
An Information-Theoretic Perspective on Credit Assignment in Reinforcement Learning

Dilip Arumugam

Department of Computer Science
Stanford University
dilip@cs.stanford.edu

Peter Henderson

Department of Computer Science
Stanford University
phend@cs.stanford.edu

Pierre-Luc Bacon

Mila - University of Montreal
pierre-luc.bacon@mila.quebec

Abstract

How do we formalize the challenge of credit assignment in reinforcement learning? Common intuition would draw attention to reward sparsity as a key contributor to difficult credit assignment and traditional heuristics would look to temporal recency for the solution, calling upon the classic eligibility trace. We posit that it is not the sparsity of the reward itself that causes difficulty in credit assignment, but rather the *information sparsity*. We propose to use information theory to define this notion, which we then use to characterize when credit assignment is an obstacle to efficient learning. With this perspective, we outline several information-theoretic mechanisms for measuring credit under a fixed behavior policy, highlighting the potential of information theory as a key tool towards provably-efficient credit assignment.

1 Introduction

The *credit assignment* problem in reinforcement learning [Minsky, 1961, Sutton, 1985, 1988] is concerned with identifying the contribution of past actions on observed future outcomes. Of particular interest to the reinforcement-learning (RL) problem [Sutton and Barto, 1998] are observed future returns and the value function, which quantitatively answers “*how does choosing an action a in state s affect future return?*” Indeed, given the challenge of sample-efficient RL in long-horizon, sparse-reward tasks, many approaches have been developed to help alleviate the burdens of credit assignment [Sutton, 1985, 1988, Sutton and Barto, 1998, Singh and Sutton, 1996, Precup et al., 2000, Riedmiller et al., 2018, Harutyunyan et al., 2019, Hung et al., 2019, Arjona-Medina et al., 2019, Ferret et al., 2019, Trott et al., 2019, van Hasselt et al., 2020].

The long-horizon, sparse-reward problems examined by many existing works on efficient credit assignment are often recognized as tasks that require prolonged periods of interaction prior to observing non-zero feedback; in the extreme case, these are “goal-based” problems with only a positive reward at terminal states and zero rewards everywhere else. To see why the sparsity of rewards cannot serve as a true hardness measure of credit assignment in RL, consider any sparse-reward MDP. Notice that for any fixed constant $c > 0$, we can have a RL agent interact with the same MDP except under the reward function $\tilde{\mathcal{R}}(s, a) = \mathcal{R}(s, a) + c$. Clearly, this new reward function is no longer sparse; in fact, it is guaranteed to offer non-zero feedback on every timestep. And yet, it is also clear that this modification has neither changed the optimal policy π^* nor has it alleviated any of the burdens of credit assignment that stand in the way of efficiently learning π^* . This simple example illustrates how rectifying the sparsity of rewards does not yield a reduction in the difficulty of credit

assignment. And yet, there are well-known examples of how a prudent choice of reward bonus can substantially accelerate learning [Ng et al., 1999].

While it seems rather easy to show that the sparsity of reward is not the key factor that determines the hardness of credit assignment, the intuitive connection between reward sparsity and the difficulty of credit assignment persists; several works decompose problem difficulties into dichotomies of sparse- and dense-reward tasks [Romoff et al., 2019, Bellemare et al., 2013]. In this work, we maintain that while sparse-reward problems may serve as quintessential examples of decision-making problems where credit assignment is challenging, the underlying mechanism that drives this hardness can be more aptly characterized using information theory. We make this precise by introducing *information sparsity* as a formalization of the real driving force behind the credit assignment challenge; namely, a lack of information between behavior (actions taken under a particular behavior policy) and observed returns, yielding a case of information scarcity.

Beyond clarifying what makes credit assignment difficult, our goal is to show that information theory can also serve as a tool for facilitating efficient credit assignment. To that end, we offer several information-theoretic measures with which an agent may quantitatively allocate credit to specific state-action pairs, under a fixed behavior policy. Each of our proposed measures quantifies a precise relationship between behavior and trajectory returns. We then expand our consideration to not just the single full return of a trajectory, but the entire sequence of returns encountered at each state-action pair, exposing a connection with causal information theory. Our work leaves open the question of how these information-theoretic connections with credit assignment may integrate into existing RL algorithms. More broadly, we hope that the insights presented here inspire subsequent research on the role of information theory in analyzing efficient credit assignment in RL.

Our paper proceeds as follows: we define our problem formulation in Section 2, introduce our notion of information sparsity in Section 3, outline further directions for information-theoretic credit assignment in Section 4, and conclude with discussions of future work in Section 5. Due to space constraints, background on information theory, related work, and all proofs have been relegated to the appendix.

2 Problem Formulation

We consider a finite-horizon Markov Decision Process (MDP) [Bellman, 1957, Puterman, 1994] defined by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, H, \beta, \gamma \rangle$ where \mathcal{S} denotes the state space, \mathcal{A} is the action set, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is a (deterministic) reward function, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ is the transition function producing a distribution over next states given the current state-action pair, H is the finite, fixed episode horizon, $\beta \in \Delta(\mathcal{S})$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. We assume that both \mathcal{S} and \mathcal{A} are finite and use $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$ to denote their respective sizes. At each timestep $h \in [H]$ of the current episode, the agent observes the current state s_h and samples an action a_h according to its current (stochastic) policy $\pi_h : \mathcal{S} \mapsto \Delta(\mathcal{A})$. The agent’s objective is to find a policy so as to maximize the expected sum of future discounted rewards $\mathbb{E}[\sum_{h=1}^H \gamma^{h-1} \mathcal{R}(s_h, a_h)]$, where the expectation is taken with respect to randomness in the initial state, environment transitions, and policy. The value function of a (non-stationary) policy $\pi = (\pi_1, \dots, \pi_H)$ denotes the expected future return by following the policy from a given state s at timestep h , $V_h^\pi(s) = \mathbb{E}[\sum_{h'=h}^H \gamma^{h'-h} \mathcal{R}(s_{h'}, a_{h'}) | s_h = s]$, where the expectation is taken with respect to the policy π and the environment transition dynamics \mathcal{T} . Similarly, we use the Bellman equation to define the action-value function $Q^\pi(s, a)$ representing the expected future return from timestep h , taking action a from state s , and following policy π thereafter, $Q_h^\pi(s, a) = \mathcal{R}(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(\cdot | s, a)} [V_{h+1}^\pi(s')]$, where $V_{H+1}^\pi(s) = 0$.

A fixed behavior policy π induces a stationary visitation distribution over states and state-action pairs denoted as $d^\pi(s)$ and $d^\pi(s, a)$, respectively. Moreover, we let $\rho^\pi(\tau)$ denote the distribution over trajectories generated by a policy π with $\rho^\pi(\tau | s)$ and $\rho^\pi(\tau | s, a)$ conditioning on a particular choice of start state s or starting state-action pair (s, a) respectively. Following the notation from distributional RL [Bellemare et al., 2017], we let $Z(\tau)$ be a random variable denoting the random return obtained after completing trajectory $\tau \sim \rho^\pi(\cdot)$ under behavior policy π ; analogously, $Z \triangleq Z(s, a)$ is a random variable denoting the return observed at state s having taken a and then

following π thereafter¹. Given a trajectory $\tau = \{(s_1, a_1), (s_2, a_2), \dots, (s_{H-1}, a_{H-1}), (s_H, a_H)\}$, we may “index” into its constituent state-action pairs using the following notation: $\tau_h = (s_h, a_h)$, $\tau^h = \{(s_1, a_1), (s_2, a_2), \dots, (s_h, a_h)\}$, $\tau_i^j = \{(s_i, a_i), (s_{i+1}, a_{i+1}), \dots, (s_{j-1}, a_{j-1}), (s_j, a_j)\}$, and $\tau^{-h} = \{(s_1, a_1), \dots, (s_{h-1}, a_{h-1}), (s_{h+1}, a_{h+1}), \dots, (s_H, a_H)\}$.

3 Information Sparsity

The sparsity of rewards is a property of MDPs often mentioned when describing “difficult” decision-making problems. While most works offer a verbal explanation of what constitutes a sparse reward MDP, few papers [Riedmiller et al., 2018, Trott et al., 2019] offer a precise definition through the specification of the reward function as $\mathcal{R}(s, a, s') = \delta_{s_g}(s')$ if $d(s', s_g) \leq \epsilon$ and $\mathcal{R}(s, a, s') = 0$ otherwise, where $d : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}_{\geq 0}$ is a metric on the state space, ϵ is a small constant, and $\delta_{s_g}(s')$ is an arbitrary function defining the reward structure for states within ϵ distance of some goal state s_g , as measured by d ; the common choice is to have $\delta_{s_g}(s')$ be a constant function (for instance, $\delta_{s_g}(s') = 1$).

While a complete lack of feedback across several timesteps demands judicious exploration to encounter the first nontrivial reward signal, exploration is not the only complication. Even after an agent has acquired the first non-zero reward, it still faces a demanding credit-assignment challenge wherein it must decide which step(s) of a long trajectory were critical to the observed outcome. Stemming from this fact, reward sparsity and the credit-assignment challenge are often discussed together leading to a notion that the former is itself a driving force behind the difficulty of the latter.

In this work, we maintain that this phenomenon can be explained via information theory. Recalling the example posed in the introduction, we call attention to how the addition of a positive constant to a sparse reward function, while eliminating sparsity, offers no useful information. In contrast, a more careful choice of, for example, negated distance to goal removes sparsity in an informative way. We can make this precise by examining the following quantity:

$$\mathcal{I}_{s,a}^\pi(Z) = D_{KL}(p(Z|s,a)||p(Z|s)), \quad (1)$$

where $p(Z|s,a)$ denotes the distribution over returns for a random state-action pair conditioned on a particular realization of the state and action. Analogously, $p(Z|s) = \sum_{a \in \mathcal{A}} \pi(a|s)p(Z|s,a)$ denotes the distribution over the random returns for the state-action pair conditioned on a particular realization of the state. The quantity $\mathcal{I}_{s,a}^\pi(Z)$ is itself a random variable depending on the particular realization of the state-action pair (s, a) .

Intuitively, Equation 1 measures how much the distribution over returns of a given state-action pair shifts relative to the distribution over returns for the particular state, marginalizing over all actions. Recalling that $Q^\pi(s, a) = \mathbb{E}_{p(Z|s,a)}[Z]$ and $V^\pi(s) = \mathbb{E}_{p(Z|s)}[Z]$, one may interpret Equation 1 as distributional analogue to the advantage function $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$. To connect this quantity with information theory, we need only apply an expectation:

$$\mathcal{I}(A; Z|S) = \mathbb{E}_{(s,a) \sim d^\pi} [D_{KL}(p(Z|s,a)||p(Z|s))] \quad (2)$$

This quantity carries a very intuitive meaning in the context of credit assignment: conditioned upon states visited by policy π , how much information do the actions of π carry about the returns of those state-action pairs? Difficulties with overcoming the credit-assignment challenge in long-horizon problems arise when $\mathcal{I}(A; Z|S)$ is prohibitively small. That is, a decision-making problem where the actions of the initial policy have little to no dependence on returns cannot acquire the signal needed to learn an optimal policy; sparse-reward problems are a natural example of this. More formally, we can define this notion of *information sparsity* as follows:

¹For clarity, we ignore issues that arise from the mismatch between the continuous random variables $Z(\tau)$, Z and discrete variables involving states, actions, and trajectories. Instead, we think of returns as discrete random variables obtained from a sufficiently fine quantization of the real-valued return.

Definition 1 (Information Sparsity): *Given an MDP \mathcal{M} with non-stationary policy class Π^H , let $\Pi_0 \subset \Pi^H$ denote the set of initial policies employed at the very beginning of learning in \mathcal{M} . For a small constant $\varepsilon > 0$, we classify \mathcal{M} as ε -information-sparse if*

$$\sup_{\pi_0 \in \Pi_0} \mathcal{I}^{\pi_0}(A; Z|S) \leq \varepsilon$$

Under Definition 1, ε -sparse MDPs with small parameter ε (inducing higher sparsity of information) represent a formidable credit-assignment challenge. Using sparse reward problems as an illustrative example and taking information sparsity to be the core obstacle to efficient credit assignment, we may consider how various approaches to dealing with such tasks also resolve information sparsity. Perhaps the most common approach is to employ some form of intrinsic motivation or reward shaping [Ng et al., 1999, Chentanez et al., 2005] with heuristics such as the distance to goal, curiosity, or random network distillation [Pathak et al., 2017, Burda et al., 2018]. In all of these cases, reward sparsity is resolved in a manner that also corrects for information sparsity; to help visualize this, consider a sufficiently-large gridworld MDP with actions for each cardinal direction and a goal-based reward function. Prior to reward augmentation, sparse rewards would likely result in returns of zero across the entire space of state-action pairs visited by a uniform random policy. In contrast, by using a reward bonus equal to, for instance, the negated distance to goal, individual actions taken in almost every state can create meaningful deviations between the distributions $p(Z|s, a)$ and $p(Z|s)$, translating into an increase in the available bits of information measured by information sparsity. A similar comment can also be made for approaches that invoke Thompson sampling [Thompson, 1933] as a tool for facilitating deep exploration [Osband et al., 2016, 2019]; the random noise perturbations used by such approaches translate into excess information that accumulates in the $\mathcal{I}^{\pi_0}(A; Z|S)$ term.

Alternatively, there are other techniques for handling credit assignment that either change the problem formulation altogether or address the long-horizon aspect of decision making. In the latter category, the options framework [Sutton et al., 1999, Bacon et al., 2017] has served as a powerful tool for accommodating efficient RL over long horizons by adopting a two-timescale approach. Similar to a judicious choice of reward shaping function, provision of the right options to an agent can eliminate the difficulty of credit assignment that stems from having a long horizon. In our framework, this can be seen as picking a new (hierarchical) policy class Π_0 to resolve information sparsity. Finally, some approaches simply shift to the multi-task setting and assume access to a function that can identify failed trajectories of one task as successful behaviors for other tasks [Andrychowicz et al., 2017]. As long as these hindsight approaches can generate informative feedback for some subset of the task distribution, they can bootstrap learning of more complicated tasks.

To conclude this section, we consider the computability of information sparsity and recall that, for any function $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$,

$$\mathbb{E}_{(s,a) \sim d^\pi} [f(s, a)] = \mathbb{E}_{\tau \sim \rho^\pi} \left[\sum_{h=1}^H \gamma^{h-1} f(s_h, a_h) \right] \quad (3)$$

Taking $f(s, a) = D_{KL}(p(Z|s, a) || p(Z|s))$, it follows that we need only choose an algorithm for recovering $p(Z|s, a)$ [Morimura et al., 2010, 2012, Bellemare et al., 2017] to compute information sparsity.

4 Information-Theoretic Credit Assignment

In this section, we present potential quantities of interest for deciding how to award credit to an individual state-action pair given the outcome (return) of an entire trajectory.

4.1 Measuring Credit

One possible choice for deciding how responsible or culpable a single step of behavior is for the outcome of the whole trajectory is by performing a sort of sensitivity analysis wherein a single point in the trajectory is varied while all other points are held fixed. This first proposition for measuring the credit of state-action pairs embodies this idea exactly using conditional mutual information.

Proposition 1. Let π be a fixed behavior policy such that $\tau \sim \rho^\pi$. Let R_h be a random variable denoting the reward observed at timestep h (where the randomness of the deterministic reward follows from the randomness of the state-action pair at h , τ_h). It follows that:

$$\mathcal{I}(Z(\tau); \tau_h | \tau^{h-1}) = \mathcal{H}(R_h | \tau^{h-1})$$

The proof is provided in Appendix C.1.

The left-hand side of Proposition 1 is a conditional mutual information term quantifying the information between a single state-action pair τ_h and the policy return $Z(\tau)$, conditioned on the entire trajectory excluding timestep h . The statement of Proposition 1 shows that this measure of credit is equal to the entropy in rewards conditioned on the trajectory up to timestep $h-1$, τ^{h-1} . In practice, this encourages an approach that is reminiscent of RUDDER [Arjona-Medina et al., 2019] wherein a recurrent neural network learns a representation of τ^{h-1} and is trained as a reward classifier (for some discretization of the reward interval); the entropy of the resulting classifier can then be used as a weighting strategy for policy parameter updates or to bias exploration as a reward bonus.

Alternatively, it may be desirable to examine the importance of the current state-action pair towards policy returns conditioned only on the past history, τ^{h-1} . To help facilitate such a measure, it is useful to recall the hindsight distribution $h(a|s, Z(\tau))$ of Harutyunyan et al. [2019] that captures the probability of having taken action a from state s conditioned on the observed trajectory return $Z(\tau)$.

Proposition 2. Let π be a fixed behavior policy such that $\tau \sim \rho^\pi$ and let $h(a|s, Z(\tau))$ be the hindsight distribution as defined above. We have that

$$\mathcal{I}(Z(\tau); \tau_h | \tau^{h-1}) = \mathbb{E}_{\tau^h} \left[\mathbb{E}_{Z(\tau) | \tau^h} \left[\log \left(\frac{h(a_h | s_h, Z(\tau))}{\pi(a_h | s_h)} \right) \right] \right]$$

Moreover,

$$\mathcal{I}(Z(\tau); \tau) = \sum_{h=1}^H \mathbb{E}_{\tau^h} \left[\mathbb{E}_{Z(\tau) | \tau^h} \left[\log \left(\frac{h(a_h | s_h, Z(\tau))}{\pi(a_h | s_h)} \right) \right] \right]$$

The proof is provided in Appendix C.2.

Proposition 2 tells us that learning the hindsight distribution as proposed in Harutyunyan et al. [2019] can also be effectively used to tackle credit assignment in an information-theoretic manner by estimating the conditional mutual information between individual state-action pairs and returns, conditioned on the trajectory up to the previous timestep. While this measure captures a useful quantity intuitively, how to best incorporate such an estimate into an existing RL algorithm remains an open question for future work.

4.2 Causal Information Theory & Hindsight

In the previous section, we examined the information content between a trajectory and its return, leveraging the fact that the trajectory random variable is a sequence of random variables denoting the individual state-action pairs. In this section, we draw attention to the fact that individual returns, like the state-action pairs of a trajectory, are also random variables that appear at each timestep. Typically, we are largely concerned with only one of these random variables, attributed to the first timestep of the trajectory $Z(\tau) \triangleq Z_1$, since returns are computed in hindsight. Naturally, of course, there is an entire sequence of these return random variables Z_1, \dots, Z_H at our disposal. Accordingly, a quantity that may be of great interest when contemplating issues of credit assignment in RL is the following:

$$\mathcal{I}(\tau, Z_1, \dots, Z_H) = \mathcal{I}(\tau_1, \dots, \tau_H; Z_1, \dots, Z_H)$$

which captures all information between a completed trajectory and the sequence of observed returns at each timestep. Recall the chain rule of mutual information:

$$\mathcal{I}(X_1, \dots, X_n; Y) = \sum_{i=1}^n \mathcal{I}(X_i; Y | X^{i-1})$$

where $X^{i-1} = (X_1, \dots, X_{i-1})$ and $X^{-1} = \emptyset$. In the previous section, this allowed for a decomposition of the trajectory in temporal order so that we could examine a current state-action pair τ_h conditioned on the history τ^{h-1} . The analogous step for the sequence of return variables creates a slight oddity where we have the return at a timestep Z_h conditioned on the returns of previous timesteps Z^{h-1} . Here, the temporal ordering that was advantageous in breaking apart a trajectory now results in conditioning on returns that are always computed after observing Z_h . Fortunately, multivariate mutual information is not sensitive to any particular ordering of the random variables, a fact which can be demonstrated quickly for the two-variable case:

Fact 1. *Let X, Y, Z be three random variables.*

$$\mathcal{I}(X; Y, Z) = \mathcal{I}(X; Z, Y)$$

Fact 1 implies that we have a choice between a forward view (for processing variables in temporal order) and a backwards view (for processing in hindsight). This fact by itself is an interesting property of information theory that may deserve more attention in its own right as it blurs the line between the forward-looking perspective of RL and the opposing retrospective view used by credit-assignment techniques for supervised learning [Ke et al., 2018]. It is also reminiscent of the forward and backwards views of the widely-studied eligibility trace [Sutton, 1985, 1988, Sutton and Barto, 1998, Singh and Sutton, 1996]. Since we would like to consider the impact of the entire trajectory on each individual return, we can begin by decomposing the return random variables in hindsight:

$$\mathcal{I}(\tau_1, \dots, \tau_H; Z_1, \dots, Z_H) = \sum_{h=1}^H \mathcal{I}(Z_h; \tau_1, \dots, \tau_H | Z_{h+1}^H)$$

Further expansion of the above multivariate mutual information gives rise to the following proposition that draws a direct connection to causal information theory.

Proposition 3. *Let $\tau = (\tau_1, \dots, \tau_H)$ be a H -step trajectory and let $Z^H = (Z_H, Z_{H-1}, \dots, Z_1)$ be the associated time-synchronized sequence of return random variables. Then, we have that*

$$\mathcal{I}(\tau; Z_1, \dots, Z_H) = \mathcal{I}(\tau^H \rightarrow Z^H)$$

The proof is provided in Appendix C.3.

Notice that, in general, Proposition 3 is not always true for two arbitrary, time-synchronized stochastic processes as the multivariate mutual information and directed information obey a conservation law [Massey and Massey, 2005].

By following the first steps from the proof of Proposition 3, we can also recover an analog to Proposition 1 that prescribes a connection between $\mathcal{I}(\tau; Z_1, \dots, Z_H)$ and individual rewards R_h .

Proposition 4.

$$\mathcal{I}(\tau; Z_1, \dots, Z_H) = \mathcal{I}(\tau^H \rightarrow Z^H) = \sum_{h=1}^H \mathcal{H}(R_h | Z_{h+1}^H)$$

The proof is provided in Appendix C.3.

Taken together, Propositions 3 and 4 offer an interesting connection between information theory and causal inference, a link which has appeared before [Amblard and Michel, 2013]. We leave the question of how an agent might leverage such quantities to actively reason about the underlying causal structure of the environment to future work.

5 Discussion & Conclusion

In this work, we take an initial step towards a rigorous formulation of the credit-assignment problem in reinforcement learning. At the core of our approach is information theory, which we find naturally suited for obtaining quantitative answers to the core question facing an agent when dealing with credit assignment: *how does choosing an action a in state s affect future return?* While this work offers preliminary ideas for how information theory can then be used to understand credit assignment, it remains to be seen how these measures can be integrated into existing RL algorithms.

References

- 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, pages 3134–3142, 2019.
- Ziv Aharoni, Oron Sabag, and Haim Henri Permuter. Reinforcement learning evaluation and solution for the feedback capacity of the ising channel with large alphabet. *arXiv preprint arXiv:2008.07983*, 2020a.
- Ziv Aharoni, Dor Tsur, Ziv Goldfeld, and Haim Henry Permuter. Capacity of continuous channels with memory via directed information neural estimator. *arXiv preprint arXiv:2003.04179*, 2020b.
- Pierre-Olivier Amblard and Olivier JJ Michel. The relation between granger causality and directed information theory: A review. *Entropy*, 15(1):113–143, 2013.
- Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in neural information processing systems*, pages 5048–5058, 2017.
- Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. In *Advances in Neural Information Processing Systems*, pages 13544–13555, 2019.
- Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- 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*, 47: 253–279, 2013.
- Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 449–458. JMLR. org, 2017.
- Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.
- Toby Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.
- Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2018.
- Nuttapong Chentanez, Andrew G Barto, and Satinder P Singh. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 1281–1288, 2005.
- Kun-Jen Chung and Matthew J Sobel. Discounted mdp’s: Distribution functions and exponential utility maximization. *SIAM journal on control and optimization*, 25(1):49–62, 1987.
- Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- Johan Ferret, Raphaël Marinier, Matthieu Geist, and Olivier Pietquin. Credit assignment as a proxy for transfer in reinforcement learning. *arXiv preprint arXiv:1907.08027*, 2019.
- Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 202–211, 2016.
- Alexandre Galashov, Siddhant M Jayakumar, Leonard Hasenclever, Dhruva Tirumala, Jonathan Schwarz, Guillaume Desjardins, Wojciech M Czarnecki, Yee Whye Teh, Razvan Pascanu, and Nicolas Heess. Information asymmetry in kl-regularized rl. *arXiv preprint arXiv:1905.01240*, 2019.

- Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1352–1361. JMLR. org, 2017.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, 2018.
- 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. In *Advances in neural information processing systems*, pages 12467–12476, 2019.
- Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Gheshlaghi Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *AAAI*, 2018.
- Rein Houthoofd, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.
- Chia-Chun Hung, Timothy Lillicrap, Josh Abramson, Yan Wu, Mehdi Mirza, Federico Carnevale, Arun Ahuja, and Greg Wayne. Optimizing agent behavior over long time scales by transporting value. *Nature communications*, 10(1):1–12, 2019.
- Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.
- Nan Rosemary Ke, Anirudh Goyal, Olexa Bilaniuk, Jonathan Binas, Michael C Mozer, Chris Pal, and Yoshua Bengio. Sparse attentive backtracking: Temporal credit assignment through reminding. In *NeurIPS*, 2018.
- Hyungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. Emi: Exploration with mutual information. In *International Conference on Machine Learning*, pages 3360–3369, 2019.
- George Konidaris, Scott Niekum, and Philip S Thomas. Td_gamma: Re-evaluating complex backups in temporal difference learning. In *Advances in Neural Information Processing Systems*, pages 2402–2410, 2011.
- Gerhard Kramer. Directed information for channels with feedback. *Ph.D. thesis, ETH Zurich*, 1998.
- Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.
- Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics*, 2006.
- Simon Li, Ashish Khisti, and Aditya Mahajan. Information-theoretic privacy for smart metering systems with a rechargeable battery. *IEEE Transactions on Information Theory*, 64(5):3679–3695, 2018.
- James Massey. Causality, feedback and directed information. In *Proc. 1990 Int. Symp. on Inform. Theory and its Applications*, pages 27–30, 1990.
- James L Massey and Peter C Massey. Conservation of mutual and directed information. In *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.*, pages 157–158. IEEE, 2005.
- Marvin Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961.
- Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *ICML*, 2010.

- Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.
- Andrew Y Ng, Daishi Harada, and Stuart J Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287. Morgan Kaufmann Publishers Inc., 1999.
- Daniel Alexander Ortega and Pedro Alejandro Braun. Information, utility and bounded rationality. In *International Conference on Artificial General Intelligence*, pages 269–274. Springer, 2011.
- Pedro A Ortega and Daniel A Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683, 2013.
- Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pages 4026–4034, 2016.
- Ian Osband, Benjamin Van Roy, Daniel J Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019.
- Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.
- Haim Permuter, Paul Cuff, Benjamin Van Roy, and Tsachy Weissman. Capacity of the trapdoor channel with feedback. *IEEE Transactions on Information Theory*, 54(7):3150–3165, 2008a.
- Haim H Permuter, Young-Han Kim, and Tsachy Weissman. On directed information and gambling. In *2008 IEEE International Symposium on Information Theory*, pages 1403–1407. IEEE, 2008b.
- Doina Precup, Richard S Sutton, and Satinder P Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766. Morgan Kaufmann Publishers Inc., 2000.
- Martin L. Puterman. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.
- Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degraeve, Tom Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing solving sparse reward tasks from scratch. In *International Conference on Machine Learning*, pages 4344–4353, 2018.
- Joshua Romoff, Peter Henderson, Ahmed Touati, Emma Brunskill, Joelle Pineau, and Yann Ollivier. Separating value functions across time-scales. *arXiv preprint arXiv:1902.01883*, 2019.
- Jonathan Rubin, Ohad Shamir, and Naftali Tishby. Trading value and information in mdps. In *Decision Making with Imperfect Decision Makers*, pages 57–74. Springer, 2012.
- Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.
- Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.
- Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.*, March 1959, 4:142–163, 1959.
- Satinder P Singh and Richard S Sutton. Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1-3):123–158, 1996.
- Matthew J Sobel. The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

- Richard S Sutton. Temporal credit assignment in reinforcement learning. 1985.
- 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. 1998.
- Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.
- Sekhar Tatikonda. A markov decision approach to feedback channel capacity. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 3213–3218. IEEE, 2005.
- Sekhar Tatikonda and Sanjoy Mitter. The capacity of channels with feedback. *IEEE Transactions on Information Theory*, 55(1):323–349, 2008.
- Philip S Thomas, Scott Niekum, Georgios Theodorou, and George Konidaris. Policy evaluation using the ω -return. In *Advances in Neural Information Processing Systems*, pages 334–342, 2015.
- William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.
- Stas Tiomkin and Naftali Tishby. A unified bellman equation for causal information and value in markov decision processes. *arXiv preprint arXiv:1703.01585*, 2017.
- Dhruva Tirumala, Hyeonwoo Noh, Alexandre Galashov, Leonard Hasenclever, Arun Ahuja, Greg Wayne, Razvan Pascanu, Yee Whye Teh, and Nicolas Heess. Exploiting hierarchy for learning and transfer in kl-regularized rl. *arXiv preprint arXiv:1903.07438*, 2019.
- Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer, 2011.
- Emanuel Todorov. Linearly-solvable markov decision problems. In *Advances in neural information processing systems*, pages 1369–1376, 2007.
- Marc Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th annual international conference on machine learning*, pages 1049–1056, 2009.
- Alexander Trotter, Stephan Zheng, Caiming Xiong, and Richard Socher. Keeping your distance: Solving sparse reward tasks using self-balancing shaped rewards. In *Advances in Neural Information Processing Systems*, pages 10376–10386, 2019.
- Hado van Hasselt, Sephora Madjiheurem, Matteo Hessel, David Silver, André Barreto, and Diana Borsa. Expected eligibility traces. *arXiv preprint arXiv:2007.01839*, 2020.
- Brian D Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Carnegie Mellon University, 2010.

A Background

In this section, we review standard quantities in information theory as well as causal information theory. For more background on information theory, see [Cover and Thomas \[2012\]](#).

A.1 Information Theory

Definition 2 (Entropy & Conditional Entropy): *For a discrete random variable X with density function $p(x)$ and support $\text{supp}(p(x)) = \mathcal{X}$, the entropy of X is given by*

$$\begin{aligned}\mathcal{H}(X) &= -\mathbb{E}_{p(x)}[\log(p(x))] \\ &= -\sum_{x \in \mathcal{X}} p(x) \log(p(x))\end{aligned}$$

Similarly, the conditional entropy of a discrete random variable Y given X with density $p(y)$ ($\text{supp}(p(y)) = \mathcal{Y}$) and joint density $p(x, y)$ is given by

$$\begin{aligned}\mathcal{H}(Y|X) &= -\mathbb{E}_{p(x,y)}[\log(p(y|x))] \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log(p(y|x))\end{aligned}$$

Definition 3 (Chain Rule for Entropy): *For a collection of random variables X_1, \dots, X_n , the joint entropy can be decomposed as a sum of conditional entropies:*

$$\mathcal{H}(X_1, \dots, X_n) = \sum_{i=1}^n \mathcal{H}(X_i | X_1, \dots, X_{i-1})$$

Definition 4 (Kullback-Leibler Divergence): *The KL-divergence between two probability distributions p, q with identical support \mathcal{X} is*

$$D_{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

Definition 5 (Mutual Information): *The mutual information between two random variables X and Y is given by*

$$\begin{aligned}\mathcal{I}(X; Y) &= D_{KL}(p(x, y) || p(x)p(y)) \\ &= \mathcal{H}(X) - \mathcal{H}(X|Y) \\ &= \mathcal{H}(Y) - \mathcal{H}(Y|X)\end{aligned}$$

Also, recognizing that $p(x|y) = \frac{p(x,y)}{p(y)}$, we have that

$$\mathcal{I}(X; Y) = \mathbb{E}_Y[D_{KL}(p(x|y) || p(x))]$$

Definition 6 (Conditional Mutual Information): *The conditional mutual information between two variables X and Y , given a third random variable Z is*

$$\begin{aligned}\mathcal{I}(X; Y|Z) &= \mathbb{E}_Z[D_{KL}(p(x, y|z)||p(x|z)p(y|z))] \\ &= \mathcal{H}(X|Z) - \mathcal{H}(X|Y, Z) \\ &= \mathcal{H}(Y|Z) - \mathcal{H}(Y|X, Z) \\ &= \mathcal{I}(X; Y, Z) - \mathcal{I}(X; Z)\end{aligned}$$

Also, recognizing that $p(x|y, z) = \frac{p(x, y|z)}{p(y|z)}$, we have that

$$\mathcal{I}(X; Y|Z) = \mathbb{E}_Z[\mathbb{E}_{Y|Z}[D_{KL}(p(x|y, z)||p(x|z))]]$$

Definition 7 (Multivariate Mutual Information): *The multivariate mutual information between three random variables X , Y , and Z is given by:*

$$\mathcal{I}(X; