Rethinking the Foundations for Continual Reinforcement Learning

Anonymous authors
Paper under double-blind review

Keywords: Continual reinforcement learning, hindsight rationality, history process, MDPs

Summary

In the traditional view of reinforcement learning, the agent's goal is to find an optimal policy that maximizes its expected sum of rewards. Once the agent finds this policy, the learning ends. This view contrasts with *continual reinforcement learning*, where learning does not end, and agents are expected to continually learn and adapt indefinitely. Despite the clear distinction between these two paradigms of learning, much of the progress in continual reinforcement learning has been shaped by foundations rooted in the traditional view of reinforcement learning. In this paper, we first examine whether the foundations of traditional reinforcement learning are suitable for the continual reinforcement learning paradigm. We identify four key pillars of the traditional reinforcement learning foundations that are antithetical to the goals of continual learning: the Markov decision process formalism, the focus on atemporal artifacts, the expected sum of rewards as an evaluation metric, and episodic benchmark environments that embrace the other three foundations. We then propose a new formalism that sheds the first and the third foundations and replaces them with the history process as a mathematical formalism and a new definition of deviation regret, adapted for continual learning, as an evaluation metric. Finally, we discuss possible approaches to shed the other two foundations.

Contribution(s)

1. We identify four foundational principles and practices that shape and constrain our thinking about RL. We argue that these foundations, shaped by the traditional framing of RL, are antithetical to the purported goals of continual reinforcement learning and may be holding us back from making progress toward continual learning.
   Context: Previous work by Abel et al. (2024b) has discussed three dogmas that shape most reinforcement learning research. The second dogma overlaps with the second foundation that we argue against as part of the foundations of traditional reinforcement learning. We also base most of our arguments in alignment with the big world hypothesis that Javed & Sutton (2024) originally presented.
2. We present a new formalism that replaces two of the foundations with the history process as a mathematical formalism and deviation regret as an evaluation metric.
   Context: The history process foundation is built on earlier work by Bowling et al. (2023), and the deviation regret is an extension to earlier work by Morrill et al. (2021b) to the continual learning setting.
3. We present experimental results suggesting that the current RL algorithms fail to learn continually and that our proposed measure of evaluation can evaluate those failures.
   Context: Platanios et al. (2023) showed similar results for agents failing to learn continually, which aligns with our experimental findings.

Rethinking the Foundations for Continual Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

In the traditional view of reinforcement learning, the agent's goal is to find an optimal policy that maximizes its expected sum of rewards. Once the agent finds this policy, the learning ends. This view contrasts with *continual reinforcement learning*, where learning does not end, and agents are expected to continually learn and adapt indefinitely. Despite the clear distinction between these two paradigms of learning, much of the progress in continual reinforcement learning has been shaped by foundations rooted in the traditional view of reinforcement learning. In this paper, we first examine whether the foundations of traditional reinforcement learning are suitable for the continual reinforcement learning paradigm. We identify four key pillars of the traditional reinforcement learning foundations that are antithetical to the goals of continual learning: the Markov decision process formalism, the focus on atemporal artifacts, the expected sum of rewards as an evaluation metric, and episodic benchmark environments that embrace the other three foundations. We then propose a new formalism that sheds the first and the third foundations and replaces them with the history process as a mathematical formalism and a new definition of deviation regret, adapted for continual learning, as an evaluation metric. Finally, we discuss possible approaches to shed the other two foundations.

1 Introduction

"Consider a Markov decision process defined by the tuple ..." starts many background sections of reinforcement learning (RL) papers. The Markov Decision Process (MDP) formalism, among other foundational concepts, has long shaped how we think about agents, algorithms, and evaluation in RL. However, these foundational concepts stemmed from a classical framing of the RL problem: *an agent's goal is to find an optimal policy that maximizes its expected sum of rewards*. Once this policy is found, the learning ends – the agent no longer needs to adapt because the policy is, by definition, optimal. That traditional view influenced many of the foundations and standard practices in the field. For example, a direct consequence of the view that learning ends with finding an optimal solution is to have a separate training phase with the goal of finding that optimal solution and then have a deployment phase where no more learning is happening. Another consequence is the emphasis on the artifacts that the training process produces and overlooking the behavior of the agent during learning.

There are many decision-making problems where the traditional framing of RL is a shortcoming. For example, agents acting in a world that is much bigger and more complex than themselves, such that they cannot perceive or represent its true underlying state, will neither be able to represent the value of the states they find themselves in nor find an optimal policy (Javed & Sutton, 2024). Such agents can only rely on approximate solutions that continually adapt to perceived changes in their environment and improve as they accumulate more knowledge by interacting with the world. In these types of decision-making problems, learning is no longer about finding an optimal solution but about continual and never-ending adaption. This class of decision-making problems, where continual adaption is necessary, is called continual reinforcement learning (Abel et al., 2024a).

While the traditional and the continual learning views of RL share some similarities — they both tackle the problem of learning by interacting with the world — they have a crucial difference: framing learning as a means to find optimal artifacts versus learning as an indefinite process of adaption. Given this core difference, it is essential to reflect on whether the foundations that has stemmed from the traditional view still hold and are helpful when addressing the continual learning problem. Or could these traditional foundations hold us back from thinking most usefully about the problem?

In the first part of this paper, we identify four traditional foundational principles and practices that shape and constrain our thinking about RL. We argue that these foundations, shaped by the traditional framing of RL, are antithetical to the purported goals of continual reinforcement learning and may be holding us back from making progress toward continual learning. Moreover, these foundations are self-reinforcing: each depends upon and holds up the others, such that when attempting to replace one, the others constrain you to keep it.

In the second part of the paper, we propose a new formalism that sheds two of these foundations, and we discuss possible alternatives for replacing the remaining two foundations.

2 Four Foundations of Traditional RL

Most reinforcement learning research, along with recent progress in continual reinforcement learning, make the following assumptions, implicity or explicitly:

1. **Formalism:** The appropriate mathematical formalism is the *Markov decision process*.
2. **Objective:** The goal of RL algorithms is to produce *atemporal artifacts* (such as an optimal policy or value function).
3. **Evaluation:** The ideal measure of evaluation is the *expected sum of rewards*.
4. **Benchmarking:** Most benchmarks for comparing RL algorithms are *episodic environments*.

These assumptions are the pillars of the traditional RL foundations and remain pervasive within modern RL research. Celebrated results such as DQN reaching human-level performance in Atari (Mnih et al., 2015), AlphaGo (Silver et al., 2016), GT-Sophy (Wurman et al., 2022), balloons in the stratosphere (Bellemare et al., 2020), and DeepStack beating professional poker players (Moravčík et al., 2017) all embody these foundations. They undergo a separate training phase in episodic environments respecting common MDP assumptions such as ergodicity and communicating dynamics. This training process generates atemporal artifacts (policy or value functions) that are considered optimal or near-optimal. These artifacts are then evaluated according to their expected sum of rewards in an evaluation phase where no more learning occurs.

While these foundations were behind most of the advancement of traditional RL research, do they give us an appropriate structure to pursue continual reinforcement learning? Continual reinforcement learning does not have a consensus definition (Ring, 1994; Abel et al., 2024a). However, its very name implies that learning should continue. We now discuss that this conclusion alone is enough to create cracks in those four foundations, and we will briefly summarize the alternatives that could replace those traditional foundations.

Foundation One: MDPs as a Mathematical Formalism. This foundation is concerned with the assumptions on the environment that typically accompany the MDP formalism. We often make ergodicity assumptions, such as the MDP being unichain or communicating, which imply some characteristics of the environment. For example, we may implicitly assume every state is reachable from every other state or that the state distribution converges to some stationary distribution. Furthermore, we usually presume some properties of the MDP, such as finite state and action spaces or compact spaces with continuity assumptions. There are some problems where these assumptions hold and the MDP formalism works well. In grid-world environments, for instance, an agent can revisit any state as often as needed. In Go, repeatedly playing the game in episodes guarantees a form of ergodicity since it allows the agent to repeatedly visit previous game states by replaying the same sequence of moves. However, an important observation is that these are also examples where continual learning is unnecessary.

In contrast, the need for continual learning arises in settings with unpredictable non-stationarity in the environment (Khetarpal et al., 2022) or those that align with the *big world hypothesis* (Javed & Sutton, 2024). The *big world hypothesis* suggests that even if the real world is stationary, its complexity is much richer than the representational capacity of any agent in it. Hence, the world will appear unpredictably non-stationary. When acting in a much more complex world or when there are constraints on the computational resources of the agent, continual learning is needed (Kumar et al., 2023; Dong et al., 2022), even if the underlying world is stationary (Sutton et al., 2007). In these settings, the predictable stationarity of MDPs is invalid. Moreover, real-world settings do not allow one to reset the world into repeatable episodes or revisit states previously visited. You, the reader, can never revisit the state before you read these words. This inability to revisit states renders ergodicity assumptions unrealistic for real-world settings.

