j)$.

Then, the variance of the the former function $\hat{\mu}_{P/Q}$ can be bounded by

$$\mathrm{Var}_{x \sim Q}[\hat{\mu}_{P/Q}] \leq \frac{1}{N} \|f\|_{\infty}^2 \, d_2(P \parallel Q)$$

where d_2 is the exponentiated 2-Rényi divergence defined as

$$d_2(P \parallel Q) = \exp\left(\log \int_{\mathcal{X}} \left(\frac{dP}{dQ} \right)^2 dQ \right) = \int_{\mathcal{X}} q(x) \left(\frac{p(x)}{q(x)} \right)^2 dx$$

and ESS represents the effective sample size, where $\mathrm{ESS}(P \parallel Q, N)$ is approximately equal to the number of samples drawn from P so that $\tilde{\mu}_{P/P}$ has the same variance as $\tilde{\mu}_{P/Q}$ with N samples.

Instead of optimizing these two estimators directly (which would give a P maximizing the probability mass centered at the maximum sampled $f(x_i)$), the authors seek to optimize a lower bound on the expected value $\mathbf{E}_{x \sim P}[f(x)]$ that would hold with some high probability. The stated reasoning for why this doesn't fall into the same trap is that these overfit probability distributions would display a high variance and thus be punished by a loss function that accounts for estimator variance.

To determine such a function, they use the following theorem:

Theorem 3.1. Consider P, Q probability measures on the measurable space $(\mathcal{X}, \mathcal{F})$ with $P \ll Q$ and $d_2(P \parallel Q) < \infty$. Then, letting $x_1, \cdots x_n$ be i.i.d random variables sampled from Q and $f : \mathcal{X} \rightarrow \mathbf{R}$ a bounded function, for all $0 < \delta \leq 1$, $N > 0$, with probability at least $1 - \delta$:

$$\mathbf{E}[f(x)] \geq \hat{\mu}_{P/Q} - \|f\|_{\infty} \sqrt{\frac{(1-\delta)(d_2(P \parallel Q))}{\delta N}}$$

and maximize this value instead with the side effect of leaving δ as a hyperparameter.

Then, the authors employ this surrogate loss (which can be empirically estimated from the samples and the current policy) in order to develop two optimization algorithms, centered around parameterized and differentiable policy spaces.

Unfortunately, there were not theoretical guarantees found to back up the empirical results and the empirical results did not significantly outperform the state of the art, so it is unlikely that this particular use for importance sampling will be further explored.

3.2 Sensitivity Sampling

One of the innovations crucial to the work of Ruosong et al. is the modification of the bonus function to not only estimate the confidence interval for the Q-function but also do so in a stable way in order to increase accuracy [13].

In this case, the authors employ the technique of sensitivity sampling, with the goal of reducing the size of the dataset while keeping the confidence region intact.

Definition 3.2. (Sensitivity) For a set of state action pairs $\mathcal{Z} \subseteq \mathcal{S} \times \mathcal{A}$ and function class $\mathcal{F}$, for each $z = (s, a) \in \mathcal{Z}$ let the λ-sensitivity of z with respect to $\mathcal{Z}$ and $\mathcal{F}$ be

$$\text{sensitivity}_{\mathcal{Z},\mathcal{F},\lambda}(s,a) = \max_{f,f' \in \mathcal{F}, \|f-f'\|_{\mathcal{Z}}^2 \geq \lambda} \frac{(f(s,a) - f'(s,a))^2}{\|f - f'\|_{\mathcal{Z}}^2}$$

The work then uses this to gradually create a subsample of a set of state action pairs $\mathcal{Z} \subseteq \mathcal{S} \times \mathcal{A}$ with certain desirable properties.

We can interpret sensitivity as the importance of (s, a) in $\mathcal{Z}$ by measuring its maximal contribution to the norm $\|f - f'\|$ over any two functions $f, f' \in \mathcal{F}$.

Returning to the work done by Ruosong et al., we consider their use of sensitivity sampling in order to increase the stability of the bonus function used in their variant of UCB.

Their particular subsampling strategy creates a subsample $\mathcal{Z}'$ of $\mathcal{Z}$ by independently deciding for each $z \in \mathcal{Z}$ to include $\frac{1}{p_z}$ copies of z with probability p_z (and not including any copies otherwise).

Here, p_z is defined to be the smallest real number that is the reciprocal of an integer and

$$p_z \geq \min \left\{ 1, \text{sensitivity}_{\mathcal{Z},\mathcal{F},\lambda}(z) \cdot \frac{72}{\epsilon^2} \ln \left(4\mathcal{N} \left(\mathcal{F}, \frac{\epsilon}{72} \cdot \sqrt{\frac{\lambda}{\delta |\mathcal{Z}|}} \right) \right) \right\}$$

Although the expected number of elements in $\mathcal{Z}'$ is $|\mathcal{Z}|$, there will be fewer distinct elements, namely $\sum_{z \in \mathcal{Z}} p_z$. Using the fact that for all real numbers x at least 1, there is an integer in $[x, 2x]$ we have

$$\sum_{z \in \mathcal{Z}} p_z \leq 2 \cdot \frac{72}{\epsilon^2} \ln \left(4\mathcal{N} \left(\mathcal{F}, \frac{\epsilon}{72} \cdot \sqrt{\frac{\lambda}{\delta |\mathcal{Z}|}} \right) \right) \sum_{z \in \mathcal{Z}} \text{sensitivity}_{\mathcal{Z},\mathcal{F},\lambda}(z)$$

Then, the authors use an upper bound on the sum of the sensitivities of the data in terms of the eluder dimension and the log-covering number of $\mathcal{S} \times \mathcal{A}$, so that the number of distinct elements in $\mathcal{Z}'$ is probably (with probability $1 - \delta$) much smaller than the number of elements in $\mathcal{Z}$.

Combining all this work to create this subsample with simply rounding samples in $\mathcal{Z}'$ to a specified precision creates a set of functions that is both relatively close to the original sample while being drastically reduced in complexity.

More specifically, the subsampling strategy (with high probability) bounds the cardinality of $\mathcal{Z}'$ by $\sum_{z \in \mathcal{Z}} p_Z$ (whose upper bound we shall abbreviate as N) and the rounding limits each $z \in \mathcal{Z}'$ to be a respresentative in the covering set of $\mathcal{S} \times \mathcal{A}$. Then, letting the rounding be to a precision of $\epsilon > 0$, then the total possible number of such sets $\mathcal{Z}'$ with rounded values is at most $\mathcal{N}(\mathcal{S} \times \mathcal{A}, \epsilon)^N$.

This bound, while extremely large, is finite when the covering number is finite, and is used to then establish the stability of the bonus function (and thus the regret) with high probability.

3.3 Covering Number

It is important to note the role of covering number in the regret bounds for several of the given algorithms. Here we consider the covering numbers for certain spaces in order to extract more explicit regret bounds for such spaces.

Definition 3.3. (Covering Number) The covering number of a metric space (X, d) given a parameter $\epsilon > 0$, denoted as $\mathcal{N}(X, d, \epsilon)$, is the minimal cardinality over any set S of balls of radius ϵ such that for all $x \in X$, there is some $s \in S$ such that $d(s, x) \leq \epsilon$. This may also be applied to subsets A of X, which we similarly denote as $\mathcal{N}(A, d, \epsilon)$.

Often times the metric is omitted and the notation $\mathcal{N}(X, \epsilon)$ is used instead.

The use of covering number in regret bounds is much more non-obvious than the use of eluder dimension, but emerges as a result of a particular strategy that Ruosong et al. employ in their implementation of the bonus function. In short, the algorithm subsamples from the set of samples and rounds these subsamples to a lower degree of precision in order to (with probability $1 - \delta$) obtain certain desirable properties of its bonus function. The explicit rounding immediately elicits use of covering numbers, and the subsampling method also eventually involves such values.

Here we shall consider the space of generlized linear models with a Lipschitz link function over the domain $[-1, 1]$.

In particular, for the domain $[-1, 1]^d$ under the Euclidean metric we can bound the covering number from above by

$$\mathcal{N}([-1, 1]^d, \epsilon) \leq \left(\frac{2\sqrt{d}}{\epsilon}\right)^d$$

Which we obtain by packing the space with cubes of side length $\frac{\epsilon}{\sqrt{r}}$ that are strictly contained inside each open ball of radius ϵ.

Thus, under appropriate restrictions for generalized linear models (such as restrictions of domains to be $[-1, 1]^d$ and the requirement for the ink function to be Lipschitz) we can establish the log-covering number the function classes and the state-action pairs to be $\tilde{O}(d)$ so that applying the general theorem gives a bound of

$$\begin{aligned} R(T) &\leq \sqrt{\iota \cdot H^2 \cdot T} \\ &= \tilde{O}(\sqrt{\dim_E^2(\mathcal{F}, \delta/T^3) \cdot d \cdot d \cdot T \cdot H^2 \cdot T}) \\ &= \tilde{O}(\sqrt{d^2 \cdot d \cdot d \cdot T \cdot H^2 \cdot T}) \\ &= \tilde{O}(\sqrt{d^4 \cdot H^2 \cdot T}) \end{aligned}$$

Where we use the $O(d^2)$ bound on the eluder dimension of generalized linear spaces from before.

This is somewhat worse than state of the art provable regret bounds, which use properties specific to generalized linear functions. The authors remark that a more careful analysis can improve the bounds for the same algorithm by careful analysis and reiterate that this algorithm is for general function classes, not just generalized linear classes.

For a more interesting class of functions we are able to establish similar bounds under the right conditions, as is shown by Kühn [6]:

Theorem 3.4. (Covering Number of Gaussian Kernel Spaces) Denoting the RKHS of the Gaussian Kernel $K(x, y) = e^{-\sigma^2 \|x-y\|_2^2}$ over the unit box $[0, 1]^d$ as $H_\sigma([0, 1]^d)$, we consider this space as a subspace of continuous functions using the embedding $I_{\sigma,d} : H_\sigma([0, 1]^d) \longrightarrow C([0, 1]^d)$. Then, denoting the value $\mathcal{N}(I_{\sigma,d}(H_\sigma([0, 1]^d), \epsilon)$ as N,

$$0 < \liminf_{\epsilon \longrightarrow 0} \log \mathcal{N}(I_{\sigma,d}(H_\sigma([0, 1]^d), \epsilon) \cdot \left(\frac{\log(1/\epsilon)^{d+1}}{(\log \log 1/\epsilon)^d}\right)^{-1}$$

and

$$\limsup_{\epsilon \longrightarrow 0} \log \mathcal{N}(I_{\sigma,d}(H_\sigma([0, 1]^d), \epsilon) \cdot \left(\frac{\log(1/\epsilon)^{d+1}}{(\log \log 1/\epsilon)^d}\right)^{-1} < \infty$$

Informally, this means that the log-covering number of the Gaussian RKHS as a subspace of the set of continuous functions varies roughly according to

$$\frac{\log(1/\epsilon)^{d+1}}{(\log \log 1/\epsilon)^d}$$

Under this restriction of the domain, the state-action space already has a polynomial log-covering number, so if the eluder dimension for this space is reasonable (polynomial in d), then the algorithm given by Ruosong et al. will guarantee a polynomial regret bound.

While there is not yet a proven eluder dimension bound for Gaussian Kernel spaces, extremely recent work such as by Yang et al [14] have used alternate methods to find regret bounds for reinforcement learning over a generic RKHS. However, since this paper uses methods outside the scope of this work, we shall acknowledge the proposal of such algorithms but refrain from exploring further.

4 Conclusion

The background underlying provable bounds for reinforcement learning still has the nascent qualities of the machine learning field as a whole in the sense that there are still many independent directions for research that are rapidly defining and building off of both new and old ideas.

We observe that at the moment few spaces have proven upper bounds on the eluder dimension, so while these algorithms might perform well in toy cases with well studied properties, their general applicability is still contingent on how well behaved more general spaces will be in terms of this dimension. However, given these bounds, there are already significant results and potentially useful algorithms that use this dimension in order to demonstrate their efficiency. On the other hand, we also observe that in the past importance sampling did not seem affect algorithm efficiency, and only recently does it seem to have any impact on provable regret bounds. As a result, its usefulness is highly uncertain, but it shows potential in establishing theory due to its potential to reduce variance in samples.

We thus find that although the general methods of encouraging exploration are quite well established, there is still a diversity of strategies and problems that can be explored within these methods. All in all, we think that there is promise in using these strategies despite their qualities, we hope that algorithms based on such theory will prove themselves empirically as well.

References

[1] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1283–1294, Virtual, 13–18 Jul 2020. PMLR.

[2] Simon S. Du, Jason D. Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic q-learning with function approximation in deterministic systems: Tight bounds on approximation error and sample complexity, 2020.

[3] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies, 2017.

[4] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2681–2691, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[5] Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling, 2019.

[6] Thomas Kühn. Covering numbers of gaussian reproducing kernel hilbert spaces. *Journal of Complexity*, 27(5):489 – 499, 2011.

[7] Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 5442–5454. Curran Associates, Inc., 2018.

[8] OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik's cube with a robot hand. *CoRR*, abs/1910.07113, 2019.

[9] Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 1333–1342. Curran Associates, Inc., 2017.

[10] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling, 2013.

[11] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 2256–2264. Curran Associates, Inc., 2013.

[12] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[13] Ruosong Wang, Ruslan Salakhutdinov, and Lin F. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension, 2020.

[14] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I. Jordan. Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations, 2020.