INEXPERIENCED RL AGENTS CAN'T GET IT RIGHT: LOWER BOUNDS ON REGRET AT FINITE SAMPLE COMPLEXITY

Maia Fraser
Department of Computer Science
University of Ottawa
Canada
mfrase8@uottawa.ca

Vincent Létourneau
Department of Computer Science
University of Ottawa
Canada
vletour2@uottawa.ca

ABSTRACT

We consider a family $\mathcal{M}$ of MDPs over given state and action spaces, and an agent that is sequentially confronted with tasks from $\mathcal{M}$. Although stated for this stepwise change in distributions, the insight we develop is informative for continually changing distributions as well. In order to study how structure of $\mathcal{M}$, viewed as a learning environment, impacts the learning efficiency of the agent, we formulate an RL analog of fat shattering dimension for MDP families and show that this implies a nontrivial lower bound on regret as long as insufficiently many steps have been taken. More precisely, for some constant c which depends on shattering d states, an inexperienced agent that has explored the learning environment for fewer than d steps will necessarily have regret above c on some MDP in the family.

1 INTRODUCTION

Lifelong learning in biological or artificial agents consists in continually learning in such a way that previous experience produces an inductive bias useful for current and future learning; see (Parisi et al., 2019) for a recent overview, or (Thrun & Mitchell, 1995). While this kind of learning is a hallmark of biological life, it remains a key challenge in machine learning. In reinforcement learning (RL), the lifelong learning formalisms mainly considered are of two general types - successive or continuous - either the environment and task change at discrete time points so the agent faces new tasks/environments in succession, or there is an ongoing task that continues while the environment continuously changes. Experience from earlier tasks/phases can be transferred to future learning by various means, including methods that model the environment, for example state representation Lesort et al. (2018) or abstraction such as an interaction graph McGovern & Barto (2001); Stolle & Precup (2002); Menache et al. (2002); Şimşek et al. (2005) also used for creation of macros or options that encode useful skills (Bacon, 2018; Sutton et al., 1999). Many new lifelong RL approaches have appeared along these lines, and in some cases regret upper bounds are given Fruit & Lazaric (2017); Ortner et al. (2019). The present paper does not offer a new algorithm for extracting knowledge useful to future learning, recording it in some manner or deploying it. Rather, if provides a brief theoretical side note showing how structure of a class $\mathcal{M}$ of MDPs can imply that any RL agent with only limited experience will do poorly on some MDP of the class, and moreoever this "bad" MDP may be chosen from among members of $\mathcal{M}$ that are indistinguishible to the agent, i.e., with prescribed behaviour at the parts of state space the agent has visited. This is not an asymptotic result, but a lower bound on finite-horizon regret after finite exploration. It sheds light on how properties of a class of MDPs induce limitations on how much an agent can learn, i.e. how close to optimal its learned inductive bias can be, after only brief experience in the environment.

Despite the importance of representation learning there are few learning-theoretic results accompanying its use, even in the setting of supervised learning. In RL, several upper bounds on regret have appeared for specific algorithms that use representation, e.g. Ortner et al. (2019) in the case of state representation, or Fruit & Lazaric (2017) for options. The latter also proves a lower bound on regret that applies to that algorithm, an options variant of UCLR.

A commonly used notion of sample complexity in RL, or *sample complexity of exploration*, is the number of times an agent does not act near-optimally (Kakade, 2003; Brunskill & Li, 2013). We use a weaker notion, namely the total number of steps performed - suboptimally or not - by the agent. This is a closer analog of sample complexity in supervised learning and it allows us to speak separately of the experience that the agent had up until now in its life, vs. the performance of the agent which we only evaluate on future learning. We are focused on this transfer. What does the structure of the MDP say about how good the acquired inductive bias of an agent can be at finite sample complexity?

There has so far been relatively little development of learning-theoretic tools and results in RL, although some key works have defined and studied learning-theoretic notions that apply to algorithms, for instance if an algorithm A is (ϵ, δ)-PAC or its high-probability regret or expected regret can be controlled Dann et al. (2017). Such properties are focused on long-term convergence to optimality. In contrast, we focus on lower bounds for regret and are concerned not with convergence in the long run, but suboptimality in the short run.

2 Complexity of MDPs - lower bounding the regret of arbitrary algorithms

To achieve the stated implication we introduce an RL-analog of fat shattering dimension that applies to a *family* $\mathcal{M}$ of MDPs with fixed state and action spaces S, A such that transition probabilities are from a class $\mathcal{T}$ and reward functions from a class $\mathcal{R}$. This approach parallels the use of concept classes in supervised learning and has the advantage that one can obtain min-max lower bounds of the form *for any algorithm, there is some "bad" MDP in the family $\mathcal{M}$ such that the algorithm has regret above $c > 0$*. Indeed, by measuring an algorithm's performance by how well it can perform across the range of environments from $\mathcal{M}$, rather than just a single MDP $M \in \mathcal{M}$, it is possible to obtain meaningful lower bounds that apply to arbitrary algorithms. Such bounds would be out of reach if focused on a single MDP, where - in the absence of additional assumptions such as available information - an algorithm could in principle be hardcoded with an optimal policy.

2.1 Setting and notation

In this first exploration of complexity for MDP families, in this paper we restrict to finite state space S and finite action space A, but we do not require the transition and reward classes $\mathcal{T}$ and $\mathcal{R}$ to be finite. We consider deterministic, stationary policies drawn from a space Π. This is assumed to be the space $\Pi = S^A$ of all maps $\pi : S \to A$, but it is possible to restrict to a smaller Π (i.e. weaken both the definition of our complexity measure and the lower bound we prove), though in this case we still require that Π have product structure, i.e., $\Pi = \prod_{s \in S} \Pi_s$; this is a form of locality, saying that Π constrains policies at each state s independently of what the policy does at other states, a property which is trivially satisfied when $\Pi = S^A$.

Note that the elements of $\mathcal{M}$ correspond uniquely to dynamics-reward pairs $(\tau, R) \in \mathcal{T} \times \mathcal{R}$. For convenience we identify $\mathcal{M}$ with $\mathcal{T} \times \mathcal{R}$ in designating its elements. Also, for a given $(\tau, R) \in \mathcal{M}$ and $s_0 \in S$, we write τ_{s_0} and R_{s_0} for the real-valued functions on $A \times S$ obtained by putting $s = s_0$ as first coordinate in $\tau(s, a)$ resp. $R(s, a, s')$. Recalling that $\tau(s, a)$ is a probability mass function on S, this function will sometimes be denoted $\tau_s(a)$. For use later, we write $\mathcal{M}_s := \{(\tau_s, R_s) : (\tau, R) \in \mathcal{T} \times \mathcal{R}\}$ to record the range of local structure at s that is available in MDPs $M \in \mathcal{M}$. It is helpful to observe that for a specific MPD (τ, R), if we fix a policy π, then (τ_s, R_s), completely determines the local reward (distribution) at s, whereas the distribution of the number of visits made to s depends on τ globally, and is independent of R. Our (semi-)supervised-inspired complexity measure treats these two distributions as analogs respectively of $p(\cdot|x)$ and $p_X(\cdot)$.

For convenience, given a policy $\pi \in \Pi$, we write s_t for the time-t state of the Markov process defined by the MDP-policy pair (M, π). This is a random variable, whose distribution is determined by π and the dynamics τ of M, although we have abusively used lower case s_t.

Analogous to s_t, we write s_t^* for the time-t state of (M, π^*) where π^* is an optimal policy for $M = (\tau, R) \in \mathcal{M}$, i.e. a policy achieving maximal expected return according to the definition in the next sections. Like s_t, s_t^* is a random variable; its distribution is determined by π^* and the dynamics τ of M.

2.2 Expected total return and regret

We focus on finite-horizon analysis and define *return*, i.e. *cumulative reward*, as T-step average. Assume a particular initial distribution ρ on S. *Expected return* is then:

$$\mathbb{E}_{s_1 \sim \rho,\ s_t \sim \tau \text{ for } t>1} \frac{1}{T} \sum_{t=1}^{T} R(s_t, \pi(s_t), s_{t+1}).$$

An alternative would be to consider a discounted version of this, also T-step; this can be handled by modifying the proof we give for average reward. The finite-horizon choice, on the other hand, is linked to our definition of uniform shattering, the RL form of complexity we propose. In that definition we place a condition on distributions of $s_t, t = 1, \ldots, T$. To instead study infinite horizon return, and prove lower bounds for that quantity, would require a correspondingly stronger condition for all $s_t, t > 0$.

