On Temporal Credit Assignment and Data-Efficient Reinforcement Learning

Dilip Arumugam[1], **Thomas L. Griffiths**[1,2]
dilip.a@cs.princeton.edu, tomg@princeton.edu

[1]**Department of Computer Science, Princeton University**
[2]**Department of Psychology, Princeton University**

Abstract

The challenge of temporal credit assignment in reinforcement learning (RL) can be articulated as a simple question about the behavior of a sequential decision-making agent: *how does the execution of particular actions from specific states impact observed future outcomes?* Typically, one asks this question for each state-action pair sampled along a full trajectory within the environment and the future outcome of interest is the cumulative return obtained by an agent. Temporal credit assignment stands as the defining challenge of the RL paradigm, distinguishing it from supervised learning and bandit learning settings, where the data-efficiency challenges of generalization and exploration also arise. Nevertheless, a precise and formal characterization of the credit assignment problem remains elusive. In this work, we make an initial effort to formally define the credit assignment problem through the introduction of a performance measure for RL algorithms, quantifying the overall accuracy of credit attribution (or lack thereof) between the policies generated by an RL algorithm and the optimal policy. To define this novel performance criterion, we draw upon foundational information-theoretic and game-theoretic tools for the partial decomposition of information and the allocation of group compensation among individual team members.

Twenty-seven lawyers in the room, anybody know '*post hoc, ergo propter hoc*?'

President Josiah Bartlet, The West Wing
Season 1 – Episode 2: Post Hoc, Ergo Propter Hoc

1 Introduction

Ultimately, it is President Bartlet who provides the answer to his own question and translates the Latin *post hoc, ergo propter hoc*: "after it, therefore because of it. It means one thing follows the other, therefore it was caused by the other, but it's not always true. In fact, it's hardly ever true" (Sorkin & Schlamme, 1999). Fundamental to the reinforcement-learning (RL) problem (Sutton & Barto, 1998; Kaelbling et al., 1996; Littman, 2015) is the challenge of temporal credit assignment (Minsky, 1961; Sutton, 1984), wherein an agent strives to understand the impact of individual steps of behavior on temporally-delayed future outcomes. Historically, the classic mechanisms for addressing credit assignment in RL are temporal-difference (TD) learning and eligibility traces (Sutton, 1988; Klopf, 1972; Sutton, 1984). Unfortunately, the heuristic codified within these seminal methods quintessentially embodies a *post hoc, ergo propter hoc* philosophy as temporal recency governs which states and actions are credited for the occurrence of unexpected outcomes.

It is perhaps not too difficult to envision a sequential decision-making problem in which recency-based credit assignment paves the way for inefficient learning. The Behavior Suite for RL (Osband et al., 2019) offers one such example as part of the `bsuite` unit tests evaluating how well an RL agent copes with the challenge of credit assignment. A so-called "umbrella problem" MDP is given

where (metaphorically) the agent observes a forecast from which it may elect to pick up an umbrella at the very first timestep; after proceeding with a long-horizon task (modeled as a N-state chain) where all decisions made are inconsequential, the agent arrives at a terminal state where it may be raining. In the face of rain, a prepared agent opting to have left the initial state with the umbrella is given positive reward whereas a damp agent without the umbrella is given a negative reward. As the free evaluation parameter N increases, it is natural to expect standard TD-methods struggle with uncovering the correct relationship between the very first action and terminal reward.

We posit that progress on addressing temporal credit assignment problem in RL has stalled due to a lack of clarity and a missing formal articulation of the credit assignment problem. As we point out, this is in contrast to generalization and exploration, where clear performance measures for RL algorithms exist and admit notions of statistical efficiency. To resolve this gap in the literature, we offer a first take on what statistically-efficient credit assignment means through the introduction of an information-theoretic performance measure for RL algorithms, quantifying the overall accuracy of credit attribution (or lack thereof) for the policies generated by an RL algorithm relative to the optimal policy.

2 Problem Formulation

For any arbitrary set $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the set of all probability distributions with support on $\mathcal{X}$. For any arbitrary set A, we use $\mathcal{P}(A)$ to denote the power set of A. For any $N \in \mathbb{N}$, we denote the index set as $[N] = \{1, 2, \ldots, N\}$.

We formulate a sequential decision-making problem as a finite-horizon, episodic Markov Decision Process (MDP) (Bellman, 1957; Puterman, 1994) defined by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \beta, H \rangle \in \mathfrak{M}$. $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a reward function providing evaluative feedback in the unit interval, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition function prescribing distributions over next states, $\beta \in \Delta(\mathcal{S})$ is an initial state distribution, and $H \in \mathbb{N}$ is the maximum episode length or horizon. Within each of $K \in \mathbb{N}$ total episodes, the agent acts for H steps beginning with an initial state $s_1 \sim \beta(\cdot)$ and, at each timestep $h \in [H]$, observes the current state $s_h \in \mathcal{S}$, selects an action $a_h \in \mathcal{A}$, enjoys a reward $r_h = \mathcal{R}(s_h, a_h)$, and transitions to a next state $s_{h+1} \sim \mathcal{T}(\cdot \mid s_h, a_h)$.

An agent is characterized by its non-stationary, stochastic policy $\pi : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$, which encodes a pattern of behavior by mapping individual states and the current timestep to a probability distribution over actions. We assess the performance of a policy π in MDP $\mathcal{M}$ at timestep $h \in [H]$ when starting at state $s \in \mathcal{S}$ and taking action $a \in \mathcal{A}$ by its associated action-value function $Q^{\pi}_{\mathcal{M},h}(s, a) = \mathbb{E}\left[\sum_{h'=h}^{H} \mathcal{R}(s_{h'}, a_{h'}) \;\middle|\; s_h = s, a_h = a\right]$. Taking the value function as $V^{\pi}_{\mathcal{M},h}(s) = \mathbb{E}_{a \sim \pi_h(\cdot \mid s)}\left[Q^{\pi}_{\mathcal{M},h}(s, a)\right]$, we define the optimal policy $\pi^{\star}$ as achieving supremal value $V^{\star}_{\mathcal{M},h}(s) = \sup_{\pi \in \Pi} V^{\pi}_{\mathcal{M},h}(s)$ for all $s \in \mathcal{S}$, $h \in [H]$ where Π denotes the class of all non-stationary, stochastic policies. For any episode $k \in [K]$, we let $\tau_k = (s_1^{(k)}, a_1^{(k)}, \ldots, s_H^{(k)}, a_H^{(k)}, s_{H+1}^{(k)}) \sim \rho^{\pi^{(k)}}$ denote the random trajectory experienced by the agent executing its policy $\pi^{(k)}$ in the environment. Meanwhile, $H_k = \{\tau_1, \tau_2, \ldots, \tau_{k-1}\} \in \mathcal{H}$ is the entire random history of agent interaction at the start of the kth episode.

3 Toward Statistically-Efficient Temporal Credit Assignment

In this section, we begin by offering some context that leads to the introduction of a new performance criterion for RL algorithms. We then proceed with the definition of our performance measure for assessing RL agents solely along the axis of temporal credit assignment. As our performance measure depends on solutions to an information-theoretic sub-problem, we dedicate subsequent sections to discussing the sub-problem itself as well as our proposed game-theoretic resolution. Notably, our

solution concept ties back to our performance criterion, giving a precise intuition for what is being measured and credence to the resulting notion of statistically-efficient credit assignment.

3.1 Motivation

The primary contribution of this paper is a novel performance measure for RL algorithms. Abstractly, any RL algorithm can be represented via a sequence of policies $\{\pi^{(k)}\}_{k\in[K]}$ applied to a MDP across K episodes, where the policy deployed at each episode $\pi^{(k)}$ is a function of the current history H_k. Thus, any measure of a RL algorithm's proficiency is some abstract function $\Lambda : \Pi^K \times \mathfrak{M} \to \mathbb{R}_{\geq 0}$ whereby the numerical value $\Lambda(\{\pi^{(k)}\}_{k\in[K]}, \mathcal{M})$ offers some quantitative evaluation of an RL algorithm with respect to MDP $\mathcal{M} \in \mathfrak{M}$.

The RL literature already offers a number of established choices for the performance criterion Λ. We begin with brief and informal overviews of two popular types of theoretical analysis in RL (value-loss and PAC-MDP sample complexity) before giving a more formal presentation of a third performance criterion (cumulative regret) to juxtapose against our novel performance measure in the following sub-section. Value-loss analyses (Porteus, 1971; 1975; Singh & Yee, 1994) prove upper bounds on the difference between the optimal value function $V^{\star}_{\mathcal{M},1}$ and the value function $V^{\overline{\pi}}_{\mathcal{M},1}$ induced by some well-chosen alternative policy $\overline{\pi}$. Thus, if one explicitly provides (or even simply posits the existence of) an algorithm to obtain $\overline{\pi}$ within K episodes, a value-loss analysis implies a Λ only focused on the final policy $\pi^{(K)}$ without regard for the full learning trajectory. The value-loss bound itself establishes an upper bound for this particular Λ, guaranteeing approximately-optimal behavior at termination. The classic simulation lemma (Kearns & Singh, 2002; Lobel & Parr, 2024) is one such result, which provides the theoretical foundation for model-based RL by ensuring that a policy obtained via an approximately-precise model of the true MDP yields approximately-optimal performance in the true MDP. In the context of data-efficient RL, one notable use of value-loss bounds is to establish the virtues of state abstraction (Li et al., 2006; Abel, 2020) in addressing generalization, where the alternative policy of interest is the optimal policy obtained under some aggregated or otherwise compressed state space (Tsitsiklis & Van Roy, 1996; Van Roy, 2006; Abel et al., 2016; 2018; 2019; Arumugam & Singh, 2022; Turner et al., 2025).

The sample complexity of a RL algorithm (Kakade, 2003) is defined as the total number of timesteps for which the policy employed by a RL algorithm is worse than ε-suboptimal, for an arbitrary choice of $\varepsilon > 0$; mirroring the seminal Probably Approximately Correct (PAC) learning framework of Valiant (1984), PAC-MDP analyses establish that, for any $\varepsilon, \delta > 0$, the sample complexity of a RL algorithm can be upper bounded with probability at least $1 - \delta$ by a polynomial in ε^{-1}, δ^{-1}, and the relevant MDP quantities (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002; Kakade, 2003; Strehl et al., 2006; Strehl & Littman, 2008; Strehl et al., 2009; Lattimore & Hutter, 2012; Dann & Brunskill, 2015). Just as with value-loss bounds, PAC-MDP algorithms implicitly engage with and upper bound a performance measure Λ that sums a binary indicator function $\mathbb{1}(V^{\pi^{(k)}} < V^{\star}_{\mathcal{M},1} - \varepsilon)$ across all episodes denoting whether each policy in $\{\pi^{(k)}\}_{k\in[K]}$ is worse than ε-optimal. Unlike the generalization focus of value-loss bounds, PAC-MDP analysis is primarily oriented around assessing data efficiency through strategic exploration. That said, various works have succeeded in leveraging such analysis techniques to simultaneously tackle exploration and generalization (Kakade et al., 2003; Pazis & Parr, 2016; Krishnamurthy et al., 2016; Jiang et al., 2017; Du et al., 2019; 2021; Jin et al., 2021). Moreover, such PAC-MDP analysis techniques are extensible to the Bayesian RL setting (Kolter & Ng, 2009).

As a point of contrast with our performance criterion that quantifies the cumulative misallocation of credit by an RL algorithm relative to the optimal policy of a MDP, we give a slightly more detailed view into another well-established performance criterion: the cumulative regret of a RL algorithm. Regret measures the total expected performance shortfall between an agent's chosen policy and the optimal policy across all episodes: $\textsc{Regret}(\{\pi^{(k)}\}_{k\in[K]}, \mathcal{M}) = \mathbb{E}\left[\sum_{k=1}^{K}\left(V^{\star}_{\mathcal{M},1}(s_1) - V^{\pi^{(k)}}_{\mathcal{M},1}(s_1)\right)\right]$. It is inevitable that an agent must incur some amount of regret

in order to explore the environment and synthesize optimal behavior. Thus, cumulative regret essentially quantifies the speed with which an RL algorithm solves the exploration challenge.

The Bayesian RL setting (Bellman & Kalaba, 1959; Duff, 2002; Ghavamzadeh et al., 2015) recognizes that the underlying MDP is entirely unknown to the agent and, therefore, a random variable. The agent is thus endowed with a prior distribution $\mathbb{P}(\mathcal{M} \in \cdot)$ to reflect initial uncertainty in the true MDP. An alternative performance criterion considerate of this reality is the Bayesian regret, which simply integrates out the randomness in $\mathcal{M}$ with respect to an agent's (well-specified) prior: $\textsc{BayesRegret}(\{\pi^{(k)}\}_{k\in[K]}) = \mathbb{E}\left[\textsc{Regret}(\{\pi^{(k)}\}_{k\in[K]}, \mathcal{M})\right]$.

A broad range of RL algorithms have been shown to yield statistically-efficient exploration through frequentist (Agrawal & Jia, 2017; Dann et al., 2017; Jin et al., 2018) and/or Bayesian (Osband et al., 2013; Osband & Van Roy, 2014; Abbasi-Yadkori & Szepesvari, 2014; Ouyang et al., 2017; Osband & Van Roy, 2017; Lu & Van Roy, 2019; Arumugam & Van Roy, 2022; Lu et al., 2023) regret analysis. Similar to PAC-MDP analyses, a number of works also find additional opportunities to integrate consideration for generalization as well; for instance, via state abstraction (Dong et al., 2019) or the Eluder dimension (Russo & Van Roy, 2013; Wen & Van Roy, 2013; Osband & Van Roy, 2014; Wang et al., 2020b; Huang et al., 2021; Li et al., 2022).

A surprising observation is that none of the aforementioned analyses seem to engage with the challenge of temporal credit assignment in an explicit, tangible manner. That is, to the best of the authors' knowledge, there is no specific point in any of the aforementioned analyses that one could substantively point to as "handling" or "accounting for" how an RL algorithm in question deals with credit assignment. One plausible takeaway from this realization is that, perhaps, temporal credit assignment is not central to statistically-efficient RL. While not an impossibility, such a hypothesis would be counterintuitive, especially at a moment in time when one of the primary applications of RL involves selecting hundreds upon thousands of actions (tokens) only to be met with a binary terminal reward (Stiennon et al., 2020; Ouyang et al., 2022); arguably, the challenge of temporal credit assignment is more pivotal now than ever before in the history of RL. An alternative takeaway is that all the aforementioned performance criteria are so widely used because, at some point, one or more RL researchers merely decided those measures captured a dimension of efficient decision-making that matters and was worth studying. Consequently, if the current issue is that we lack a formal and coherent performance measure for evaluating the efficacy of credit assignment, then perhaps it is simply incumbent upon the RL community to make one. It is in this spirit that we here introduce the *misallocation* of a RL algorithm.

3.2 Defining Misallocation

Paralleling regret as an evaluation metric for RL algorithms, our proposed performance criterion will be a cumulative discrepancy across all episodes between the optimal MDP policy $\pi^\star$ and the episode policy $\pi^{(k)}$. Unlike regret, however, the discrepancy measured will not be with respect to performance via the value function induced by each policy. Instead, we will define an information-theoretic statistic for each policy that quantifies how much each step of behavior impacts the cumulative return. Under the premise that the optimal policy represents the ideal treatment of the credit assignment problem, it then follows that any deviation from the statistic computed for the optimal policy constitutes an error, analogous to episodic regret, which our performance criterion will accumulate. We begin by defining this statistic for capturing the credit assigned by a fixed policy about the cumulative return to individual steps of behavior. For brief background on information theory, please see Appendix A.

For a fixed policy within a MDP, a key observation is that the assignment of credit or blame represents an instance of a more fundamental statistical challenge: decomposing the influence or information that one collection of random variables exert over a single target random variable. Any policy π in MDP $\mathcal{M}$ has an associated distribution over trajectories ρ^π, for which we let $\tau_\pi \sim \rho^\pi$ be a random variable denoting a single trajectory obtained by executing π in $\mathcal{M}$. Similar to the literature

on distributional RL (Bellemare et al., 2017), we use $Z(\tau_\pi) = \sum_{h=1}^{H} \mathcal{R}(s_h, a_h)$ to denote the random cumulative return obtained with trajectory τ_π. Intuitively, this return random variable can be related back to the classic value function induced by π via an expectation: $V^\pi_{\mathcal{M},1}(s) = \mathbb{E}\left[Z(\tau_\pi) \mid s_1 = s\right]$. One way to formalize the total influence of agent behavior on observed returns is via the mutual information $\mathbb{I}(Z(\tau_\pi); \tau_\pi)$ between a random trajectory generated by the policy and the random cumulative return. However this quantity, while comprehensive, is monolithic and blurs the contributions of individual steps of behavior and overall performance. The trajectory is itself a collection of state-action valued random variables $\tau = (s_1, a_1, \ldots, s_H, a_H, s_{H+1})$ and, for brevity, we may define one random variable $\xi^\pi_h = (s_h, a_h) \in \mathcal{S} \times \mathcal{A}$ for each timestep $h \in [H]$[1] (whose distribution is obtained by marginalizing ρ^π).

If all $\mathbb{I}(Z(\tau_\pi); \xi^\pi_1, \ldots, \xi^\pi_H)$ bits of information could be decomposed on a per-timestep basis, it would yield an accurate profile of how the constituent states and decisions in τ_π from policy π impact the cumulative trajectory return $Z(\tau_\pi)$. This problem of *partial information decomposition* (PID) (Williams & Beer, 2010) has been studied for many years in the information-theory literature. Solving a PID problem instance is itself an open research challenge that continues to be an active area of study in recent years (Bertschinger et al., 2013; 2014; Griffith & Koch, 2014; Timme et al., 2014; Griffith & Ho, 2015; Olbrich et al., 2015; Banerjee & Griffith, 2015; James & Crutchfield, 2017; Lizier et al., 2018; Ay et al., 2021; Kolchinsky, 2022; Venkatesh et al., 2023; Kolchinsky, 2024; Murphy & Bassett, 2024). For the moment, we defer the finer details of PID and its solution so that we may proceed with the definition of our performance measure. In the next sub-section, we present a particular method for addressing PID such that a corresponding solution yields a natural interpretation in the context of RL and credit assignment.

We define a black-box function $\texttt{PID} : \Pi \times \mathfrak{M} \to \mathbb{R}^H_{\geq 0}$ that computes any partial information decomposition, such that each component $h \in [H]$ of the vector $\texttt{PID}(\pi, \mathcal{M}) \in \mathbb{R}^H_{\geq 0}$ quantifies the average dependence (measured in bits of information) between the random state-action pair ξ^π_h visited by policy π at timestep h and the cumulative return $Z(\tau_\pi)$ obtained. Assuming we can solve any PID problem instance, we are able to introduce our novel performance criterion for assessing how well an RL algorithm addresses the temporal credit assignment problem. We define the cumulative misallocation of an RL algorithm, or MALLOC for short, as

$$\textsc{Malloc}(\{\pi^{(k)}\}_{k\in[K]}, \mathcal{M}) = \mathbb{E}\left[\sum_{k=1}^{K} ||\texttt{PID}(\pi^\star, \mathcal{M}) - \texttt{PID}(\pi^{(k)}, \mathcal{M})||_1\right].$$

If the optimal policy yields a particular dependency structure between visited states; selected actions; and optimal returns, then it is natural to interpret good credit assignment as expediently arriving at policies that adhere to a similar (if not identical) structure. The PID solution of each policy $\pi^{(k)}$ with respect to MDP $\mathcal{M}$ gives a statistic that encodes this dependency structure quantitatively, while the misallocation simply accumulates the discrepancy between those structures and that of the optimal policy. As the mapping between policies and value functions is many-to-one (Dadashi et al., 2019), a more technically-correct definition of misallocation might (charitably) benchmark with respect to nearest optimal policy (formally, taking an infimum over all optimal policies $\pi^\star \in \Pi^\star \triangleq \{\pi \in \Pi \mid ||V^\star_{\mathcal{M},1} - V^\pi_{\mathcal{M},1}||_\infty \leq 0\}$), though we omit this more verbose definition for clarity.

Just as the (frequentist) regret presumes full knowledge of the underlying MDP transition function and reward function, so too does our misallocation criterion; adopting the Bayesian RL setting and mirroring the Bayesian regret, we may analogously define the Bayesian misallocation as $\textsc{BayesMalloc}(\{\pi^{(k)}\}_{k\in[K]}) = \mathbb{E}\left[\textsc{Malloc}(\{\pi^{(k)}\}_{k\in[K]}, \mathcal{M})\right]$. In the next sub-section, we examine fundamental tools from game theory to elucidate the partial information decomposition problem and its resulting solution used within the definition of misallocation.

[1] Note this drops the terminal state s_H which, due to the dependence of rewards solely on the state-action pair $\mathcal{R}(s, a)$, is irrelevant. Transition-based rewards $\mathcal{R}(s, a, s')$ can be accounted for by adding a ξ_{H+1} random variable with a null action.

3.3 Partial Information Decomposition

The previous section establishes misallocation as an information-theoretic measure for assessing the efficacy of an RL algorithm in addressing temporal credit assignment, provided access to an oracle solution for PID problems. In this section, we clarify how and why the PID between a policy's cumulative return $Z(\tau_\pi)$ and state-action pairs visited in each timestep $\{\xi_h^\pi\}_{h\in[H]}$ conveys something meaningful and quantitatively substantive about assigning credit or blame to behavior. For the ease of exposition, it is natural to begin with restricted focus on the simplest case of $H = 2$, yielding the smallest possible temporal delay between behavior and outcomes.

With a horizon $H = 2$, the total information or influence between states and actions of any policy π and cumulative returns is given by the mutual information: $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi)$. The chain rule of mutual information yields two identities for decomposing this global information term:

$$\mathbb{I}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi) = \mathbb{I}(Z(\tau_\pi); \xi_1^\pi) + \mathbb{I}(Z(\tau_\pi); \xi_2^\pi \mid \xi_1^\pi) = \mathbb{I}(Z(\tau_\pi); \xi_2^\pi) + \mathbb{I}(Z(\tau_\pi); \xi_1^\pi \mid \xi_2^\pi).$$

One might wonder why such a decomposition into a (more-refined) mutual information and conditional mutual information terms fails to provide a solution to PID. While this certainly is one kind of decomposition, the fidelity of this decomposition to the individual contributions of ξ_1^π and ξ_2^π is entirely dependent on the relationship between either $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi)$ and $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi \mid \xi_2^\pi)$ or $\mathbb{I}(Z(\tau_\pi); \xi_2^\pi)$ and $\mathbb{I}(Z(\tau_\pi); \xi_2^\pi \mid \xi_1^\pi)$.

In the case when either $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi) \geq \mathbb{I}(Z(\tau_\pi); \xi_1^\pi \mid \xi_2^\pi)$ or $\mathbb{I}(Z(\tau_\pi); \xi_2^\pi) \geq \mathbb{I}(Z(\tau_\pi); \xi_2^\pi \mid \xi_1^\pi)$, we may upper bound the chain rule expansion above and obtain $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi) \leq \mathbb{I}(Z(\tau_\pi); \xi_1^\pi) + \mathbb{I}(Z(\tau_\pi); \xi_2^\pi)$. This inequality conveys that, while the cumulative information or influence between behavior and returns is $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi)$, a subset of those bits are redundantly present in both ξ_1^π and ξ_2^π individually. Consequently, aggregating information via $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi) + \mathbb{I}(Z(\tau_\pi); \xi_2^\pi)$ ends up double counting these redundant bits. In the alternative case when either $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi) \leq \mathbb{I}(Z(\tau_\pi); \xi_1^\pi \mid \xi_2^\pi)$ or $\mathbb{I}(Z(\tau_\pi); \xi_2^\pi) \leq \mathbb{I}(Z(\tau_\pi); \xi_2^\pi \mid \xi_1^\pi)$, we may lower bound the chain rule expansion above and obtain $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi) \geq \mathbb{I}(Z(\tau_\pi); \xi_1^\pi) + \mathbb{I}(Z(\tau_\pi); \xi_2^\pi)$. This inequality conveys that the combined information offered individually by ξ_1^π and ξ_2^π in isolation is insufficient to account for the overall information between behavior and returns $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi)$. This implies that the residual, unaccounted bits of information are only holistically accessible through the combination of (ξ_1^π, ξ_2^π) and cannot otherwise be obtained without this synergy.

Recognizing these deficiencies in classic information-theoretic quantities, pioneering work by Williams & Beer (2010) introduced new information-theoretic quantities at the requisite level of granularity needed for PID. In the context of the $H = 2$ case, there is the *unique information* $\mathcal{U}(Z(\tau_\pi); \xi_1^\pi \mid \xi_2^\pi)$ provided by ξ_1^π about $Z(\tau_\pi)$ that is not offered by ξ_2^π; conversely, we also have the unique information $\mathcal{U}(Z(\tau_\pi); \xi_2^\pi \mid \xi_1^\pi)$ provided by ξ_2^π about $Z(\tau_\pi)$ that is not offered by ξ_1^π. There is the *redundant information* $\mathcal{R}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi)$ identically provided by each of ξ_1^π and ξ_2^π about $Z(\tau_\pi)$. Finally, there is the *synergistic information* $\mathcal{S}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi)$ jointly provided by ξ_1^π and ξ_2^π about $Z(\tau_\pi)$ that cannot be obtained from either one in isolation. As the above chain rule decompositions illustrate, standard mutual information blends distinct types of information

$$\begin{aligned}
\mathbb{I}(Z(\tau_\pi); \xi_1^\pi) &= \mathcal{U}(Z(\tau_\pi); \xi_1^\pi \mid \xi_2^\pi) + \mathcal{R}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi) \\
\mathbb{I}(Z(\tau_\pi); \xi_2^\pi \mid \xi_1^\pi) &= \mathcal{U}(Z(\tau_\pi); \xi_2^\pi \mid \xi_1^\pi) + \mathcal{S}(Z(\tau_\pi); \xi_2^\pi, \xi_1^\pi),
\end{aligned}$$

with analogous equations holding *mutatis mutandis* for $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi)$ and $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi \mid \xi_2^\pi)$. In short, research into PID aims to provide a definition for one of the three more-granular information quantities (Bertschinger et al., 2014; Griffith & Koch, 2014; Griffith & Ho, 2015), such that the others may be obtained systematically using the previous two equations and the expansion for $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi)$. For an arbitrary horizon H, one may obtain similar decompositions of the monolithic $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi, \ldots, \xi_H^\pi)$ by recursively applying the pairwise definitions above; just as the chain rule of mutual information admits multiple decompositions of $\mathbb{I}(Z(\tau_\pi); \xi_1^\pi, \ldots, \xi_H^\pi)$, there are many equivalent decompositions into unique, redundant, and synergistic information terms.

3.4 Shapley Values for Information-Theoretic Credit Assignment

Our proposed approach for solving arbitrary instances of the PID problem is to reduce each instance to a coalition game and leverage a classic game-theoretic mechanism for equitably distributing group compensation among individual team members. A coalition is defined by a group of $H \in \mathbb{N}$ individuals and a profit function $\omega : \mathcal{P}([H]) \to \mathbb{R}$, mapping coalitions of team members to real values and where $\omega(\emptyset) = 0$. Intuitively, for any coalition of team members $A \subseteq [H]$, $\omega(A)$ quantifies the value of payoff obtained by those members of the coalition A as a result of their cooperation. A natural question arises when, after all team members cooperate together and obtain a cumulative value of $\omega([H])$, how this collective payoff should be fairly dispensed to the individual participants? A classic resolution from the economics and game-theory literatures to this credit assignment problem is prescribed by Shapley values (Shapley, 1953) which, under a fixed and known payoff function ω, distributes credit according to a function $\varphi : [H] \to \mathbb{R}$ defined as

$$\varphi(h) = \frac{1}{H} \sum_{A \subseteq [H] \setminus \{h\}} \binom{H-1}{|A|}^{-1} \left(\omega(A \cup \{h\}) - \omega(A)\right), \qquad \forall h \in [H].$$

In words, the Shapley value associated with an individual $h \in [H]$ averages the contributions of the individual h across all possible teams that could be formed without h from the remaining members of $[H] \setminus \{h\}$. While this intuition is convenient, widespread adoption of Shapley values as a solution to this problem largely stems from its *unique* adherence (Dubey, 1975) to several desiderata that one ought to naturally demand from any measure of credit attribution. Due to space constraints, we defer a review of these desiderata to Appendix B and proceed with leveraging Shapley values to solve PID problems.

We reduce any instance of PID $\mathbb{I}(Z(\tau_\pi); \{\xi_h^\pi\}_{h \in H})$ for temporal credit assignment to a coalition game with H players and payoff function $\omega(A) = \mathbb{I}(Z(\tau_\pi); \{\xi_{h'}^\pi\}_{h' \in A})$, for any $A \in \mathcal{P}([H])$. In doing so, we then obtain a solution for PID via the individual Shapley values associated with each timestep: $\texttt{PID}(\pi, \mathcal{M}) = [\varphi^\pi(1), \ldots, \varphi^\pi(H)] \in \mathbb{R}^H$. While we are not the first to establish a connection between these ideas (Ay et al., 2021), we do maintain fidelity to the original definition of Shapley values (without imposing constraints on the ordering/hierarchy of random variables that call for an alternative generalization of Shapley values (Faigle & Kern, 1992)) and, to the best of our knowledge, are the first to bring these ideas to bear on temporal credit assignment in single-agent RL. Crucially, the definition of Shapley values given above then yields a natural and interpretable meaning to the individual components of the $\texttt{PID}(\pi, \mathcal{M})$ vector in terms of unique, redundant, and synergistic information (please see Appendix D for the proof).

Proposition 1. *For any policy π; MDP $\mathcal{M}$; and timestep $h \in [H]$, define $\xi_{-h}^\pi \triangleq \{\xi_{h'}^\pi\}_{h' \in [H] \setminus \{h\}}$ and, for any subset $A \subseteq [H]$, $\xi_A^\pi \triangleq \{\xi_{h'}^\pi\}_{h' \in A}$. The Shapley values for PID at timestep h satisfy*

$$\varphi^\pi(h) = \mathcal{U}(Z(\tau_\pi); \xi_h^\pi \mid \xi_{-h}^\pi) + H^{-1} \left(\mathcal{R}(Z(\tau_\pi); \xi_h^\pi, \xi_{-h}^\pi) + \sum_{A \subseteq [H] \setminus \{h\}} \binom{H-1}{|A|}^{-1} \mathcal{S}(Z(\tau_\pi); \xi_h^\pi, \xi_A^\pi) \right).$$

Proposition 1 highlights that, according to the desiderata uniquely satisfied by Shapley values for coalition games, the credit assigned to the states and actions visited by policy π and timestep h for cumulative returns $Z(\tau_\pi)$ is equal to the unique information provided by behavior at that timestep (not contained anywhere else in the trajectory) as well as equitable portions of the redundant and synergistic information provided by behavior at h and other timesteps along the trajectories generated by π. For the simple $H = 2$ case, this yields similar results for each timestep, with $h = 1$ as $\varphi(1) = \mathcal{U}(Z(\tau_\pi); \xi_1^\pi \mid \xi_2^\pi) + \frac{1}{2}\left(\mathcal{R}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi) + \mathcal{S}(Z(\tau_\pi); \xi_1^\pi, \xi_2^\pi)\right)$.

We have made a first attempt at a formal articulation of the temporal credit assignment problem via the cumulative misallocation of a RL algorithm. Through a game-theoretic solution for defining this information-theoretic performance criterion, we offer one notion of what it means for a RL algorithm to achieve statistically-efficient credit assignment.

A Information Theory

Here we introduce various concepts in probability theory and information theory (Shannon, 1948) used throughout this paper. We encourage readers to consult (Cover & Thomas, 2012; Gray, 2011; Polyanskiy & Wu, 2022; Duchi, 2024) for more background.

All random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We define the mutual information between any two random variables X, Y through the Kullback-Leibler (KL) divergence:

$$\mathbb{I}(X;Y) = D_{\mathrm{KL}}(\mathbb{P}((X,Y) \in \cdot) \,||\, \mathbb{P}(X \in \cdot) \times \mathbb{P}(Y \in \cdot)) \qquad D_{\mathrm{KL}}(P \,||\, Q) = \begin{cases} \int \log\left(\frac{dP}{dQ}\right) dP & P \ll Q \\ +\infty & P \not\ll Q \end{cases},$$

where P and Q are both probability measures on the same measurable space and $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative of P with respect to Q. An analogous definition of conditional mutual information holds through the expected KL-divergence for any three random variables X, Y, Z:

$$\mathbb{I}(X;Y \mid Z) = \mathbb{E}\left[D_{\mathrm{KL}}(\mathbb{P}((X,Y) \in \cdot \mid Z) \,||\, \mathbb{P}(X \in \cdot \mid Z) \times \mathbb{P}(Y \in \cdot \mid Z))\right].$$

With these definitions in hand, we may define the entropy and conditional entropy for any two random variables X, Y as

$$\mathbb{H}(X) = \mathbb{I}(X;X) \qquad \mathbb{H}(Y \mid X) = \mathbb{H}(Y) - \mathbb{I}(X;Y).$$

This yields the following identities for mutual information and conditional mutual information for any three arbitrary random variables X, Y, and Z:

$$\begin{aligned} \mathbb{I}(X;Y) &= \mathbb{H}(X) - \mathbb{H}(X \mid Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X), \\ \mathbb{I}(X;Y|Z) &= \mathbb{H}(X|Z) - \mathbb{H}(X \mid Y, Z) = \mathbb{H}(Y|Z) - \mathbb{H}(Y|X, Z). \end{aligned}$$

Through the chain rule of the KL-divergence and the fact that $D_{\mathrm{KL}}(P \,||\, P) = 0$ for any probability measure P, we obtain another equivalent definition of mutual information,

$$\mathbb{I}(X;Y) = \mathbb{E}\left[D_{\mathrm{KL}}(\mathbb{P}(Y \in \cdot \mid X) \,||\, \mathbb{P}(Y \in \cdot))\right],$$

as well as the chain rule of mutual information: $\mathbb{I}(X;Y_1,\ldots,Y_n) = \sum_{i=1}^{n} \mathbb{I}(X;Y_i \mid Y_1,\ldots,Y_{i-1})$.

B Coalition Games & Shapley Values

Consider a team comprised of $N \in \mathbb{N}$ individuals and a profit function $\omega : \mathcal{P}([N]) \to \mathbb{R}$ mapping coalitions of team members to real values with $\omega(\emptyset) = 0$. Intuitively, for any coalition of team members $A \subseteq [N]$, $\omega(A)$ quantifies the value or payoff obtained by the members of the coalition as a result of their cooperation. A natural question arises when, after all team members cooperate together and obtain a value of $\omega([N])$, how this collective payoff should be fairly dispensed to the individual participants? A classic resolution to this credit assignment problem from the economics and game-theory literature is prescribed by Shapley values (Shapley, 1953) which, for a fixed and known payoff function ω, distributes credit according to a function $\varphi : [N] \to \mathbb{R}$ defined as

$$\varphi(i) = \frac{1}{N} \sum_{A \subseteq [N] \setminus \{i\}} \binom{N-1}{|A|}^{-1} \left(\omega(A \cup \{i\}) - \omega(A)\right), \qquad \forall i \in [N].$$

In words, the Shapley value associated with an individual $i \in [N]$ averages the contributions of the individual i across all possible teams that could be formed without i from the remaining members of $[N] \setminus \{i\}$. While this intuition is convenient, widespread adoption of Shapley values as a solution to this problem largely stems from its *unique* adherence (Dubey, 1975) to the following five desiderata that one ought to naturally demand from any measure of credit attribution.

Fact 1 (Efficiency). *For any $N \in \mathbb{N}$,*

$$\omega([N]) = \sum_{i=1}^{N} \varphi(i).$$

That is, the sum of payoffs allocated to each individual team member in $[N]$ is equal to the total payoff attributed to the entire team, $\omega([N])$.

Fact 2 (Symmetry). *If*

$$\omega(A + \{i\}) = \omega(A + \{j\}), \qquad \forall A \subseteq [N] \setminus \{i, j\},$$

then $\varphi(i) = \varphi(j)$. In words, if any two individuals $i, j \in [N]$ offer the same marginal contribution to any team formed by the remaining members of $[N]$, then their individual Shapley values will be the same.

Fact 3 (Dummy). *If*

$$\omega(A + \{i\}) = \omega(A), \qquad \forall A \subseteq [N] \setminus \{i\},$$

then $\varphi(i) = 0$. In words, an individual $i \in [N]$ whose marginal contribution to any team not including themselves is negligible accordingly has a Shapley value of zero.

Fact 4 (Additivity). *If there are two distinct payoff functions, $\omega_1, \omega_2 \in \{\mathcal{P}([N]) \to \mathbb{R}\}$, yielding Shapley values $\varphi_1, \varphi_2 \in \{[N] \to \mathbb{R}\}$, respectively, then the payoff function*

$$\omega(A) = \omega_1(A) + \omega_2(A), \qquad \forall A \subseteq [N]$$

has Shapley values

$$\varphi(i) = \varphi_1(i) + \varphi_2(i), \qquad \forall i \in [N].$$

Fact 5 (Linearity). *If a payoff function $\omega : \mathcal{P}([N]) \to \mathbb{R}$ has Shapley values $\varphi : [N] \to \mathbb{R}$, then the payoff function*

$$\omega_\alpha(A) = \alpha \cdot \omega(A), \qquad \forall A \subseteq [N]$$

has Shapley values

$$\varphi_\alpha(i) = \alpha \cdot \varphi(i), \qquad \forall i \in [N]$$

for all $\alpha \in \mathbb{R}$.

C Related Work

Tackling temporal credit assignment via TD-learning and eligibility traces (Sutton, 1988; Klopf, 1972; Sutton, 1984) was, arguably, the starting point for computational RL research. Both found widespread success in the early days of reinforcement learning, spanning empirical (Barto et al., 1983; Tesauro, 1991; 1992; Watkins & Dayan, 1992; Lin, 1992; Peng & Williams, 1994) and theoretical (Sutton, 1984; Dayan, 1992; Jaakkola et al., 1994; Tsitsiklis, 1994; Dayan & Sejnowski, 1994; Sutton & Singh, 1994; Singh & Sutton, 1996; Bradtke & Barto, 1996; Kearns & Singh, 2000) contributions. In the years since, however, a lack of holistic satisfaction with classic TD-learning and eligibility traces has given rise to numerous extensions and adaptations.

Among these is a blend of both theoretical and empirical work on extending TD-methods to off-policy learning (with or without function approximation) (Precup et al., 2000; 2001; Sutton et al., 2008; Kolter, 2011; Sutton et al., 2014; Van Hasselt et al., 2014; Seijen & Sutton, 2014; Mahmood et al., 2014; van Hasselt & Sutton, 2015; Chelu et al., 2022) and gradient-based function approximation (Maei et al., 2009; Sutton et al., 2009; Maei, 2011). Additionally, there have been similar efforts to extend eligibility traces as well (Pitis, 2018; van Hasselt et al., 2021; Gupta et al., 2024); the expected eligibility trace of van Hasselt et al. (2021) bears particular relevant to the information-theoretic perspective on credit assignment adopted in this work, by recognizing the value of assigning credit to the trajectory random variable rather than a single realization of a random trajectory.

While these methods maintain fidelity to classic TD-learning as the underlying mechanism for credit assignment, another line of work takes DQN (Mnih et al., 2015) with experience replay (Lin, 1992) as a starting point and considers extensions that yield more scalable solutions to temporal credit assignment (Schaul et al., 2016; Daley & Amato, 2019; Elelimy et al., 2025; Pignatelli et al., 2023). The overall key distinction between these works and our contributions is an adherence to classic TD-learning an making an effort to operate within the confines of what it affords; meanwhile, this work advocates for a fundamental break from these classic ideas in favor of a more general treatment of credit assignment that might clarify a precise connection with data-efficient RL.

Fundamental challenges in the core of TD-learning have already been raised in the literature. Konidaris et al. (2011) elucidate how the λ-return central to TD(λ) can be derived as a maximum-likelihood estimator of the return under three faulty assumptions that realistically never hold in practice and, along with Thomas et al. (2015), offer potential remedies via alternative weighted combinations of n-step returns. Daley et al. (2024) conduct a deep theoretical dive into the recency heuristic that underlies TD-learning and finds it to be inherent to any convex combination of n-step returns and, furthermore, find it to be inextricable from current TD-learning without risk of divergence. Perhaps the most cogent efforts to resolve the dependence on temporal recency within TD-learning comes in the form of emphatic TD (Sutton et al., 2016; Mahmood et al., 2015; Yu, 2015; Hallak et al., 2016; Jiang et al., 2021; Klissarov et al., 2022), where an interest function modulates the contributions of individual states to each TD-update. Unfortunately, this interest function is deemed to be user-specified, leaving open the question of how the saliency of states to TD-updates ought to be determined in a data-driven way to facilitate sample-efficient RL overall.

In the same spirit of this paper to challenge the core foundations of credit assignment, there are a few diverse works that appeal to a variety of alternative ideas including importance sampling (Harutyunyan et al., 2019; Velu et al., 2023), return redistribution (Arjona-Medina et al., 2019), stochastic computation graphs (Weber et al., 2019), and information theory (Arumugam et al., 2021). While the final information-theoretic lens is most related to our perspective, it only succeeds in establishing useful identities and connections between information theory and credit assignment. Overall, all of these works still leave open the question of how these connections genuinely impact data-efficient RL.

Finally, we note in passing that, while the focus of this work is exclusively on single-agent RL, multi-agent RL (Albrecht et al., 2024) has emerged as a setting where the topic of credit assignment is discussed with perhaps greater frequency (Chang et al., 2003) in the context of attributing credit among a team of cooperating agents. In this context, existing work has been done to leverage Shapley values for performing such credit assignment (Wang et al., 2020a). While this additional axis to the credit assignment problem in multi-agent RL undoubtedly has impact on data efficiency in that setting, it is orthogonal to the data efficiency concerns studied in this work. Nevertheless, one might naturally hope that clarity on the relationship between temporal credit assignment and sample efficiency in single-agent RL might yield additional promising insights for multi-agent RL as well.

D Proof of Proposition 1

Proposition 1. *For any policy π; MDP $\mathcal{M}$; and timestep $h \in [H]$, define $\xi^\pi_{-h} \triangleq \{\xi^\pi_{h'}\}_{h' \in [H] \setminus \{h\}}$ and, for any subset $A \subseteq [H]$, $\xi^\pi_A \triangleq \{\xi^\pi_{h'}\}_{h' \in A}$. The Shapley values for PID at timestep h satisfy*

$$\varphi^\pi(h) = \mathcal{U}(Z(\tau_\pi); \xi^\pi_h \mid \xi^\pi_{-h}) + H^{-1} \left(\mathcal{R}(Z(\tau_\pi); \xi^\pi_h, \xi^\pi_{-h}) + \sum_{A \subseteq [H] \setminus \{h\}} \binom{H-1}{|A|}^{-1} \mathcal{S}(Z(\tau_\pi); \xi^\pi_h, \xi^\pi_A) \right).$$

Proof. Recall that the Shapley value for a coalition game consisting of H players is defined as

$$\varphi(h) = \frac{1}{H} \sum_{A \subseteq [H] \setminus \{h\}} \binom{H-1}{|A|}^{-1} \left(\omega(A \cup \{h\}) - \omega(A)\right), \qquad \forall h \in [H].$$

Further recall that our profit function $\omega : \mathcal{P}([H]) \to \mathbb{R}$ is defined as

$$\omega(A) = \mathbb{I}(Z(\tau_\pi); \xi_A^\pi), \qquad \forall A \subseteq [H].$$

Just by substituting into the definition of the Shapley value and applying the chain rule of mutual information, we see that

$$\begin{aligned}
\varphi^\pi(h) &= \frac{1}{H} \sum_{A \subseteq [H]\setminus\{h\}} \binom{H-1}{|A|}^{-1} (\omega(A \cup \{h\}) - \omega(A)) \\
&= \frac{1}{H} \sum_{A \subseteq [H]\setminus\{h\}} \binom{H-1}{|A|}^{-1} \left(\mathbb{I}(Z(\tau_\pi); \xi_{A\cup\{h\}}^\pi) - \mathbb{I}(Z(\tau_\pi); \xi_A^\pi)\right) \\
&= \frac{1}{H} \sum_{A \subseteq [H]\setminus\{h\}} \binom{H-1}{|A|}^{-1} (\mathbb{I}(Z(\tau_\pi); \xi_A^\pi) + \mathbb{I}(Z(\tau_\pi); \xi_h^\pi \mid \xi_A^\pi) - \mathbb{I}(Z(\tau_\pi); \xi_A^\pi)) \\
&= \frac{1}{H} \sum_{A \subseteq [H]\setminus\{h\}} \binom{H-1}{|A|}^{-1} \mathbb{I}(Z(\tau_\pi); \xi_h^\pi \mid \xi_A^\pi)
\end{aligned}$$

At this point, we note that we may (equivalently) decompose the Shapley value into a sum over subsets of fixed size:

$$\varphi^\pi(h) = \frac{1}{H} \sum_{A \subseteq [H]\setminus\{h\}} \binom{H-1}{|A|}^{-1} \mathbb{I}(Z(\tau_\pi); \xi_h^\pi \mid \xi_A^\pi) = \frac{1}{H} \sum_{k=0}^{H-1} \binom{H-1}{k}^{-1} \sum_{\substack{A \subseteq [H]\setminus\{h\} \\ |A|=k}} \mathbb{I}(Z(\tau_\pi); \xi_h^\pi \mid \xi_A^\pi).$$

Pulling out the first term of the sum and applying the identity for mutual information in terms of unique and redundant information, we have

$$\begin{aligned}
\varphi^\pi(h) &= \frac{1}{H} \sum_{k=0}^{H-1} \binom{H-1}{k}^{-1} \sum_{\substack{A \subseteq [H]\setminus\{h\} \\ |A|=k}} \mathbb{I}(Z(\tau_\pi); \xi_h^\pi \mid \xi_A^\pi) \\
&= \frac{1}{H} \mathbb{I}(Z(\tau_\pi); \xi_h^\pi) + \frac{1}{H} \sum_{k=1}^{H-1} \binom{H-1}{k}^{-1} \sum_{\substack{A \subseteq [H]\setminus\{h\} \\ |A|=k}} \mathbb{I}(Z(\tau_\pi); \xi_h^\pi \mid \xi_A^\pi) \\
&= \frac{1}{H} \left(\mathcal{U}(Z(\tau_\pi); \xi_h^\pi \mid \xi_{-h}^\pi) + \mathcal{R}(Z(\tau_\pi); \xi_h^\pi, \xi_{-h}^\pi)\right) + \frac{1}{H} \sum_{k=1}^{H-1} \binom{H-1}{k}^{-1} \sum_{\substack{A \subseteq [H]\setminus\{h\} \\ |A|=k}} \mathbb{I}(Z(\tau_\pi); \xi_h^\pi \mid \xi_A^\pi).
\end{aligned}$$

Similarly, we may expand the conditional mutual information term with the identity for unique and synergistic information:

$$\frac{1}{H} \sum_{k=1}^{H-1} \binom{H-1}{k}^{-1} \sum_{\substack{A \subseteq [H]\setminus\{h\} \\ |A|=k}} \mathbb{I}(Z(\tau_\pi); \xi_h^\pi \mid \xi_A^\pi) = \frac{1}{H} \sum_{k=1}^{H-1} \binom{H-1}{k}^{-1} \sum_{\substack{A \subseteq [H]\setminus\{h\} \\ |A|=k}} \left(\mathcal{U}(Z(\tau_\pi); \xi_h^\pi \mid \xi_{-h}^\pi) + \mathcal{S}(Z(\tau_\pi); \xi_h^\pi, \xi_A^\pi)\right).$$

Focusing on the unique information term in isolation, we may simplify to obtain

$$\begin{aligned}
\frac{1}{H} \sum_{k=1}^{H-1} \binom{H-1}{k}^{-1} \sum_{\substack{A \subseteq [H]\setminus\{h\} \\ |A|=k}} \mathcal{U}(Z(\tau_\pi); \xi_h^\pi \mid \xi_{-h}^\pi) &= \mathcal{U}(Z(\tau_\pi); \xi_h^\pi \mid \xi_{-h}^\pi) \cdot \frac{1}{H} \sum_{k=1}^{H-1} \binom{H-1}{k}^{-1} \underbrace{\sum_{\substack{A \subseteq [H]\setminus\{h\} \\ |A|=k}} 1}_{=\binom{H-1}{k}} \\
&= \mathcal{U}(Z(\tau_\pi); \xi_h^\pi \mid \xi_{-h}^\pi) \cdot \frac{1}{H} \sum_{k=1}^{H-1} \binom{H-1}{k}^{-1} \cdot \binom{H-1}{k} \\
&= \mathcal{U}(Z(\tau_\pi); \xi_h^\pi \mid \xi_{-h}^\pi) \cdot \frac{1}{H} \sum_{k=1}^{H-1} 1 \\
&= \frac{H-1}{H} \mathcal{U}(Z(\tau_\pi); \xi_h^\pi \mid \xi_{-h}^\pi).
\end{aligned}$$

Putting everything together, we have

$$\varphi^{\pi}(h) = \frac{1}{H}\left(\mathcal{U}(Z(\tau_\pi);\xi_h^\pi \mid \xi_{-h}^\pi) + \mathcal{R}(Z(\tau_\pi);\xi_h^\pi,\xi_{-h}^\pi)\right) + \frac{1}{H}\sum_{k=1}^{H-1}\binom{H-1}{k}^{-1}\sum_{\substack{A\subseteq[H]\setminus\{h\}\\|A|=k}}\mathbb{I}(Z(\tau_\pi);\xi_h^\pi \mid \xi_A^\pi)$$

$$= \mathcal{U}(Z(\tau_\pi);\xi_h^\pi \mid \xi_{-h}^\pi) + \frac{1}{H}\left(\mathcal{R}(Z(\tau_\pi);\xi_h^\pi,\xi_{-h}^\pi) + \sum_{A\subseteq[H]\setminus\{h\}}\binom{H-1}{|A|}^{-1}\mathcal{S}(Z(\tau_\pi);\xi_h^\pi,\xi_A^\pi)\right).$$

□

References

Yasin Abbasi-Yadkori and Csaba Szepesvari. Bayesian Optimal Control of Smoothly Parameterized Systems: The Lazy Posterior Sampling Algorithm. *arXiv preprint arXiv:1406.3926*, 2014.

David Abel. *A Theory of Abstraction in Reinforcement Learning*. PhD thesis, Brown University, 2020.

David Abel, David Hershkowitz, and Michael Littman. Near Optimal Behavior via Approximate State Abstraction. In *International Conference on Machine Learning*, volume 48, pp. 2915–2923. PMLR, 2016.

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State Abstractions for Lifelong Reinforcement Learning. In *International Conference on Machine Learning*, pp. 10–19. PMLR, 2018.

David Abel, Dilip Arumugam, Kavosh Asadi, Yuu Jinnai, Michael L Littman, and Lawson LS Wong. State Abstraction as Compression in Apprenticeship Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3134–3142, 2019.

Shipra Agrawal and Randy Jia. Optimistic Posterior Sampling for Reinforcement Learning: Worst-Case Regret Bounds. In *Advances in Neural Information Processing Systems*, pp. 1184–1194, 2017.

Stefano V Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024.

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

Dilip Arumugam and Satinder Singh. Planning to the Information Horizon of BAMDPs via Epistemic State Abstraction. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Dilip Arumugam and Benjamin Van Roy. Deciding What to Model: Value-Equivalent Sampling for Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35:9024–9044, 2022.

Dilip Arumugam, Peter Henderson, and Pierre-Luc Bacon. An Information-Theoretic Perspective on Credit Assignment in Reinforcement Learning. *arXiv preprint arXiv:2103.06224*, 2021.

Nihat Ay, Daniel Polani, and Nathaniel Virgo. Information Decomposition Based on Cooperative Game Theory. *Kybernetika*, 56(5):979–1014, 2021.

Pradeep Kr Banerjee and Virgil Griffith. Synergy, Redundancy and Common Information. *arXiv preprint arXiv:1509.03706*, 2015.

Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems. *IEEE transactions on systems, man, and cybernetics*, pp. 834–846, 1983.

Marc G Bellemare, Will Dabney, and Rémi Munos. A Distributional Perspective on Reinforcement Learning. In *International Conference on Machine Learning*, pp. 449–458. PMLR, 2017.

Richard Bellman. A Markovian Decision Process. *Journal of Mathematics and Mechanics*, pp. 679–684, 1957.

Richard Bellman and Robert Kalaba. On Adaptive Control Processes. *IRE Transactions on Automatic Control*, 4(2):1–9, 1959.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, and Jürgen Jost. Shared Information — New Insights and Problems in Decomposing Information in Complex Systems. In *Proceedings of the European Conference on Complex Systems 2012*, pp. 251–269. Springer, 2013.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying Unique Information. *Entropy*, 16(4):2161–2183, 2014.

Steven J Bradtke and Andrew G Barto. Linear Least-Squares Algorithms for Temporal Difference Learning. *Machine Learning*, 22:33–57, 1996.

Ronen I Brafman and Moshe Tennenholtz. R-MAX - A General Polynomial Time Algorithm for Near-Optimal Reinforcement Learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Yu-Han Chang, Tracey Ho, and Leslie Kaelbling. All Learning is Local: Multi-Agent Learning in Global Reward Games. *Advances in Neural Information Processing Systems*, 16, 2003.

Veronica Chelu, Diana Borsa, Doina Precup, and Hado van Hasselt. Selective Credit Assignment. *arXiv preprint arXiv:2202.09699*, 2022.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

Robert Dadashi, Adrien Ali Taiga, Nicolas Le Roux, Dale Schuurmans, and Marc G Bellemare. The Value Function Polytope in Reinforcement Learning. In *International Conference on Machine Learning*, pp. 1486–1495. PMLR, 2019.

Brett Daley and Christopher Amato. Reconciling λ-returns with Experience Replay. *Advances in Neural Information Processing Systems*, 32, 2019.

Brett Daley, Marlos C Machado, and Martha White. Demystifying the Recency Heuristic in Temporal-Difference Learning. *arXiv preprint arXiv:2406.12284*, 2024.

Christoph Dann and Emma Brunskill. Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Peter Dayan. The Convergence of TD(λ) for General λ. *Machine Learning*, 8:341–362, 1992.

Peter Dayan and Terrence J Sejnowski. TD(λ) Converges with Probability 1. *Machine Learning*, 14:295–301, 1994.

Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Provably Efficient Reinforcement Learning with Aggregated States. *arXiv preprint arXiv:1912.06366*, 2019.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably Efficient RL with Rich Observations via Latent State Decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear Classes: A Structural Framework for Provable Generalization in RL. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

Pradeep Dubey. On the Uniqueness of the Shapley Value. *International Journal of Game Theory*, 4 (3):131–139, 1975.

John C. Duchi. *Lecture Notes for Statistics 311/Electrical Engineering 377*. Stanford University, 2024. URL https://web.stanford.edu/class/stats311/lecture-notes.pdf.

Michael O'Gordon Duff. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts Amherst, 2002.

Esraa Elelimy, Brett Daley, Andrew Patterson, Marlos C Machado, Adam White, and Martha White. Deep Reinforcement Learning with Gradient Eligibility Traces. In *Reinforcement Learning Conference*, 2025.

Ulrich Faigle and Walter Kern. The Shapley Value for Cooperative Games Under Precedence Constraints. *International Journal of Game Theory*, 21:249–266, 1992.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian Reinforcement Learning: A Survey. *Foundations and Trends in Machine Learning*, 8(5-6):359–483, 2015.

Robert M. Gray. *Entropy and Information Theory*. Springer Science & Business Media, 2011.

Virgil Griffith and Tracey Ho. Quantifying Redundant Information in Predicting a Target Random Variable. *Entropy*, 17(7):4644–4653, 2015.

Virgil Griffith and Christof Koch. Quantifying Synergistic Mutual Information. In *Guided Self-Organization: Inception*, pp. 159–190. Springer, 2014.

Dhawal Gupta, Scott M Jordan, Shreyas Chaudhari, Bo Liu, Philip S Thomas, and Bruno Castro da Silva. From Past to Future: Rethinking Eligibility Traces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12253–12260, 2024.

Assaf Hallak, Aviv Tamar, Rémi Munos, and Shie Mannor. Generalized Emphatic Temporal Difference Learning: Bias-Variance Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, et al. Hindsight Credit Assignment. *Advances in Neural Information Processing Systems*, 32, 2019.

Kaixuan Huang, Sham M Kakade, Jason D Lee, and Qi Lei. A Short Note on the Relationship of Information Gain and Eluder Dimension. *arXiv preprint arXiv:2107.02377*, 2021.

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. On the Convergence of Stochastic Iterative Dynamic Programming Algorithms. *Neural Computation*, 6(6):1185–1201, 1994.

Ryan G James and James P Crutchfield. Multivariate Dependence Beyond Shannon Information. *Entropy*, 19(10):531, 2017.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual Decision Processes with Low Bellman Rank are PAC-Learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Ray Jiang, Tom Zahavy, Zhongwen Xu, Adam White, Matteo Hessel, Charles Blundell, and Hado Van Hasselt. Emphatic Algorithms for Deep Reinforcement Learning. In *International Conference on Machine Learning*, pp. 5023–5033. PMLR, 2021.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-Learning Provably Efficient? *Advances in Neural Information Processing Systems*, 31, 2018.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms. *Advances in Neural Information Processing Systems*, 34:13406–13418, 2021.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Sham Kakade, Michael J Kearns, and John Langford. Exploration in Metric State Spaces. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 306–312, 2003.

Sham Machandranath Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University of London, University College London (United Kingdom), 2003.

Michael Kearns and Satinder Singh. Near-Optimal Reinforcement Learning in Polynomial Time. *Machine Learning*, 49:209–232, 2002.

Michael J Kearns and Satinder P Singh. Bias-Variance Error Bounds for Temporal Difference Updates. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pp. 142–147, 2000.

Martin Klissarov, Rasool Fakoor, Jonas W Mueller, Kavosh Asadi, Taesup Kim, and Alexander J Smola. Adaptive Interest for Emphatic Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35:95–108, 2022.

A Harry Klopf. *Brain Function and Adaptive Systems: A Heterostatic Theory*. Air Force Cambridge Research Laboratories, Air Force Systems Command, United States Air Force, 1972.

Artemy Kolchinsky. A Novel Approach to the Partial Information Decomposition. *Entropy*, 24(3): 403, 2022.

Artemy Kolchinsky. Partial Information Decomposition as Information Bottleneck. *arXiv e-prints*, pp. arXiv–2405, 2024.

J Kolter. The Fixed Points of Off-Policy TD. *Advances in Neural Information Processing Systems*, 24, 2011.

J Zico Kolter and Andrew Y Ng. Near-Bayesian Exploration in Polynomial Time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 513–520, 2009.

George Konidaris, Scott Niekum, and Philip S Thomas. TD_γ: Re-Evaluating Complex Backups in Temporal Difference Learning. *Advances in Neural Information Processing Systems*, 24, 2011.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC Reinforcement Learning with Rich Observations. *Advances in Neural Information Processing Systems*, 29, 2016.

Tor Lattimore and Marcus Hutter. PAC Bounds for Discounted MDPs. In *23rd International Conference on Algorithmic Learning Theory*, pp. 320–334. Springer, 2012.

Gene Li, Pritish Kamath, Dylan J Foster, and Nati Srebro. Understanding the Eluder Dimension. *Advances in Neural Information Processing Systems*, 35:23737–23750, 2022.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a Unified Theory of State Abstraction for MDPs. In *International Symposium on Artificial Intelligence and Mathematics*, volume 4, pp. 5, 2006.

Long-Ji Lin. Self-Improving Reactive Agents Based on Reinforcement Learning, Planning and Teaching. *Machine Learning*, 8:293–321, 1992.

Michael L Littman. Reinforcement Learning Improves Behaviour From Evaluative Feedback. *Nature*, 521(7553):445–451, 2015.

Joseph T Lizier, Nils Bertschinger, Jürgen Jost, and Michael Wibral. Information Decomposition of Target Effects From Multi-Source Interactions: Perspectives on Previous, Current and Future Work, 2018.

Sam Lobel and Ronald Parr. An optimal tightness bound for the simulation lemma. In *Reinforcement Learning Conference*, 2024.

Xiuyuan Lu and Benjamin Van Roy. Information-Theoretic Confidence Bounds for Reinforcement Learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, and Zheng Wen. Reinforcement Learning, Bit by Bit. *Foundations and Trends in Machine Learning*, 16(6): 733–865, 2023.

Hamid Maei, Csaba Szepesvari, Shalabh Bhatnagar, Doina Precup, David Silver, and Richard S Sutton. Convergent Temporal-Difference Learning with Arbitrary Smooth Function Approximation. *Advances in Neural Information Processing Systems*, 22, 2009.

Hamid Reza Maei. *Gradient Temporal-Difference Learning Algorithms*. PhD thesis, University of Alberta, 2011.

A Rupam Mahmood, Hado P Van Hasselt, and Richard S Sutton. Weighted Importance Sampling for Off-Policy Learning with Linear Function Approximation. *Advances in Neural Information Processing Systems*, 27, 2014.

A Rupam Mahmood, Huizhen Yu, Martha White, and Richard S Sutton. Emphatic Temporal-Difference Learning. *arXiv preprint arXiv:1507.01569*, 2015.

Marvin Minsky. Steps Toward Artificial Intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-Level Control Through Deep Reinforcement Learning. *nature*, 518(7540):529–533, 2015.

Kieran A Murphy and Dani S Bassett. Information Decomposition in Complex Systems via Machine Learning. *Proceedings of the National Academy of Sciences*, 121(13):e2312988121, 2024.

Eckehard Olbrich, Nils Bertschinger, and Johannes Rauh. Information Decomposition and Synergy. *Entropy*, 17(5):3501–3517, 2015.

Ian Osband and Benjamin Van Roy. Model-Based Reinforcement Learning and the Eluder Dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

Ian Osband and Benjamin Van Roy. Why is Posterior Sampling Better than Optimism for Reinforcement Learning? In *International Conference on Machine Learning*, pp. 2701–2710, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) Efficient Reinforcement Learning via Posterior Sampling. *Advances in Neural Information Processing Systems*, 26:3003–3011, 2013.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvari, Satinder Singh, et al. Behaviour Suite for Reinforcement Learning. *arXiv preprint arXiv:1908.03568*, 2019.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training Language Models to Follow Instructions with Human Feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning Unknown Markov Decision Processes: A Thompson Sampling Approach. *Advances in Neural Information Processing Systems*, 30, 2017.

Jason Pazis and Ronald Parr. Efficient PAC-Optimal Exploration in Concurrent, Continuous State MDPs with Delayed Updates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Jing Peng and Ronald J Williams. Incremental Multi-Step Q-Learning. In *Machine Learning Proceedings 1994*, pp. 226–232. Elsevier, 1994.

Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, Olivier Pietquin, and Laura Toni. A Survey of Temporal Credit Assignment in Deep Reinforcement Learning. *arXiv preprint arXiv:2312.01072*, 2023.

Silviu Pitis. Source Traces for Temporal Difference Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Learning to Coding*. Cambridge University Press, 2022.

Evan L Porteus. Some Bounds for Discounted Sequential Decision Processes. *Management Science*, 18(1):7–11, 1971.

Evan L Porteus. Bounds and Transformations for Discounted Finite Markov Decision Chains. *Operations Research*, 23(4):761–784, 1975.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility Traces for Off-Policy Policy Evaluation. In *ICML*, volume 2000, pp. 759–766. Citeseer, 2000.

Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-Policy Temporal-Difference Learning with Function Approximation. In *ICML*, pp. 417–424, 2001.

Martin L. Puterman. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.

Daniel Russo and Benjamin Van Roy. Eluder Dimension and the Sample Complexity of Optimistic Exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized exerience replay. In *International Conference on Learning Representations*, 2016.

Harm Seijen and Rich Sutton. True Online TD(λ). In *International Conference on Machine Learning*, pp. 692–700. PMLR, 2014.

Claude E Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Lloyd S Shapley. A Value for n-Person Games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, Princeton, 1953.

Satinder P Singh and Richard S Sutton. Reinforcement Learning with Replacing Eligibility Traces. *Machine learning*, 22(1-3):123–158, 1996.

Satinder P Singh and Richard C Yee. An Upper Bound on the Loss from Approximate Optimal-Value Functions. *Machine Learning*, 16:227–233, 1994.

Aaron Sorkin and Thomas Schlamme. The West Wing – Post Hoc, Ergo Propter Hoc, 1999.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to Summarize with Human Feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Alexander L Strehl and Michael L Littman. An Analysis of Model-Based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. PAC Model-Free Reinforcement Learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 881–888, 2006.

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement Learning in Finite MDPs: PAC Analysis. *Journal of Machine Learning Research*, 10(11), 2009.

Rich Sutton, Ashique Rupam Mahmood, Doina Precup, and Hado Hasselt. A New Q(λ) with Interim Forward View and Monte Carlo Equivalence. In *International Conference on Machine Learning*, pp. 568–576. PMLR, 2014.

Richard S. Sutton. *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 1984.

Richard S Sutton. Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, 3(1):9–44, 1988.

Richard S Sutton and Andrew G Barto. *Introduction to Reinforcement Learning*. MIT Press, 1998.

Richard S Sutton and Satinder P Singh. On Step-Size and Bias in Temporal-Difference Learning. In *Proceedings of the Eighth Yale Workshop on Adaptive and Learning Systems*, 1994.

Richard S Sutton, Hamid Maei, and Csaba Szepesvári. A Convergent $o(n)$ Temporal-Difference Algorithm for Off-Policy Learning with Linear Function Approximation. *Advances in Neural Information Processing Systems*, 21, 2008.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 993–1000, 2009.

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(73): 1–29, 2016.

Gerald Tesauro. Practical Issues in Temporal Difference Learning. *Advances in Neural Information Processing Systems*, 4, 1991.

Gerald Tesauro. Temporal Difference Learning of Backgammon Strategy. In *Machine Learning Proceedings 1992*, pp. 451–457. Elsevier, 1992.

Philip S Thomas, Scott Niekum, Georgios Theocharous, and George Konidaris. Policy Evaluation Using the Ω-Return. *Advances in Neural Information Processing Systems*, 28, 2015.

Nicholas Timme, Wesley Alford, Benjamin Flecker, and John M Beggs. Synergy, Redundancy, and Multivariate Information Measures: An Experimentalist's Perspective. *Journal of Computational Neuroscience*, 36:119–140, 2014.

John N Tsitsiklis. Asynchronous Stochastic Approximation and Q-Learning. *Machine Learning*, 16:185–202, 1994.

John N Tsitsiklis and Benjamin Van Roy. Feature-Based Methods for Large Scale Dynamic Programming. *Machine Learning*, 22(1):59–94, 1996.

Cameron Rouse Turner, Dilip Arumugam, Logan Nelson, and Thomas L Griffiths. Trade-Offs Between Tasks Induced by Capacity Constraints Bound the Scope of Intelligence. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

Leslie G Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Hado van Hasselt and Richard S Sutton. Learning to Predict Independent of Span. *arXiv preprint arXiv:1508.04582*, 2015.

Hado Van Hasselt, A Rupam Mahmood, and Richard S Sutton. Off-Policy TD(λ) with a True Online Equivalence. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence, Quebec City, Canada*, pp. 330–339, 2014.

Hado van Hasselt, Sephora Madjiheurem, Matteo Hessel, David Silver, André Barreto, and Diana Borsa. Expected Eligibility Traces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9997–10005, 2021.

Benjamin Van Roy. Performance Loss Bounds for Approximate Value Iteration with State Aggregation. *Mathematics of Operations Research*, 31(2):234–244, 2006.

Akash Velu, Skanda Vaidyanath, and Dilip Arumugam. Hindsight-DICE: Stable Credit Assignment for Deep Reinforcement Learning. *arXiv preprint arXiv:2307.11897*, 2023.

Praveen Venkatesh, Keerthana Gurushankar, and Gabriel Schamberg. Capturing and Interpreting Unique Information. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 2631–2636. IEEE, 2023.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley Q-value: A Local Reward Approach to Solve Global Reward Games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7285–7292, 2020a.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement Learning with General Value function Approximation: Provably Efficient Approach via Bounded Eluder Dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020b.

Christopher JCH Watkins and Peter Dayan. Q-Learning. *Machine Learning*, 8:279–292, 1992.

Théophane Weber, Nicolas Heess, Lars Buesing, and David Silver. Credit Assignment Techniques in Stochastic Computation Graphs. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2650–2660. PMLR, 2019.

Zheng Wen and Benjamin Van Roy. Efficient Exploration and Value Function Generalization in Deterministic Systems. *Advances in Neural Information Processing Systems*, 26, 2013.

Paul L Williams and Randall D Beer. Nonnegative Decomposition of Multivariate Information. *arXiv preprint arXiv:1004.2515*, 2010.

Huizhen Yu. On Convergence of Emphatic Temporal-Difference Learning. In *Conference on Learning Theory*, pp. 1724–1751. PMLR, 2015.