The Alternative: History Processes as a Mathematical Formalism. Beyond the agent-environment interface, this formalism has few assumptions about the process since the *big world hypothesis* does not allow the agent to assume a priori structure or regularity about the environment. We expand on this foundation formally in Section 3.

Foundation Two: Focus on Atemporal Artifacts. Artifacts refer to any atemporal representation of an agent's learned knowledge, such as policies, value functions, options, or features. We often give considerable concern to the notion of optimal value functions and optimal policies. The assumption that learning should produce those fixed representations leads us to think of algorithms having a "training" period wherein they aim to converge to optimal artifacts and follow that with a "testing" phase to evaluate the generated artifacts. These artifacts exist for some problems, such as the grid world and chess examples, but they do not exist for problems that require continual learning.

Environments of interest to continual learning rarely admit fixed optimal artifacts. The assumption that an agent can converge to an optimal policy or a value function contradicts the very need for continual adaption since such an atemporal artifact would be the end of learning rather than requiring its continuation. For example, consider an agent with computational constraints that cannot fully represent the values of all possible states in its environment. For that agent, even if a fixed optimal value function theoretically exists, it cannot represent it, compute it, or store it. Instead, such an agent must rely on an approximation of this value function that evolves over time, deciding which information to retain and which to discard. In this context, the most useful value representation is continually adapting and time-dependent, not atemporal. As a result, a focus on fixed atemporal artifacts should be replaced with a focus on the continual adaption of the agent's behavior. This foundation is also notably critiqued as *Dogma Two* by Abel et al. (2024b).

The Alternative: Focus on Behaviour. The goal of RL algorithms is to produce behavior in response to experience. In the continual learning setting, there is no difference between training and testing. All the past experience is training, and all future experience is testing. The focal point is how an agent behaves in response to its experience.

Foundation Three: Expected Sum of Rewards as an Evaluation Measure. In episodic environments, this is the episodic return, and we desire that during training, we see the episodic return approach the return of the optimal policy. Episodes allow drawing i.i.d. samples of this return for any stationary policy, which is how evaluation is usually performed during the testing phase. Hence, maximizing the episodic return during training often leads to better performance during the testing phase.

A salient feature of real-world settings that require continual learning is the inability to reset the world or revisit previous states, i.e., the MDP may not be communicating, as discussed in Foundation One. A ramification of this feature is that it is not even possible to estimate an *expected sum of* *rewards* as it would require the environment to be repeatedly reset to something akin to an initial state so that the agent can reliably try different actions in the same states to achieve the optimal performance criteria. One might think the average reward criterion in a continuing environment is a solution to this criticism. However, without the communicating assumption, a high average reward may be more a property of how fortunate the agent is to end up in a particular communicating class of states with a high average reward. For continual learning settings, we need a measure of evaluation that does not depend on having such a repeatability assumption.

The Alternative: Deviation Regret as an Evaluation Measure. We propose deviation regret as an evaluation measure for continual learning agents. Deviation regret was proposed as the evaluation measure for defining hindsight rationality, originally introduced in the context of strategic games (Morrill et al., 2021b). We further develop this concept for continual reinforcement learning. The essence is that agents should be evaluated on the "situations" they find themselves in, not against some optimal, unrealizable sequence of actions. We formulate deviation regret for continual learning in Section 4.

Foundation Four: Episodic Benchmarks. Common environments, such as classic control tasks and the Arcade Learning Environment (ALE, Bellemare et al. (2013)), are episodic and, therefore, are communicating MDPs. Other naturally continuing environments, such as Mujoco (Todorov et al., 2012) and Minecraft, are often truncated during training, converting them into episodic tasks. A few examples of continuing, never-ending environments, such as Jelly Bean World (Platanios et al., 2023) exist but have not been widely adopted.

Most of these traditional *benchmarks* are problematic when considering the goal of continual learning. They reinforce the idea that environments can always be thought of as ergodic and episodic and exhibit an optimal policy, which is the assumed goal of traditional RL training.

The Alternative: Benchmark Environments Without a Clear Markov State or Episode Reset. We will not expand on this much beyond recognizing that it as an issue. In summary, we should not expect to see continual learning algorithms differentiate themselves in environments where continual learning is unnecessary. Additionally, more work is needed to design environments where continual learning is needed. To make progress, we should have benchmarks that align with the big world hypothesis. Ideally, we should test our agents in the complex, big, real world, but this is impractical for algorithmic development and scientific repeatability. An alternative is to constrain our agents' representational capacity and use more modest-sized environments such that the constraints simulate the big world hypothesis and allow for the development of agents that can cope in such continual learning settings.

Final Remarks on the Traditional Foundations. The four foundations we discussed are self-reinforcing. Just presuming the goal of artifacts immediately suggests the MDP formalism to support the existence of an optimal policy and necessary assumptions to ensure it can be learned, with benchmark environments that fit these assumptions. Similarly, our common benchmark environments have a clear notion of optimal policy, making the focus be on algorithms that produce such an artifact. It is no simple task to tear down any one of these foundations when the others demand its reinstatement. Hence, our proposed alternatives seek to replace all four of these foundations.

3 History Processes as a Mathematical Formalism

For the new formalism to support the goals of continual RL, we need to place as few constraints on the environment as possible. Ideally, constraints would be limited to the interface between the environment and the agent (e.g., actions, observations, rewards) but not on the properties of the environment or its dynamics (e.g., Markovianity, ergodicity). One might consider this as an impossible approach as there needs to be some structure or repeatability in the environment to make learning possible. We will resolve this by making post hoc statements as is common with bandit algorithms, e.g., this agent performs nearly as well as the single best arm in hindsight. Such statements can be made for stationary bandits (with assumptions on the environment) and for adversarial bandits (where limited assumptions are made).

We base the environment definition on the formalism introduced by Bowling et al. (2023), which had a similar aim to approach environments and goals as generally as possible. We deviate slightly from this formalism by assuming that the agent acts first, as in the work by Abel et al. (2024a). Formally, we assume a finite action space, $\mathcal{A}$, and a finite observation space $\mathcal{O}$. We can then define the space of finite-length histories as $\mathcal{H} \equiv \bigcup_{n=0}^{\infty} (\mathcal{A} \times \mathcal{O})^n$, which is the set of all possible sequences of observation-action pairs that can result from the agent-environment interaction. We then define the environment as follows:

Definition 1. An environment e is a function from finite-length histories and actions to a distribution over observations, $e : \mathcal{H} \times \mathcal{A} \to \Delta(\mathcal{O})$.

Finally, we assume that the agent's goal is a preference relation over histories that satisfies the reward hypothesis axioms (Bowling et al., 2023), including temporal γ-indifference. Hence, it can be represented as a reward function mapping from actions and observations to a real-valued number: $R : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$, where the agent's goal is to maximize the expected γ-discounted sum of rewards $R(a_t, o_t)$, summed over the transitions in its history. Since the domain of this function is the finite set of actions and observations, the range of this reward function is bounded.

We continue to follow Bowling et al. (2023) and define an agent as follows:

Definition 2. An agent λ is a function from finite-length histories to a distribution over actions, $\lambda : \mathcal{H} \to \Delta(\mathcal{A})$.

We will focus on agents that can be decomposed into a *representation of state* and a system that learns to select policies over this representation. Formally, let $\mathcal{S}$ be a finite set, which we will call states, and let $S : \mathcal{H} \to \mathcal{S}$ be some fixed partition of the histories such that $S(h) \in \mathcal{S}$ is the agent's representation of the state for history h. Using this state representation, we can specify a notion of a policy, $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, as a mapping from a state to a distribution over actions, with Π being some fixed set of such mappings. Finally, we define an agent's learning rule as follows:

Definition 3. The agent's learning rule σ is a function from finite-length histories to a distribution over policies, $\sigma : \mathcal{H} \to \Delta(\Pi)$.

To illustrate how these definitions interact, consider the history at time t, $h_t \equiv \langle a_1, o_1, \ldots, a_t, o_t \rangle$. Given that history, the agent takes an action $a_{t+1} \sim \pi_t\left(S(h_t)\right)$ where $\pi_t = \sigma(h_t)$. The environment then generates an observation $o_{t+1} \sim e(h_t, a_t)$ creating the new history h_{t+1}.

Remarks. The use of state here should not be confused with the requirements on the state as used in an MDP, such as Markovianity. It is not intended to restrict the dynamics of the environment, it is the agent's own representation of the history. One may require S to be defined in the form of a state update function, $u : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \mathcal{S}$, that defines how states evolve in a recurrent fashion with each each transition from a starting state s_0 as in Morrill et al. (2022).