Given a definition for expected return, the *regret* in M of a policy π is defined as the amount by which the expected return of π in M is suboptimal compared to other policies. In other words, letting π_* be an optimal policy for M, the regret of π is

$$\text{Я}(\tau, R, \pi) := \mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} R(s_t^*, \pi_*(s_t^*), s_{t+1}^*)] - \mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} R(s_t, \pi(s_t), s_{t+1})].$$

2.3 UNIFORM SHATTERING

This paper makes a first attempt at providing a notion of complexity for "RL problems" which, as described above, we formalize as a family of MDPs.

Fix $\alpha \in (0, 1/2)$ and $\nu \in (0, 1)$. The complexity condition we propose here is modeled after ν-uniform α-shattering, which was defined in Fraser (2016) to extend PAC-style analysis to semi-supervised learning.

Roughly speaking, ν-uniform α-shattering is a measure of complexity for a joint statistical model $\mathcal{P}$ on $X \times Y$ where it quantifies the variability achieved by conditional distributions $p(\cdot|x)$ while the corresponding marginals $p_X(\cdot)$ are constrained. More precisely, it records the existence of a collection $\underline{C} \subset \mathscr{P}(X)$ of n disjoint regions in X and a subfamily $\underline{\mathcal{P}} \subset \mathcal{P}$ for which the associated conditional distributions $p(\cdot|x)$ are able to "match" (with margin α) arbitrary choices $\epsilon \in \{0, 1\}^n$ of labels over the n regions, while corresponding marginals $p_X(\cdot)$ assign at least ν/n probability mass to each region. Here we are using *matching ϵ with margin α* as shorthand for the target map f_p associated[1] to $p(\cdot|x)$ satisfying $f_p(x) < 1/2 - \alpha$ on the regions where $\epsilon = 0$ and $f_p(x) > 1/2 + \alpha$ where $\epsilon = 1$.

In many setups (e.g. absolute loss, mean as target map), this matching condition can be re-expressed as ensuring f_p have local loss above/below α compared to the majority value of an arbitrary reference function $f : X \to \{0, 1\}$ fixed in advance. Here the choice of above/below must match ϵ viewed as an indicator function on $\underline{C}$, i.e., subset of $\underline{C}$. This observation plays a key role in proving lower bounds based on uniform shattering and helps motivate the following.

As analog to ν-uniform α-shattering, the measure of complexity we propose for the MDP family $\mathcal{M}$ records the existence of a set $\underline{S}$ of n states in S and a subfamily $\underline{\mathcal{M}} \subset \mathcal{M}$ of (τ, R) such that if we fix any reference function $\pi_0 : \underline{S} \to A$ then distributions of the random variables $R_s(a, s')$, $s' \sim \tau_s(\pi(s))$ associated to $(\tau, R) \in \underline{\mathcal{M}}$ are able to "match" (with threshold α) arbitrary subsets of the n states of $\underline{S}$ while the dynamics τ in $\underline{\mathcal{M}}$ ensure proportion of time visiting each of the states of $\underline{S}$ remains close (within $\delta > 0$) to ν/n.

The quantity $\mathbb{E}_{s' \sim \tau|_s}[R_s(\pi_0(s), s')] - \mathbb{E}_{s' \sim \tau|_s}[R_s(\pi(s), s')]$ will serve as RL-analog for local loss in the semi-supervised version that was sketched above; we refer to it as *local regret of π relative to π_0* at s under (τ_s, R_s). It only depends on the dynamic-reward pair locally at s. The expression *matching a subset of $\underline{S}$ with margin α* is now interpreted as requiring the local regret of π_0 at states inside/outside the subset to be greater-than/at-most α. By *local regret of a policy π at state s* we mean local regret of π relative to the optimal policy π_*. This is the amount by which π's expected reward is suboptimal at s for a single execution of $\pi(s)$, i.e. , $\mathbb{E}_{s' \sim \tau}[R(s, \pi_*(s), s'] - \mathbb{E}_{s' \sim \tau}[R(s, \pi(s), s']$, and does not take into account the expected number of times the state s is visited by the agent.

Definition 1. Suppose there is a collection of n states $\underline{S} \subset S$ and a subfamily $\underline{\mathcal{M}} \subset \mathcal{M}$ of size $|\underline{\mathcal{M}}| = 2^n$ with product structure $\underline{\mathcal{M}} = \prod_{s \in S} X_s$, where $\forall s \in \underline{S}$ the set $X_s \subset \mathcal{M}_s$ has exactly two elements which we denote (τ^0, R^0) and (τ^1, R^1) such that the following properties hold. Fixing any choice of policy $\pi \in \Pi$,

1. at every $s \in \underline{S}$, under π one of (τ^0, R^0) and (τ^1, R^1) induces local regret $> \alpha$ and the other local regret $\leq \alpha$;
2. (off-policy version)[2] under each MDP in $\underline{\mathcal{M}}$ and all policies, the induced distributions[3] for s_t, $t = 1, \ldots, T$ assign average probability sufficiently close to ν/n at each $s \in \underline{S}$ so that for the higher local regret in 1. a weighted version also holds

$$\left[\frac{1}{T}\sum_{t=0}^{T} P(s_t^* = s)\right] \mathbb{E}_{s' \sim \tau|_s} R_s(\pi_*(s), s') - \left[\frac{1}{T}\sum_{t=0}^{T} P(s_t = s)\right] \mathbb{E}_{s' \sim \tau|_s} R_s(\pi(s), s') > \frac{\alpha\nu}{n} \qquad (1)$$

In this case we say $\mathcal{M}$ α-shatters the set $\underline{S}$ ν-uniformly.

Remark 1. This form of complexity detects a region $\underline{S}$ of state space S which a certain subfamily $\underline{\mathcal{M}}$ of MDPs is guaranteed to visit reasonably uniformly (say, with average probability approximately ν/n at each state of $\underline{S}$) yet

[1] (often $f_p(x)$ is the mean of $p(\cdot|x)$

[2] For the on-policy version we would need this criterion for s_t, $t = N + 1, \ldots, N + T$.

[3] starting from initial distribution ρ

where the rewards are highly variable in the sense that a given action at any state $s \in \underline{S}$ will produce high rewards for half the MDPs in $\underline{\mathcal{M}}$ and low rewards for the others. Regarding the role that average probability approximately ν/n plays in achieving 2., notice that if $\nu/n < P < \nu/n(1+\delta)$ for both of the average probabilities in (1) and the local regret statement in 1. holds more strictly so $\mathbb{E}_{s'\sim\tau}[R(s,\pi_*(s),s'] - (1+\delta)\mathbb{E}_{s'\sim\tau}[R(s,\pi(s),s'] > \alpha$ then we obtain the condition (1).

Remark 2. This definition is intended as exploration; it may be interesting to develop an analog for state-action pairs, rather than working with states.

Remark 3. The MDPs of $\underline{\mathcal{M}}$ with given value of (τ_s, R_s) for $s \in \underline{S}_0$ are indistinguishable from each other for an agent that has only experienced $\underline{S}_0$ yet we'll show any chosen π will perform poorly on $\underline{S} \setminus \underline{S}_0$ for one of these MDPs.

2.4 Regret bounds

Given the assumed initial distribution ρ used in computing regret, recall $s_1 \sim \rho$ and the distributions governing the random variables s_t and s_t^*, $t \geq 1$ are induced from ρ in the respective Markov processes, i.e., those defined on S by π, τ and π^*, τ respectively. The following theorem studies how quickly an agent can improve its policy. This applies to on-policy or off-policy learning: For any agent that has N steps of experience, there is a bound on how good the learned policy can be - whether it is the current parametrization of policy in on-policy learning, or the current estimate of target policy in off-policy learning.

Theorem 1. *Let $\alpha \in (0, 1/2)$ and $\nu \in (0,1)$. Suppose the family $\mathcal{M}$ ν-uniformly α-shatters the n-set $\underline{S} \subset S$. Then as long as $N < n$, the policy obtained by any RL-agent with only N steps experience will have regret at least $c = \epsilon\alpha\nu/2$ on some MDP of $\mathcal{M}$, where $\epsilon > 0$ is defined by $N = (1-\epsilon)n$, $\epsilon \in (0,1)$. For example, $c = \nu/16$ if $\alpha = 1/4$ and $N = n/2$. Moreover, due to the product structure of $\mathcal{M}$, this holds even with (τ_s, R_s) fixed on all s previously explored.*

Proof. We address here the off-policy case where the notation and argument are simpler.

Under the behaviour policy of the agent, for each ground truth dynamic-reward pair $(\tau, R) \in \mathcal{M}$, the agent may have visited any subset of $\underline{S}$ of size at most N in the course of its first N steps. Write $\underline{S}_0 \subset \underline{S}$ for this set of explored states in $\underline{S}$ and put $l := |\underline{S}_0|$. We have $l \in \{0, \ldots, N\}$. The next computations assume the agent has seen $s \in \underline{S}_0$ and $l := |\underline{S}_0|$, although we omit writing this conditioning.

Let $\pi \in \Pi$ be the policy proposed by the agent after this exploration. Suppose the ground truth dynamic-reward pair (τ, R) belongs to $\underline{\mathcal{M}}$. The overall regret $\text{Я}(\tau, R, \pi)$ of the agent can be lower bounded as follows

$$\begin{aligned}
\text{Я}(\tau,R,\pi) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S}\Big[P(s_t^*=s)\mathbb{E}_{s'\sim\tau}[R(s,\pi_*(s),s'] \\
&\qquad\qquad - P(s_t=s)\mathbb{E}_{s'\sim\tau}[R(s,\pi(s),s']\Big] \\
&\geq \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in\underline{S}\setminus\underline{S}_0}\Big[P(s_t^*=s)\mathbb{E}_{s'\sim\tau}[R(s,\pi_*(s),s'] \\
&\qquad\qquad - P(s_t=s)\mathbb{E}_{s'\sim\tau}[R(s,\pi(s),s']\Big].
\end{aligned}$$

To lower bound the maximum of $\text{Я}(\tau, R, \pi)$ over all MDPs in $\underline{\mathcal{M}}$ we will lower bound the average over the elements of a subfamily of $\underline{\mathcal{M}}$. Averaging has the advantage of linearity so it passes into summands, while the looser lower bound thus obtained is sufficient for our purposes.

Recall that all the above quantities are implicitly conditional on the agent having seen $\underline{S}_0$. Since this may restrict the possible ground truth (τ, R), it is useful to consider, for each of the 2^l choices of (τ_s, R_s) on $\underline{S}_0$, the 2^{n-l} pairs $(\tau', R') \in \underline{\mathcal{M}}$ which coincide with (τ_s, R_s) at $s \in \underline{S}_0$, i.e. , $(\tau_s, R_s) = (\tau'_s, R'_s)$, $s \in \underline{S}_0$. Denote this subfamily of $\underline{\mathcal{M}}$ by $\underline{\mathcal{M}}^{\underline{S}_0}$; to keep notation simple, the assumed (τ_s, R_s), $s \in \underline{S}_0$ is not explicitly written.

Now, assuming the ground truth (τ, R) is determined on $\underline{S}_0$, there are 2^{n-l} pairs $(\tau', R') \in \underline{\mathcal{M}}^{\underline{S}_0}$ which agree with this fixed dynamic-rewards on $\underline{S}_0$. We will average over this family

$$\begin{aligned}
&\frac{1}{2^{n-l}} \sum_{(\tau,R)\in\underline{\mathcal{M}}^{\underline{S}_0}} \mathfrak{R}(\tau, R, \pi) \\
&\geq \frac{1}{2^{n-l}} \frac{1}{T} \sum_{t=1}^{T} \sum_{s\in\underline{S}\setminus\underline{S}_0} \sum_{(\tau,R)\in\underline{\mathcal{M}}^{\underline{S}_0}} \Big[P(s_t^* = s)\mathbb{E}_{s'\sim\tau}[R(s, \pi_*(s), s'] \\
&\qquad\qquad - P(s_t = s)\mathbb{E}_{s'\sim\tau}[R(s, \pi(s), s']\Big].
\end{aligned}$$

To sum the quantity in square brackets over $(\tau, R) \in \underline{\mathcal{M}}^{\underline{S}_0}$ we now view these (τ, R) according to what kind of regret they generate on $\underline{S} \setminus \underline{S}_0$ under the policy π. More precisely, we name $(\tau', R') \in \underline{\mathcal{M}}^{\underline{S}_0}$ according to where in $\underline{S} \setminus \underline{S}_0$ they cause the local regret of π to be above or below α: for every subset $K \subset \underline{S} \setminus L$ of the unvisited states of $\underline{S}$, let (τ^K, R^K) be the unique pair for which local regret of π at each state of K is above α while at states of $\underline{S} \setminus \underline{S}_0$ outside K it is at most α. There are $\binom{n-l}{k}$ sets K of size k, so there are $\binom{n-l}{k}$ elements in $\underline{\mathcal{M}}^{\underline{S}_0}$ that cause local regret of π above α at exactly k states of $\underline{S} \setminus \underline{S}_0$. Moreover, under (τ^K, R^K), the local regret incurred by π is at least α at each of the k states of K, so weighted regret incurred on $\underline{S} \setminus L$ exceeds $k\alpha$ times ν/n. This regret in turn lower bounds the overall regret and we have

$$\begin{aligned}
&\sum_{s\in\underline{S}\setminus\underline{S}_0} \sum_{(\tau,R)\in\underline{\mathcal{M}}^{\underline{S}_0}} \frac{1}{T} \sum_{t=1}^{T} \Big[P(s_t^* = s)\mathbb{E}_{s'\sim\tau}[R(s, \pi_*(s), s'] - P(s_t = s)\mathbb{E}_{s'\sim\tau}[R(s, \pi(s), s']\Big] \\
&> \sum_{k=0}^{n-l} \binom{n-l}{k} k\alpha \frac{\nu}{n} \\
&= (n-l) \sum_{k=0}^{n-l-1} \binom{n-l-1}{k} \frac{\alpha\nu}{n} \\
&= (n-l) 2^{n-l-1} \frac{\alpha\nu}{n} \\
&= (n-l) \frac{2^{n-l}}{2} \frac{\alpha\nu}{n} \geq 2^{n-l} \frac{\epsilon\alpha\nu}{2}.
\end{aligned}$$

Therefore:

$$\frac{1}{2^{n-l}} \sum_{(\tau,R)\in\underline{\mathcal{M}}^{\underline{S}_0}} \mathfrak{R}(\tau, R, \pi) > \frac{\epsilon\alpha\nu}{2}.$$

Notice this lower bound does not depend on $l = |\underline{S}_0|$. So, regardless of how many states of $\underline{S}$ the agent visited in its N-step exploration, there is always some MDP in $\mathcal{M}$ with the same dynamic-rewards that the agent experienced on $\underline{S}_0$ but for which the agent's learned policy will have regret above $\frac{\epsilon\alpha\nu}{2}$. □

Remark 4. If we extend our analysis to continuous state spaces then, instead of a collection $\underline{S}$ of n states, we would use a collection $\underline{S}$ of n disjoint subsets $S_i \subset S$ and the regret condition would require local regret above/at-most α on the majority of the relevant S_i; in this case, as analog to the current computations, instead of $\alpha \cdot \nu/n$ we would have $\alpha/2$ times ν/n as a lower bound on regret over a given S_i.

We now illustrate the uniform shattering property for MDP families and associated regret bounds by some examples.

Example 1. Consider an MDP family with state space $S = \{s_0, s_1, ..., s_n\}$. The action spaces A_i are all identical with 2 actions $\{a_0^i, a_1^i\}$ for all $i = 0, ..., n-1$ and A_n is empty. Taking any actions in A_i transitions to the state s_{i+1} with probability 1. There are 2^n MDPs in the family $\mathcal{M}$, indexed by the ordered tuples $u \in \{0, 1\}^n$ with M_u having the following reward structure: $a_{u_i}^i$ has reward α while the other action in A_i has reward 0. See figure 1.
We check this family 1-uniformly α-shatters the set $\underline{S} = \{s_0, ..., s_{n-1}\}$. Clearly, for any policy and each states of $\underline{S}$ there is a pair of MDPs in $\mathcal{M}$, one for which π has regret α and one with regret 0 so condition (1) of definition 1 is met. Taking the time horizon T to be equal to n, condition (2) is verified by computing the left term of inequality 1 is equal to $\frac{\alpha}{n}$ and then $\nu = 1$. Finally, assuming that an agent has experienced at most $N < n$, theorem 1 says that there is an MDP in $\mathcal{M}$ for which the regret of the policy π is at least $(1 - \frac{N}{n})\frac{\alpha}{T}$.

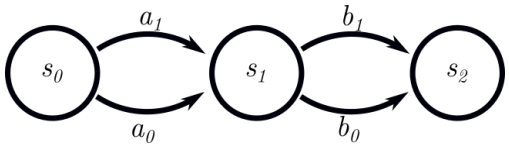

Figure 1: The simple MDP of Example 1.

More generally, in the simple case where π and π_* have the same expected proportion of time spent $\frac{1}{T}\sum_t P(s_t^* = s) = \frac{1}{T}\sum_t P(s_t = s)$ at all $s \in \underline{S}$, then condition 2 of definition 1 takes the simple form: $\frac{1}{T}\sum_t P(s_t = s) \geq \frac{\nu}{n}$

Example 2. By using clustering methods, it is possible to extend the result of theorem 1 to more general settings. Consider for example an MDP family on the path space of the directed graph of the previous example, ie a directed binary tree with $n+1$ levels. Each level $l = 0, ..., n$ contains 2^l states $s_{l,i}$ and each state has two actions transitioning to two different states with probability 1. Every reward function from an MDP in the previous example lifts to an MDP on the path space of the multi-graph and we take the MDP family $\widetilde{\mathcal{M}}$ to be exactly the set of those lifts. Although each state in the tree has a much smaller chance of being visited, because of the way the reward functions of $\widetilde{\mathcal{M}}$ were chosen, the regret bounds from the previous example still holds.

Example 3. In general condition 2 of definition 1 is trickier than condition 1. The notion of bottleneck may help identifying states that have the right bounds on the proportion of time spent on the critical states $\underline{S}$. The interaction graph of an MDP is the graph that has as nodes the states of the MDP and an edge between s and s' if there is an action in A_s that has a nonzero probability of transitioning to s'. The betweeness of a state s is

$$\mathcal{B}(s) = \frac{\sum_{u,v \neq s \in \mathcal{S}} \sigma_{u,v}(s)}{\sum_{u,v \neq s \in \mathcal{S}} \sigma_{u,v}}$$

that is the proportion of all shortest paths that contain s with $\sigma_{u,v}$ the number of shortest paths between u and v (shortest in the sense of the graph distance) and $\sigma_{u,v}(s)$ the number of shortest paths containing s. This quantity is used for example in Bacon & Precup (2013) to identify states that are useful goals of a temporal abstraction. Here we will assume such states, called bottleneck states, in particular allow the transition between two clusters of states, the initial states and another disjoint group of states that have very high reward actions within the group.

Suppose the state space $\mathcal{S}$ of an MDP family $\mathcal{M}$ decomposes as the disjoint union of three sets $\mathcal{S}_i, \mathcal{S}_b, \mathcal{S}_h$, the initial states, the bottleneck states and the mentioned high-reward states. The initial states contain the support of the initial distribution. The set $\mathcal{S}_h$ contain some states that have high reward actions such that an optimal policy of $M \in \mathcal{M}$ always ends up in $\mathcal{S}_h$. The set $\mathcal{S}_b$ contains states that are α-preshattered by a family of MDPs $\mathcal{M}$ (preshattered will mean that only condition 1 of definition 1 is met).

Now fix some policy $\pi \in \Pi$ and suppose there are n bottleneck states $\underline{S} \subset \mathcal{S}_b$ and all have to be visited to cross from $\mathcal{S}_i$ to $\mathcal{S}_h$. Since $\mathcal{M}$ α-preshatters $\mathcal{S}_b$, there is some MDP M for which each state s in $\underline{S}$ has a good action with reward r_+ and a bad action with reward r_- and $r_+ - r_- \geq \alpha$ and at each states $s \in \underline{S}$, π choses the bad action. If Π is the set of all possible policies on M, it will not be the case that all policies transit to the set $\mathcal{S}_h$ and therefore satisfy condition 2 so we will restrict admissible policies to the set Π_h of the ones that do cross to the states $\mathcal{S}_h$. In the setting described, if $s \in \underline{S}$

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T} P(s_t^* = s)\mathbb{E}s' \sim \tau|_s R_s(\pi_*(s), s') - P(s_t = s)\mathbb{E}s' \sim \tau|_s R_s(\pi(s), s') \\
&= \frac{1}{T}\sum_{t=0}^{T} P(s_t^* = s)r_+ - P(s_t = s)r_- \\
&= \frac{1}{T}(r_+ - r_-) \geq \frac{\alpha}{T}
\end{aligned}$$

and condition 2 is verified.

3 CONCLUSION

We have defined uniform shattering, a notion of complexity for a family of MDPs similar in flavour to common measures in supervised learning, such as fat shattering or VC dimension. The main use of these complexity measures in supervised learning is to derive bounds on the risk of the hypothesis produced by learning algorithms. Likewise, we find here a lower bound on the regret of a policy returned by any learning algorithm that has not had an experience broad enough to prevent some unavoidable mistakes. Our observations are mainly of theoretical interest but provide insight into how properties of an MDP class might dictate the need for a specific minimal amount of prior experience in a lifelong learning setting before an agent can count on reducing its regret below a given threshold. The tools and method we use are also more generally of of value for framing and advancing the discussion of RL learning theory; to our knowledge no such agnostic bounds have been stated before. Our lower-bound result addresses only the case of finite discrete state spaces. Further work will explore the specifics of continuous state spaces in this context and how clustering of states can be used to treat regions instead of individual states. Another avenue we wish to investigate is the interplay between complexity and representation, namely how complexity is reduced using the specific representations of the MDP environment, for example options.

REFERENCES

P.-L. Bacon. *Temporal Representation Learning*. PhD thesis, McGill University, 2018.

Pierre-Luc Bacon and Doina Precup. Using label propagation for learning temporally abstract actions in reinforcement learning. In *Proceedings of the Workshop on Multiagent Interaction Networks (MAIN'13)*, 2013.

E. Brunskill and L. Li. Sample complexity of multi-task reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

Özgür Şimşek, Alicia P. Wolfe, and Andrew G. Barto. Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 816–823, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102454. URL https://doi.org/10.1145/1102351.1102454.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5717–5727, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

M. Fraser. Multi-step learning and underlying structure in statistical models. In *NeurIPS 2016*, pp. 4815–4823, 2016.

Ronan Fruit and Alessandro Lazaric. Exploration-exploitation in mdps with options. In Aarti Singh and Xiaojin (Jerry) Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 576–584. PMLR, 2017. URL http://proceedings.mlr.press/v54/fruit17a.html.

Sham Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

T. Lesort, N. Diaz-Rodriguez, and D. Filliat J.-F. Goudou. State representation learning for control: An overview. *Neural Networks*, 108:379–392, 2018.

A. McGovern and A. G. Barto. Automatic discovery of subgoals in reinforcement learning using diverse density. In *Proceedings of the Eighteenth International Conference on Machine Learning. ICML ?01*, pp. 361–368, 2001.

I. Menache, S. Mannor, and N. Shimkin. Q-cut - dynamic discovery of sub-goals in reinforcement learning. In *Proceedings of the 13th European Conference on Machine Learning*, pp. 295–306. Springer-Verlag, 2002.

Ronald Ortner, Matteo Pirotta, Alessandro Lazaric, Ronan Fruit, and Odalric-Ambrym Maillard. Regret bounds for learning state representations in reinforcement learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/9b8b50fb590c590ffbf1295ce92258dc-Paper.pdf.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S0893608019300231.

M. Stolle and D. Precup. Learning options in reinforcement learning. In *Proceedings of the 5th International Symposium on Abstraction, Reformulation and Approximation*, pp. 212–223. Springer-Verlag, 2002.

R. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999.

Sebastian Thrun and Tom M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 15(1):25–46, 1995. ISSN 0921-8890. doi: https://doi.org/10.1016/0921-8890(95)00004-Y. URL https://www.sciencedirect.com/science/article/pii/092188909500004Y. The Biology and Technology of Intelligent Autonomous Agents.