This kind of decomposition of the agent into a fixed state representation and an adapting policy is explicitly seen in Morrill et al. (2022) and Dong et al. (2022), and implicitly in Abel et al. (2024a). In the latter, they introduce the notion of an *agent basis*: $\Lambda_b \subset \Lambda$, and a learning rule that maps histories to an element of the agent basis. We are essentially choosing Π as our agent basis Λ_b, and we allow the learning rule σ to map to a distribution over the agent basis, i.e., over the policy set Π. As with Abel et al. (2024a), we will examine the agent's learning through its learning rule σ that is adapting the choice of policy π_t from its experience, h_t.

4 Deviation Regret as an Evaluation Measure

Given the history process formalism, we now turn our attention to a measure of evaluation.

4.1 Agents as Creators of Worlds

Given an environment e and a finite-length history h, we can construct a new environment, $e_h(h', a) \equiv e(h \cdot h', a)$, which defines the set of distributions over observations that arise from actions taken after history h. This matches our mathematical formalism for an environment. Thus, as an agent acts in its environment instantiating a sequence of histories $h_1, h_2, \ldots, h_n$, it can be seen as also instantiating a sequence of *worlds*, each world is itself an environment, $e_{h_1}, e_{h_2}, \ldots, e_{h_n}$. **An effective learning agent should be well-adapted to the worlds that it finds itself in.** We will attempt to instantiate this notion using *deviation regret*, extending the notion of hindsight rationality from Morrill et al. (2021b) to continual learning.

We now define a deviation ϕ as a function that systematically applies modifications to the agent's policy. Formally, a *deviation* is defined as $\phi : \Pi \to \Pi$, where Π is the set of all possible policies. For example, a deviation might change the action taken at a singular state, or if the agent's policy is a parametrized function, it might apply a systematic perturbation to the parameters of the policy, generating a new deviation policy. As we discussed in section 3, the agent's learning rule σ generates the agent's policy at each time step given the history up to that time step, i.e., $\pi_t = \sigma(h_t)$. To study an agent under a deviation, we apply the deviation ϕ to the agent's policy in each timestep, producing the deviation policy $\phi(\pi_t)$. Hence, we can further define a function that composes the agent's learning rule with the deviation function: $\phi(\sigma) : \mathcal{H} \to \Delta(\Pi)$.

Deviation regret focuses on the notion of a systematic deviation. For any particular deviation, we care about the agent's *regret* for not applying the deviation, and we sum this regret over opportunities to apply this deviation. In our case, the sequence of opportunities is the sequence of worlds instantiated by the agent's own interaction with the environment. This gives us a deviation regret for deviation ϕ in environment e by agent λ,

$$\underbrace{\rho_T(\phi, \lambda, e)}_{\text{deviation regret}} = \frac{1}{T}\sum_{t=1}^{T}\left(\underbrace{\mathbb{E}\left[\sum_{i=t}^{t+H-1}\gamma^{(i-t)}R_i \middle| \phi(\sigma), H_{t-1}\right]}_{\text{deviation return}} - \underbrace{\mathbb{E}\left[\sum_{i=t}^{t+H-1}\gamma^{(i-t)}R_i \middle| \sigma, H_{t-1}\right]}_{\text{agent return}}\right) \quad (1)$$

where H is an evaluation horizon chosen so γ^H is sufficiently small, and H_t is the history (and corresponding world) experienced by the agent in timestep t. An important note is that we discount rewards at time i with $(i - t)$, since this new world starts at time t, with all previously accumulated rewards $r_1, \ldots, r_{t-1}$ shared by both the deviation return and the agent return (so they cancel in the difference). The purpose of discounting in this way is to treat each world equally rather than treating later worlds as discounted by the time since the beginning of the interaction.

As is common with regret notions, we are interested in whether $\rho_T(\phi, \lambda, e) \to 0$, i.e., the deviation regret is approaching zero almost surely or in expectation for any environment. And if this holds for all deviations $\phi \in \Phi$, we say that the agent is minimizes deviation regret with respect to the set of deviations Φ. What do we choose for the set Φ? This question has interesting answers in the repeated extensive-form game setting (Morrill et al., 2021b;a), but as one concrete example, we might consider Φ to be the class of *external deviations*. An external deviation is a constant function, i.e., $\phi_\pi(\cdot) \equiv \pi$. So we can consider $\Phi_{\text{ext}} = \{\phi_\pi\}_{\pi \in \Pi}$. In this case, deviation regret is comparing the agent's expected return to the expected return of a fixed policy averaged over the worlds experienced by the agent. With no additional assumptions on the environment, this would necessitate an agent that continually learns. Furthermore, as an evaluation measure, deviation regret focuses on the agent's behavior in response to its experience, shifting the focus away from artifacts.

4.2 Deviation-Regret Estimation

We now show that an agent can estimate the deviation regret given its stream of experience. The definition of deviation regret in Eq. 1 consists of two components: the agent return and the deviation return. The rewards along the trajectory of the agent directly estimate the agent return. The deviation return may seem unknowable as it requires a counterfactual estimate of the return under an alternative sequence of policies. However, just as with adversarial bandits, we can estimate the counterfactual return of having applied a deviation as long as the agent's support for policies is always closed under the deviation function, so that one can compute an importance sampling ratio $\frac{\Pr(a_i|\phi(\pi_i))}{\Pr(a_i|\pi_i)}$ and construct an unbiased estimator of the deviation return with bounded variance. This can be achieved by a sufficiently random learning rule.

A precise algorithm for the deviation regret estimate is given in Algorithm 1. While the presented algorithm uses ordinary importance sampling, practical implementations may use other importance sampling variants or variance reduction methods.

Algorithm 1: Estimating the Deviation Regret $\hat{\rho}_T(\phi, \lambda, e)$

Input: Deviation ϕ, agent λ, horizon H, trajectory $\{(h_{t-1}, a_t, r_t)\}_{t=1}^{T}$
Output: Estimated deviation regret $\hat{\rho}_T(\phi, \lambda, e)$
Initialize $\hat{G}_T \leftarrow 0$, $\hat{G}'_T \leftarrow 0$
for $t = 1$ *to* T **do**
  Compute $G_t \leftarrow \sum_{i=t}^{\min(T,t+H-1)} \gamma^{i-t} r_i$
  Compute importance weight $W_t \leftarrow \prod_{i=t}^{\min(T,t+H-1)} \frac{\phi(\pi_i)(a_i|h_{i-1})}{\pi_i(a_i|h_{i-1})}$
  Update $\hat{G}_T \leftarrow \hat{G}_T + G_t$, $\hat{G}'_T \leftarrow \hat{G}'_T + W_t G_t$
Compute $\hat{\rho}_T(\phi, \lambda, e) \leftarrow \frac{1}{T}(\hat{G}_T - \hat{G}'_T)$
return $\hat{\rho}_T(\phi, \lambda, e)$

Our main theorem states that if the agent is sufficiently random, the deviation regret estimator given is consistent, as the agent's experience grows, the agent's estimate of the deviation regret gets arbitrarily close to the true deviation regret, with probability approaching 1.

Theorem 1 (Estimating the H-step Deviation Regret)**.** *The estimator we defined above, $\hat{\rho}_T(\phi, \lambda, e)$, is a consistent estimator of deviation regret $\rho_T(\phi, \lambda, e)$ for all environments e, deviations ϕ, $\gamma \in [0, 1]$, and agents λ that take every action with probability at least $c > 0$ in every timestep. More precisely, for all $\varepsilon > 0$,*

$$\lim_{T \to \infty} \mathbb{P}\big(|\rho_T(\phi, \lambda, e) - \hat{\rho}_T(\phi, \lambda, e)| \leq \varepsilon\big) = 1, \tag{2}$$

where the probability is taken over the random behaviour of the agent acting in the environment.

Theorem 2. *There is a consistent estimator for the case where $\gamma < 1$ and $H = \infty$.*

For the proofs, see the Appendix. While the statements here are asymptotic, the appendix contains a finite sample bound for the H-step deviation return estimator.

4.3 Illustrative Experiments

In this section, we present an illustrative experiment demonstrating that deviation regret is a suitable evaluation measure for continual learning agents. We show that widely used reinforcement learning algorithms, developed around the traditional foundations, often fail in continual learning settings. We then show that when these failures occur, the agent experiences positive deviation regret — meaning there exists a deviation policy, representable by the agent, that would have avoided the failure.

Current algorithms fail to continually learn. To study agents' behaviors when there is no repeatability or resets in the environment, we modified the Swimmer environment from Mujoco (Todorov et al., 2012) and turned it into a continuing task. We then trained a PPO (Schulman et al., 2017) agent in this Continuing Swimmer environment for 50 million steps and repeated the experiment using 10 different seeds. Figure 1 shows the results of this experiment. Across all seeds, agents started learning for some time, and then they all failed. While some seeds managed to learn for longer than others, after 20 million steps of interaction with the environment, all agents had already failed.

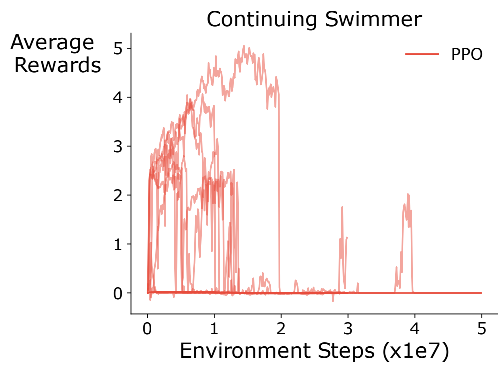

Figure 1: PPO agents failing to continually learn. The lines represent individual runs.

There exists a deviation policy when agents fail to learn. We constructed a deviation set Φ from different checkpoints of the neural network weights, each defining a different policy. These checkpoints contain the network parameters at different points during learning. We then estimated the deviation regret of the agents had they used any of these deviation policies. Finally, we selected the best deviation policy for each agent and sampled estimates for its return starting from various history points. Figure 2 shows the results of this experiment. We plotted the discounted H-step return for the agents and we plotted samples for the deviation return starting from various history points. When the deviation return sample is higher than the agent's return, then there is a positive deviation regret. We can see that when agents fail, the return from the deviation policy is almost always higher than the agent's return, meaning that if the agent had used this deviation policy, it wouldn't have failed.

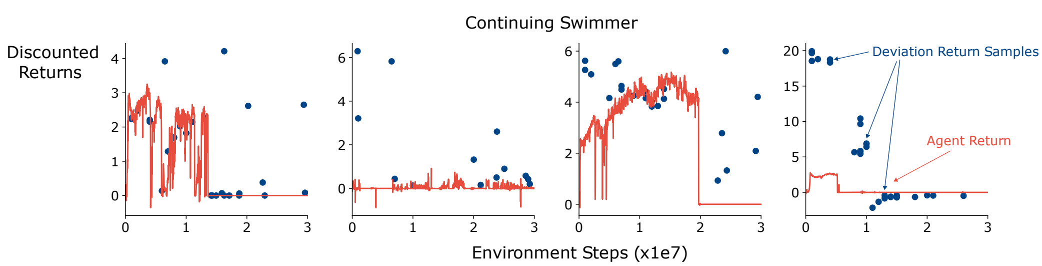

Figure 2: Deviation regret for agents that fail to continually learn.

5 Discussion

We now consider a number of objections that can be raised against this notion of deviation regret and the history process formalism.

Deviations give an alternative and unknowable sequence of worlds. A potential challenge is that systematically applying a deviation would change the distribution of worlds encountered by the agent, which is an unknowable counterfactual. A critical distinction in the choice of deviation regret is that we are not doing *policy regret* (Arora et al., 2012), where the environment within which the deviation's return is evaluated is affected by the applied deviation. However, we also are not making any "oblivious adversary" assumption that the distribution of worlds is not impacted by the agent's actions, i.e., we have an adaptive adversary. Typically, this setting is met with responses such as external regret does not admit any natural interpretation when the adversary is adaptive (Arora et al., 2018). The interpretation though is clear, it reflects how much the agent would prefer to have applied the deviation to its policy under the sequence of worlds it actually found itself in; whether that is a natural interpretation seems at least debatable. Note that a similar choice is made in off-policy reinforcement learning, where the excursion setting considers the target policy's effect on future states and rewards from the distribution of states visited by the behavior policy rather than correcting the distribution to fit the target policy's distribution if it were to be followed (Sutton et al., 2016). Furthermore, there are settings where vanishing external regret implies vanishing policy regret (Arora et al., 2018), which are exactly recovered in games where this notion was first explored. Most importantly, though, this approach does not need the unknowable counterfactual.

This distinction between policy regret and deviation regret can be observed with an environment that is constructed as a two-state MDP. The actions are STAY or SWITCH, which deterministically cause their respective transition. The reward for SWITCH is always -10 while the reward for STAY is $+1$ in state 1, and $+2$ in state 2 (the initial state), and $\gamma = \frac{1}{2}$. Policy regret would compare any agent to the policy that always chooses STAY never leaving the initial state and its discounted return is 4. However, an agent that followed this policy does not guarantee no policy regret (or deviation regret for many deviations), as the adversary could just as easily set the reward for SWITCH and STAY in state 1 high enough for it to suffer linear regret. Now consider an agent that avoids this outcome via doing some degree of exploration. At some point it will end up in state 1, and once in state 1 the best policy to maximize future discounted return is to STAY forever for a return of $+2$. Policy regret would consider this a poor outcome. However, does it really make sense to look back in time and compare the agent's future behavior from state 1 to what would have been possible if it had never ever taken the SWITCH action to leave state 2? Once in state 1, the comparison should be to what can be done to maximize discounted return in the world it finds itself in. That is the heart of deviation regret. Finally, note that γ (or the evaluation horizon H) is playing a significant role in the notion of deviation regret.[1] If γ was large enough, the optimal policy would, in fact be to SWITCH back to state 2 and STAY forever. And in such a case, policy regret and deviation regret would coincide.

Deviation regret does not order agents. A desirable property of an evaluation criteria is that you can use it to order agents. We might desire to say that if $\max_{\phi \in \Phi} \rho(\phi, \lambda, e) < \max_{\phi \in \Phi} \rho(\phi, \lambda', e)$, then λ is preferred to λ' in environment e. However, this doesn't mean what it appears to mean. Agent λ likely observes a different sequence of histories, and so a different distribution of worlds, compared to λ', and as a result, it is not at all clear what it would mean to compare the deviation regret over those worlds. Notice that the above notion of policy regret allows for this kind of comparison since the comparator in the regret term does not depend on the agent at all. This is a fair objection. It does not seem possible to construct an intuitive total ordering using these criteria (however, note that it does seem possible to make an intuitive partial ordering). Deviation regret is best used to judge if an agent is adapting effectively and to do so without making assumptions on the environment (e.g., assuming the environment is a finite ergodic MDP, where effective adaptation would necessarily converge to the MDP's optimal policy). Empirical leaderboards and benchmarks may still need to resort to expected discounted return on an environment. However, that approach has its own weaknesses, particularly if we do not require ergodicity assumptions.

We can observe these different weaknesses in a simple environment where an agent must choose between LEFT or RIGHT as its first action. Suppose LEFT deterministically results in the agent playing repeated games of rock-paper-scissors against an opponent that always chooses ROCK, so that there is a simple learning problem. While RIGHT results in the agent playing repeated games of Go against a strong but imperfect opponent, so there is a challenging learning problem. The agent is completely uninformed in this decision. However, considering simple expected discounted return on this environment, an agent that defaults to choosing its first action as its first decision will most definitely outperform any agent that orders its actions differently or chooses randomly. This is true even if this alternative agent is extremely capable at learning, and manages to eventually learn to

[1] This is in contrast to the "futility of discounting in continuing problems" from Sutton & Barto (2018, p. 254), where the choice of discount factor is shown not to affect the agent's objective. The difference from our treatment is their appeal to a *stationary distribution*, which requires an ergodicity assumption on the environment we explicitly avoid. Maybe discounting in a continuing problem such as our history process is not"futile" after all?

win the majority of its games of Go. Deviation regret, instead evaluates agents by whether they are effectively adapted, relative to some set of deviations, to the worlds in which they find themselves — whether that be a simple to learn rock-paper-scissors setting or a challenging game of Go. Since the above two agents don't see the same distribution of such worlds, it makes little sense to order the agents by this criteria. However, it makes equally little sense to order them by how they make one completely uninformed decision, which would dominate any expected return assessment.

Deviation Regret encourages agents to reach a place where no learning is possible. We can avoid reaching places where no learning is possible by sublinearly increasing the evaluation horizon H. So even if the agent reaches a "no learning place", it will incur deviation regret that encourages it to change its policy and eventually get out of it. However, if no deviation policy incurs a deviation regret along the "no learning path", then the agent is doing the best it can given the world it found itself in.

6 Conclusion and Future Work

In this paper, we described four foundations of traditional RL that are antithetical to the goals of continual reinforcement learning. Further, we presented the underpinnings of an alternative set of foundations that better conceptualize the challenges faced within continual learning. More excitingly, these foundations seem to suggest a new approach to agent and algorithm design. This will also entail the development of suitable benchmark environments that embrace these alternative foundations.

Appendix

Here, we give a proof of Theorem 1 and provide an outline for the proof of Theorem 2, establishing the consistency of the estimators. The complete proof of Theorem 2 can be found in the Supplementary Materials (7). The proofs primarily rely on the concentration of the return estimates around the true returns, shown by identifying a relevant martingale and applying the Azuma-Hoeffding inequality. The above statement is true for idealized return estimates that can access future data, while in practice such data is not available. Therefore we provide an error analysis of this difference and establish that as we see more data, the contribution of this error diminishes to 0. Following these two ideas, we can prove the consistency of the H-step deviation regret estimator. The infinite discounted return case requires some additional care, which we discuss later. In the remainder of this section we introduce the relevant notation, state some basic results used, and then give the proofs.

Any behaviour $f : \mathcal{H} \to \Delta(\mathcal{A})$ in an environment e induces a distribution over trajectories. Let R_t be the random reward at timestep t. We define $G_t = \sum_{i=t}^{t+H-1} \gamma^{i-t} R_i$, the random H-step discounted return from timestep t. Furthermore, H_t denotes the random histories induced by the agent in the environment. Then, the deviation regret may be re-expresed as

$$\rho_T(\phi, \lambda, e) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[G_t | \phi(\sigma), H_{t-1}\right] - \mathbb{E}\left[G_t | \sigma, H_{t-1}\right]. \tag{3}$$

Recall, when estimating the returns in Algorithm 1, we only had access to data up to timestep T, that is, the return estimates for the last H steps are truncated. To capture these, define

$$G_t^{[T]} = \sum_{i=t}^{\min(T, t+H-1)} \gamma^{i-t} R_i \quad \text{and} \quad W_t^{[T]} = \prod_{i=t}^{\min(T, t+H-1)} \frac{\phi(\pi_i)(A_i | H_{i-1})}{\pi_i(A_i | H_{i-1})},$$

where A_t is the random action of the agent in timestep t, $G_t^{[T]}$ is the truncated agent return estimate, and ${G'}_t^{[T]} = W_t^{[T]} G_t^{[T]}$ the truncated deviation return estimate. Recall, π_i is the policy used by the agent in timestep i. The non-truncated (idealized) agent return estimate is captured by G_t, and, defining $W_t = \prod_{i=t}^{t+H-1} \frac{\phi(\pi_i)(a_i|h_{i-1})}{\pi_i(a_i|h_{i-1})}$, $G'_t = W_t G_t$ is the non-truncated (idealized) deviation return estimate. Note, an apostrophe (or prime) denotes a quantity relevant to the deviation. With this,

$$\hat{\rho}_T(\phi, \lambda, e) = \frac{1}{T}\sum_{t=1}^{T} G'^{[T]}_t - G^{[T]}_t.$$

We will argue through the idealized estimator $\hat{\rho}^*_T(\phi, \lambda, e) = \frac{1}{T}\sum_{t=1}^{T} G'_t - G_t$.

Finally, let $r^* = \max_{a\in\mathcal{A}, o\in\mathcal{O}} |R(a,o)|$, which exists since $\mathcal{A}$ and $\mathcal{O}$ are finite. Then

$$|G_t| \le \sum_{i=t}^{t+H-1} \gamma^{i-t}|R_i| \le Hr^*, \tag{4}$$

and for an agent that takes every action with probability at least c in every timestep

$$|G'_t| = |W_t|\cdot|G_t| \le \left|\prod_{i=t}^{t+H-1} \frac{\phi(\pi_i)(a_i|h_{i-1})}{\pi_i(a_i|h_{i-1})}\right| Hr^* \le c^{-H}Hr^*. \tag{5}$$

Proof of Theorem 1. Fix any ϕ, e, and λ as in the theorem statement. Let H_t be as above. For brevity, we will simply denote the deviation regret by ρ_T, and the estimates as $\hat{\rho}_T$ and $\hat{\rho}^*_T$. We decompose the estimator error as

$$|\rho_T - \hat{\rho}_T| = |\rho_T - \hat{\rho}^*_T + \hat{\rho}^*_T - \hat{\rho}_T| \le |\rho_T - \hat{\rho}^*_T| + |\hat{\rho}^*_T - \hat{\rho}_T|, \tag{6}$$

where we used the triangle inequality. The first term is the estimation error for the idealized estimate, the second term is the difference between the idealized and the truncated estimates. We bound each of these separately.

Idealized Estimator Error. Consider

$$\Delta_T = T(\rho_T - \hat{\rho}^*_T) = \sum_{t=1}^{T}(G'_t - \mathbb{E}[G_t|\phi(\sigma), H_{t-1}]) + (G_t - \mathbb{E}[G_t|\sigma, H_{t-1}]).$$

Let $\delta_t = (G'_t - \mathbb{E}[G_t|\phi(\sigma), H_{t-1}]) + (G_t - \mathbb{E}[G_t|\sigma, H_{t-1}])$, the terms in the sum. We will show that δ_t is a martingale difference sequence (MDS), hence Δ_T is a martingale. Towards this, we need $\mathbb{E}[\delta_t|\sigma, H_{t-1}] = 0$ and $|\delta_t|$ bounded. We have

$$\begin{aligned}\mathbb{E}[\delta_t|\sigma, H_{t-1}] &= \mathbb{E}\Big[(G'_t - \mathbb{E}[G_t|\phi(\sigma), H_{t-1}]) + (G_t - \mathbb{E}[G_t|\sigma, H_{t-1}]) \,\Big|\, \sigma, H_{t-1}\Big] \\ &= \mathbb{E}[G'_t|\sigma, H_{t-1}] - \mathbb{E}\Big[\mathbb{E}[G_t|\phi(\sigma), H_{t-1}] \,\Big|\, \sigma, H_{t-1}\Big] + \mathbb{E}[G_t|\sigma, H_{t-1}] - \mathbb{E}[G_t|\sigma, H_{t-1}]\end{aligned}$$

by linearity of expectation. The last two terms cancel, and now we show the first two do as well. Since $\mathbb{E}[G_t|\phi(\sigma), H_{t-1}]$ is the expected deviation return (a constant), the second term itself is the same constant. In the first term $G'_t = W_t G_t$, where W_t is the importance sampling that corrects from the agents' behaviour to the deviation's behaviour. That is, it is well known that

$$\mathbb{E}[G'_t|\sigma, H_{t-1}] = \mathbb{E}[W_t G_t|\sigma, H_{t-1}] = \mathbb{E}[G_t|\phi(\sigma), H_{t-1}],$$

and we see that the first two terms also cancel. We conclude $\mathbb{E}[\delta_t|\sigma, H_{t-1}] = 0$.

We bound $|\delta_t|$ by noting that each term in it is bounded. More precisely, we use Eq. 4, Eq. 5, the triangle inequality, and that for any $X < c$, we also have $\mathbb{E}[X|Y] < c$ for all Y.

$$|\delta_t| \le |G'_t| + |\mathbb{E}[G_t|\phi(\sigma), H_{t-1}])| + |G_t| + |\mathbb{E}[G_t|\sigma, H_{t-1}])| \le c^{-H}Hr^* + 3Hr^*.$$

Therefore, Δ_T is a martingale with the increments, δ_t, bounded by $(c^{-H}+3)Hr^*$. Note that $\Delta_0 = 0$. By the Azuma-Hoeffding inequality, we conclude that for any $\epsilon_1 > 0$ and $T \geq 1$

$$\mathbb{P}(|\Delta_T| \geq \epsilon_1) \leq 2 \exp\left(\frac{-\epsilon_1^2}{2T\big((c^{-H}+3)Hr^*\big)^2}\right).$$

Since $|\Delta_T| \geq \epsilon_1$ is equivalent to $|\rho_T - \hat{\rho}^*_T| \geq \epsilon_1/T$, letting $\epsilon_2 = \epsilon_1/T$ we can restate the above as

$$\mathbb{P}(|\rho_T - \hat{\rho}^*_T| \geq \epsilon_2) \leq 2\exp\left(\frac{-(\epsilon_2 T)^2}{2T\big((c^{-H}+3)Hr^*\big)^2}\right) = 2\exp\left(\frac{-\epsilon_2^2 T}{2\big((c^{-H}+3)Hr^*\big)^2}\right). \tag{7}$$

We see that the idealized estimate gets close to the true deviation at an exponential rate. This completes the bound on the first error term.

Error due to truncating the estimate. By definition and the triangle inequality

$$T|\hat{\rho}^*_T - \hat{\rho}_T| = \left|\sum_{t=1}^{T}(G'^{[T]}_t - G^{[T]}_t) - (G'_t - G_t)\right| \leq \sum_{t=1}^{T}|G'^{[T]}_t - G'_t| + |G^{[T]}_t - G_t|.$$

Both $\sum_{t=1}^{T}|G'^{[T]}_t - G'_t|$ and $\sum_{t=1}^{T}|G^{[T]}_t - G_t|$ are bounded by r^*H^2. It is only the last H terms that are truncated, therefore all other terms in the sum are 0. Each of the non-zero terms are no more than H-step returns of rewards no more than r^*. Plugging this into the inequality we developed so far, we find

$$|\hat{\rho}^*_T - \hat{\rho}_T| \leq \frac{2r^*H}{T(1-\gamma)}. \tag{8}$$

As this is a uniform bound over all realizations of the random estimates, for any $\epsilon_3 > 0$, choosing $T \geq \frac{2r^*H}{\epsilon_3(1-\gamma)}$, we have $|\hat{\rho}^*_T - \hat{\rho}_T| \leq \epsilon_3$.

Combining the Error Estimates. Using the error decomposition in Eq. 6, as well as the bounds Eq. 7 and Eq. 8 developed for each error component, we can conclude that for any $\epsilon > 0$, $T \geq \frac{2r^*H}{(\epsilon/2)(1-\gamma)}$,

$$\mathbb{P}(|\rho_T - \hat{\rho}_T| \leq \epsilon) \geq 1 - 2\exp\left(\frac{-(\epsilon/2)^2 T}{2(c^{-H}+3)Hr^*}\right).$$

Letting $T \to \infty$, we conclude with the result we set out to prove,

$$\lim_{T\to\infty} \mathbb{P}(|\rho_T - \hat{\rho}_T| \leq \varepsilon) = 1. \qquad \square$$

The infinite discounted deviation return case. The proof is analogous to the H-step return case, with only two changes.

1. The errors in the truncated estimates are larger, albeit still bounded and independent of T.
2. The importance sampling weights become unbounded unless treated carefully.

The first change is straightforward. For the second change, we note that if we only estimate the deviation regret up to some $\delta > 0$ accuracy, we can truncate to a finite return of sufficient length, thus controlling the importance sampling weights. See Supplementary Materials (7) for details.

References

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

David Abel, Mark K. Ho, and Anna Harutyunyan. Three dogmas of reinforcement learning. *Reinforcement Learning Journal*, 2024b.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of International Conference on Machine Learning (ICML)*, 2012.

Raman Arora, Michael Dinitz, Teodor V. Marinov, and Mehryar Mohri. Policy regret in repeated games. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 2013.

Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 2020.

Michael Bowling, John D. Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *Proceedings of the Fortieth International Conference on Machine Learning (ICML)*, pp. 3003–3020, 2023.

Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Simple agent, complex environment: Efficient reinforcement learning with agent states. *Journal of Machine Learning Research*, 2022.

Khurram Javed and Richard S. Sutton. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 2022.

Saurabh Kumar, Henrik Marklund, Ashish Rao, Yifan Zhu, Hong Jun Jeon, Yueyang Liu, and Benjamin Van Roy. Continual learning as computationally constrained reinforcement learning. *arXiv preprint arXiv:2307.04345*, 2023.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Kirkeby Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 2017.

Dustin Morrill, Ryan D'Orazio, Marc Lanctot, Reca Sarfati, James Wright, Amy Greenwald, and Michael Bowling. Efficient deviation types and learning for hindsight rationality in extensive-form games. In *Proceedings of the Thirty-Eighth International Conference on Machine Learning (ICML)*, pp. 7818–7828, 2021a.

Dustin Morrill, Ryan D'Orazio, Reca Sarfati, Marc Lanctot, James R. Wright, Amy Greenwald, and Michael Bowling. Hindsight and sequential rationality of correlated play. In *Proceedings of the Thirty-Fifth Conference on Artificial Intelligence (AAAI)*, 2021b.

Dustin Morrill, Amy R. Greenwald, and Michael Bowling. The partially observable history process. In *The AAAI Workshop on Reinforcement Learning in Games*, 2022.

Emmanouil Antonios Platanios, Abulhair Saparov, and Tom Mitchell. Jelly bean world: A testbed for never-ending learning. In *International Conference on Learning Representations*, 2023.

Mark Bishop Ring. *Continual learning in reinforcement environments*. The University of Texas at Austin, 1994.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2018.

Richard S. Sutton, Anna Koop, and David Silver. On the role of tracking in stationary environments. In *International Conference on Machine Learning*, 2007.

Richard S. Sutton, A. Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 17:1–29, 2016.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems*, 2012.

Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, 2022.

Supplementary Materials

The following content was not necessarily subject to peer review.

7 There is a consistent estimator for the infinite discounted deviation regret

For the proof of Theorem 2, we follow the ideas outlined in Appendix. More precisely, we will define

- $\rho_{T,\delta}$, a truncated version of ρ_T that is still δ close to the deviation regret, for any $\delta > 0$.
- $\hat{\rho}^*_{T,\delta}$, an idealized estimator of the truncated regret that can access future data.

We argue that $\hat{\rho}^*_{T,\delta}$ estimates $\rho_{T,\delta}$ arbitrarily well with high probability. However, $\hat{\rho}^*_{T,\delta}$ uses future data, while a real estimator, $\hat{\rho}_{T,\delta}$ does not have access to this, resulting in an additional error. Our analysis shows that with increasing data, this error is also driven to be arbitrarily small. Finally, requiring better and better estimates over time by setting $\delta = \delta(T)$ for some $\delta(T)$ that goes to 0 with time, we arrive at our final estimator $\hat{\rho}_T = \hat{\rho}_{T,\delta(T)}$. Note that $\hat{\rho}_T$ is not exactly Algorithm 1 with $H = \infty$, as $\hat{\rho}_T$ may not immediately incorporate a new observed reward into all of its deviation return estimates.

Now that we gave an outline of the proof, we introduce all required notation. All the terms not explicitly introduced here use the definitions provided earlier. In this section, we denote by $\mathbb{P}^\lambda$ the probability measure on the trajectories induced by the agent-environment interaction, and by $\mathbb{E}^\lambda$ the corresponding expectation operator. Similarly, $\mathbb{P}^\phi$ denotes the probability measure induced by applying the deviation and $\mathbb{E}^\phi$ its expectation operator. Effectively, $\mathbb{E}^\lambda$ replaces $\mathbb{E}[\,\cdot\,|\sigma]$ and $\mathbb{E}^\phi$ replaces $\mathbb{E}[\,\cdot\,|\phi(\sigma)]$.

We define both infinite and H-step returns.

$$G_t^\infty = \sum_{i=t}^{\infty} \gamma^{i-t} R_i,$$

$$G_t^{\{H\}} = \sum_{i=t}^{t+H-1} \gamma^{i-t} R_i.$$

With this notation, the deviation regret with infinite returns is

$$\rho_T(\pi, \lambda, e) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}^\pi[G_t^\infty | H_{t-1}] - \mathbb{E}^\lambda[G_t^\infty | H_{t-1}],$$

where H_{t-1} is the random history of the agent. These quantities are bounded for $\gamma < 1$.

Fact 1. *$|G_t^\infty| \leq \frac{1}{1-\gamma} r^*$ and therefore $|\rho_T| \leq \frac{2}{1-\gamma} r^*$. Also, as before, $|G_t^{\{H\}}| \leq H r^*$.*

We want to truncate returns, while staying close to the true values. Towards this, let, for any $\delta > 0, \gamma \in [0, 1)$, let

$$H(\delta, \gamma) = \left\lceil \frac{\ln\left(\frac{r^*}{\delta(1-\gamma)}\right)}{1-\gamma} \right\rceil,$$

the *effective horizon*. Choosing $H \geq H(\delta, \gamma)$ guarantees

$$|G_t^\infty - G_t^{\{H\}}| < \delta.$$

Define the truncated regret for any $\delta > 0$ as

$$\rho_{T,\delta}(\pi, \lambda, e) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}^\pi[G_t^{\{H(\delta,\gamma)\}} | H_{t-1}] - \mathbb{E}^\lambda[G_t^\infty | H_{t-1}].$$

Then, by construction,

$$|\rho_T(\pi, \lambda, e) - \rho_{T,\delta}(\pi, \lambda, e)| < \delta. \tag{9}$$

To estimate the (truncated) deviation return we use importance sampling, that is

$$G_t'^{\{H\}} = W_{t:t+H-1} G_t^{\{H\}} \quad \text{with} \quad W_{t:i} = \prod_{k=t}^{i} \frac{\pi(A_k|H_{k-1})}{\lambda(A_k|H_{k-1})}. \tag{10}$$

For an agent that takes every action in every step with probability at least $c > 0$, $W_{t:t+H-1} \leq c^{-H}$. With this, we can bound the deviation regret estimate.

Fact 2. $|G_t'^{\{H\}}| \leq c^{-H} H r^*$.

At this point, we can define $\hat{\rho}^*_{T,\delta}$, the idealized estimator of the truncated regret,

$$\hat{\rho}^*_{T,\delta}(\phi, \lambda, e) = \frac{1}{T} \sum_{t=1}^{T} G_t'^{\{H(\delta,\gamma)\}} - G_t^{\infty}. \tag{11}$$

However, for the practical estimator of regret at timestep T all estimates will naturally be truncated at step T, that is

$$\hat{\rho}_{T,\delta}(\phi, \lambda, e) = \frac{1}{T} \sum_{t=1}^{T} G_t'^{\{\min(T-t+1, H(\delta,\gamma))\}} - G_t^{\{T-t+1\}}. \tag{12}$$

As stated in the proof outline, we will choose a $\delta(T)$ such that $\delta(T)$ decreases to 0 in the limit as $T \to \infty$, and use $\hat{\rho}_T = \hat{\rho}_{T,\delta(T)}$. The particular choice we make is

$$\delta(T) = T^{-1/|4\ln(c)|}$$

With this, we are ready to provide a proof for Theorem 2.

Proof. When the deviation ϕ, agent λ, and environment e are clear from context, they are omitted from the notation. We have seen that for any $\delta > 0$, $|\rho_T - \rho_{T,\delta}| < \delta$. We will show that for any $\epsilon > 0$,

$$\mathbb{P}^{\lambda}(|\rho_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| \leq \epsilon) \geq 1 - f(T, \epsilon), \tag{13}$$

for some f with $f(T, \epsilon) \to 0$ as $T \to \infty$. We will refer to the event

$$E = \{|\rho_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| \leq \epsilon\}$$

as the "good event". For any $\epsilon > 0$, for large enough T such that $\delta(T) < \epsilon$, on the good event, we have

$$\begin{aligned} |\rho_T - \hat{\rho}_{T,\delta(T)}| &= |(\rho_T - \rho_{T,\delta(T)}) + (\rho_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)})| \\ &\leq |\rho_T - \rho_{T,\delta(T)}| + |\rho_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| \\ &\leq \epsilon + \epsilon \\ &= 2\epsilon, \end{aligned}$$

which in turn shows that for any $\epsilon > 0$

$$\lim_{T \to \infty} \mathbb{P}^{\lambda}(|\rho_T - \hat{\rho}_{T,\delta(T)}| \leq \epsilon) = 1,$$

the statement we set out to prove.

We focus on the estimation problem in Eq. 13 for the rest of the proof. As we did in the H-step deviation regret estimation case, we argue through the idealized estimator $\hat{\rho}^*_{T,\delta(T)}$. We decompose the estimator error as

$$\begin{aligned}|\rho_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| &= |\rho_{T,\delta(T)} - \hat{\rho}^*_{T,\delta(T)} + \hat{\rho}^*_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| \\ &\leq |\rho_{T,\delta(T)} - \hat{\rho}^*_{T,\delta(T)}| + |\hat{\rho}^*_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}|, \end{aligned} \tag{14}$$

where we used the triangle inequality. The first term is the estimation error for the idealized estimate, the second term is the difference between the idealized and the truncated estimates. We bound each of these separately.

Idealized Estimator Error. We will use the shorthand $H = H(\delta, \gamma)$. Consider

$$\begin{aligned}\Delta_T &= T(\hat{\rho}^*_{T,\delta(T)} - \rho_{T,\delta(T)}) \\ &= \sum_{t=1}^{T} (G'^{\{H\}}_t - \mathbb{E}^\phi[G^{\{H\}}_t | H_{t-1}]) + (G^\infty_t - \mathbb{E}^\lambda[G^\infty_t | H_{t-1}]),\end{aligned}$$

where we regrouped the terms to capture the difference between the true returns and their random estimates. Let $\delta_t = (G'^{\{H\}}_t - \mathbb{E}^\phi[G^{\{H\}}_t | H_{t-1}]) + (G^\infty_t - \mathbb{E}^\lambda[G^\infty_t | H_{t-1}])$, the terms in the sum. We will show that δ_t is a martingale difference sequence (MDS), hence Δ_T is a martingale. Towards this, we need $\mathbb{E}^\lambda[\delta_t | H_{t-1}] = 0$ and $|\delta_t|$ bounded. We have

$$\begin{aligned}\mathbb{E}^\lambda[\delta_t | H_{t-1}] &= \mathbb{E}^\lambda\left[(G'^{\{H\}}_t - \mathbb{E}^\phi[G^{\{H\}}_t | H_{t-1}]) + (G^\infty_t - \mathbb{E}^\lambda[G^\infty_t | H_{t-1}]) \,\middle|\, H_{t-1}\right] \\ &= \mathbb{E}^\lambda[G'^{\{H\}}_t | H_{t-1}] - \mathbb{E}^\lambda\left[\mathbb{E}^\phi[G^{\{H\}}_t | H_{t-1}] \,\middle|\, H_{t-1}\right] + \mathbb{E}^\lambda[G^\infty_t | H_{t-1}] - \mathbb{E}^\lambda[G^\infty_t | H_{t-1}]\end{aligned}$$

by linearity of expectation. The last two terms cancel and now we show the first two do as well. Since $\mathbb{E}^\phi[G^{\{H\}}_t | H_{t-1}]$ is the expected deviation return (a constant), the second term itself is the same constant. In the first term $G'_t = W_{t:t+H-1} G^{\{H\}}_t$, where $W_{t:t+H-1}$ is the importance sampling that corrects from the agents' behaviour to the deviation's behaviour. That is, it is well known that

$$\mathbb{E}^\lambda[G'_t | H_{t-1}] = \mathbb{E}^\lambda[W_{t:t+H-1} G^{\{H\}}_t | H_{t-1}] = \mathbb{E}^\phi[G^{\{H\}}_t | H_{t-1}],$$

and we see that the first two terms also cancel. We conclude $\mathbb{E}^\lambda[\delta_t | H_{t-1}] = 0$.

We bound $|\delta_t|$ by noting that each term in it is bounded. More precisely, we use Facts 1 and 2 bounding the individual terms, the triangle inequality, and that for any $X < c$, we also have $\mathbb{E}[X|Y] < c$ for all Y.

$$\begin{aligned}|\delta_t| &\leq |G'^{\{H\}}_t| + |\mathbb{E}^\phi[G^{\{H\}}_t | H_{t-1}])| + |G^\infty_t| + |\mathbb{E}^\lambda[G^\infty_t | H_{t-1}])| \\ &\leq c^{-H} H r^* + H r^* + \frac{2}{1-\gamma} r^*.\end{aligned}$$

Therefore, Δ_T is a martingale with the increments, δ_t, bounded by

$$b(c, H, \gamma, r^*) = (c^{-H} H + H + 2(1-\gamma)^{-1}) r^*.$$

Note that $\Delta_0 = 0$. By the Azuma-Hoeffding inequality, we conclude that for any $\epsilon_1 > 0$ and $T \geq 1$

$$\mathbb{P}^\lambda(|\Delta_T| \geq \epsilon_1) \leq 2 \exp\left(\frac{-\epsilon_1^2}{2T b(c, H, \gamma, r^*)^2}\right).$$

Since $|\Delta_T| \geq \epsilon_1$ is equivalent to $|\rho_{T,\delta(T)} - \hat{\rho}^*_{T,\delta(T)}| \geq \epsilon_1 / T$, letting $\epsilon_2 = \epsilon_1 / T$ we can restate the above as

$$\mathbb{P}^\lambda(|\rho_{T,\delta(T)} - \hat{\rho}^*_{T,\delta(T)}| \geq \epsilon_2) \leq 2 \exp\left(\frac{-(\epsilon_2 T)^2}{2T b(c, H, \gamma, r^*)^2}\right) = 2 \exp\left(\frac{-\epsilon_2^2 T}{2 b(c, H, \gamma, r^*)^2}\right). \tag{15}$$

Note that $H \in \Theta(\ln(1/\delta))$, not considering its dependence on r^*, γ. Furthermore, $\delta(T) = T^{-1/|4\ln(c)|}$. This makes $H \approx \ln(T^{1/|4\ln(c)|}) = \frac{\ln T}{4|\ln(c)|}$, and we have for the denominator in the exponent in Eq. 15 that

$$\begin{aligned}2b(c,H,\gamma,r^*)^2 &= 2(c^{-H}H + H + 2(1-\gamma)^{-1})^2 {r^*}^2 \\ &\leq 6(c^{-2H}H^2 + H^2 + 4(1-\gamma)^{-2}){r^*}^2 \\ &\approx 6\left((c^{-2\frac{\ln T}{4|\ln(c)|}} + 1)\left(\frac{\ln T}{4|\ln(c)|}\right)^2 + 4(1-\gamma)^{-2}\right){r^*}^2.\end{aligned}$$

Here, using $a^{\log_b(x)} = x^{\log_b(a)}$,

$$c^{-2\frac{\ln T}{4|\ln(c)|}} = \exp\left(\frac{\ln T}{2\ln(c)} \cdot \ln c\right) = \exp\left(\frac{\ln T}{2}\right) = T^{\frac{1}{2}},$$

so

$$\begin{aligned}2b(c,H,\gamma,r^*)^2 &\leq C_0\left((T^{\frac{1}{2}} + 1)\left(\frac{\ln T}{4|\ln(c)|}\right)^2 + 4(1-\gamma)^{-2}\right){r^*}^2 \\ &\leq C_1\left(T^{\frac{3}{4}}|2\ln(c)|^{-2} + 4(1-\gamma)^{-2}\right){r^*}^2 \\ &\leq C_1 T^{3/4}(|2\ln(c)|^{-2} + 4(1-\gamma)^{-2}){r^*}^2,\end{aligned}$$

for some constants $C_i \in \mathbb{R}$ and for sufficiently large T. We used that $\ln T \leq C_2 T^{1/8}$ for large enough T and some $C_2 > 0$, and that $T^{3/4} \geq 1$. Plugging this back into Eq. 15 we find

$$\mathbb{P}^\lambda(|\rho_{T,\delta(T)} - \hat{\rho}^*_{T,\delta(T)}| \geq \epsilon_2) \leq 2\exp\left(\frac{-\epsilon_2^2 T^{1/4}}{C_1\left(|2\ln(c)|^{-2} + 4(1-\gamma)^{-2}\right){r^*}^2}\right). \tag{16}$$

We see that the idealized estimate gets close to the true deviation with increased interaction time T. This completes the bound on the first error term.

Error due to truncation. We turn our attention to the second term of Eq. 14, $|\hat{\rho}^*_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}|$. The analysis will use δ for $\delta(T)$ and H for $H(\delta,\gamma)$ except when the dependence on the arguments is important. By definition and some algebra,

$$\begin{aligned}T|\hat{\rho}^*_{T,\delta} - \hat{\rho}_{T,\delta}| &= \left|\sum_{t=1}^{T}\left({G'_t}^{\{H\}} - G_t^\infty\right) - \left({G'_t}^{\{\min(T-t+1,H)\}} - G_t^{\{T-t+1\}}\right)\right| \\ &= \left|\sum_{t=1}^{T}\left({G'_t}^{\{H\}} - {G'_t}^{\{\min(T-t+1,H)\}}\right) - \left(G_t^\infty - G_t^{\{T-t+1\}}\right)\right| \\ &= \left|\sum_{t=1}^{T}\left({G'_t}^{\{H\}} - {G'_t}^{\{\min(T-t+1,H)\}}\right) - \gamma^{T-t+1}G_{T+1}^\infty\right| \\ &\leq \left|\sum_{t=1}^{T}\left({G'_t}^{\{H\}} - {G'_t}^{\{\min(T-t+1,H)\}}\right)\right| + \left|\sum_{t=1}^{T}\gamma^{T-t+1}G_{T+1}^\infty\right|,\end{aligned}$$

where in the last step we used the triangle inequality. The second term is bounded as

$$\left|\sum_{t=1}^{T}\gamma^{T-t+1}G_{T+1}^\infty\right| \leq \frac{r^*}{1-\gamma}\sum_{t=1}^{T}\gamma^{T-t+1} \leq \frac{r^*}{(1-\gamma)^2}.$$

For the first term, when $H \leq T-t+1$, the difference is 0. We continue with the other case, that is, $\min(T-t+1,H) = T-t+1$.

$$\left|\sum_{t=1}^{T}{G'_t}^{\{H\}} - {G'_t}^{\{T-t+1\}}\right| = \left|\sum_{t=1}^{T}\sum_{i=t}^{t+H-1}\gamma^{i-t}W_{t:i}R_i - \sum_{i=t}^{T}\gamma^{i-t}W_{t:i}R_i\right|$$

$$
\begin{aligned}
&= \left| \sum_{t=1}^{T} \sum_{i=T+1}^{t+H-1} \gamma^{i-t} W_{t:i} R_i \right| \\
&\leq \sum_{t=1}^{T} \sum_{i=T+1}^{t+H-1} \gamma^{i-t} W_{t:i}\, r^* \\
&\leq \sum_{t=1}^{T} \sum_{i=T+1}^{t+H-1} \gamma^{i-t} \left(\prod_{k=t}^{i} \frac{1}{c} \right) r^* \\
&= \sum_{t=T-H+1}^{T} \sum_{i=T+1}^{t+H-1} \gamma^{i-t} c^{-(i-t+1)}\, r^* \\
&\leq c^{-1} \sum_{t=T-H+1}^{T} \sum_{i=T+1}^{T+H-1} \gamma^{i-t} c^{-(i-t)}\, r^* \\
&\leq c^{-1} \sum_{t=T-H+1}^{T} \sum_{h=0}^{2H} (\gamma/c)^h\, r^* \\
&= r^* c^{-1} H \sum_{h=0}^{2H} (\gamma/c)^h .
\end{aligned}
$$

We control $f(H, \gamma/c) := \sum_{h=0}^{2H} (\gamma/c)^h$ dependent on where γ/c lands compared to 1.

- If $\gamma/c < 1$, $f(H, \gamma/c) \leq (1 - \gamma/c)^{-1}$.
- If $\gamma/c = 1$, $f(H, \gamma/c) \leq 2H + 1$.
- If $\gamma/c > 1$, $f(H, \gamma/c) \leq \frac{(\gamma/c)^{2H+1} - 1}{\gamma/c - 1}$.

In conclusion, for any H, δ, T

$$T |\hat{\rho}^*_{T,\delta} - \hat{\rho}_{T,\delta}| \leq \frac{r^*}{(1-\gamma)^2} + r^* c^{-1} H f(H, \gamma/c).$$

We now make explicit the dependency of H on T through $\delta(T)$, while continuing to suppress the dependency of H on γ and r^*. As before, for $\delta(T) = T^{-1/|4 \ln(c)|}$ we have $H \approx \ln(T^{1/|4 \ln(c)|}) = \frac{\ln T}{4 |\ln(c)|}$ and

$$T |\hat{\rho}^*_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| \lesssim \frac{r^*}{(1-\gamma)^2} + r^* c^{-1} \frac{\ln T}{4 |\ln(c)|} f\left(\frac{\ln T}{4 |\ln(c)|}, \gamma/c \right),$$

where f scales the worst in H for the $\gamma/c > 1$ case. In this setting,

$$f(H, \gamma/c) \leq \frac{(\gamma/c)^{2H+1} - 1}{\gamma/c - 1} \leq \frac{(\gamma/c)\, T^{\frac{\ln(\gamma/c)}{2|\ln(c)|}} - 1}{\gamma/c - 1},$$

where, noting the range of c and γ, we see $\frac{\ln(\gamma/c)}{2|\ln(c)|} = \frac{-\ln(c)}{2(-\ln(c))} + \frac{\ln(\gamma)}{-2\ln(c)} \leq 1/2$, since the second term is negative. That is, we are guaranteed that $|\hat{\rho}^*_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| \in O(T^{-1/2})$. With this, we are ready to finish up the proof.

The Estimation Problem. We originally set out to analyze the truncated regret estimation problem introduced in Eq. 13, through the error decomposition of Eq. 14. We now provided a bound for each of the error terms and we are ready to combine them for the desired result. Eq. 14 stated that

$$|\rho_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| \leq |\rho_{T,\delta(T)} - \hat{\rho}^*_{T,\delta(T)}| + |\hat{\rho}^*_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}|.$$

We chose $\delta(T) = T^{-1/|4\ln(c)|}$, which indeed approaches 0 as $T \to \infty$. Then, we saw that for all $\epsilon_2 > 0$ there exist constants C_0, T_0 (independent of ϵ_2) such that for all $T > T_0$

$$\mathbb{P}^\lambda(|\rho_{T,\delta(T)} - \hat{\rho}^*_{T,\delta(T)}| \geq \epsilon_2) \leq 2\exp\left(\frac{-\epsilon_2^2 T^{1/4}}{C_0\left(|2\ln(c)|^{-2} + 4(1-\gamma)^{-2}\right) r^{*2}}\right).$$

Finally, we saw that there exists f_0 and T_1 such that for any $T \geq T_1$

$$|\hat{\rho}^*_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| \leq \frac{1}{\sqrt{T}} f_0(r^*, \gamma, c).$$

We can conclude that

$$|\rho_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| \leq |\rho_{T,\delta(T)} - \hat{\rho}^*_{T,\delta(T)}| + \frac{1}{\sqrt{T}} f_0(r^*, \gamma, c).$$

Then, $\forall \varepsilon > 0$, choosing $T \geq \max(T_0, T_1, (f_0(r^*, \gamma, c)/\varepsilon)^2)$, we have $\varepsilon \geq f_0(r^*, \gamma, c)/\sqrt{T}$, and

$$\begin{aligned}
\mathbb{P}^\lambda\left(|\rho_{T,\delta(T)} - \hat{\rho}_{T,\delta(T)}| \geq 2\varepsilon\right) &\leq \mathbb{P}^\lambda\left(|\rho_{T,\delta(T)} - \hat{\rho}^*_{T,\delta(T)}| + f_0(r^*, \gamma, c)/\sqrt{T} \geq 2\varepsilon\right) \\
&\leq \mathbb{P}^\lambda\left(|\rho_{T,\delta(T)} - \hat{\rho}^*_{T,\delta(T)}| \geq \varepsilon\right) \\
&\leq 2\exp\left(\frac{-\varepsilon^2 T^{1/4}}{C_0\left(|2\ln(c)|^{-2} + 4(1-\gamma)^{-2}\right) r^{*2}}\right).
\end{aligned}$$

This bound indeed goes to 0 as $T \to \infty$, so the proof is complete. □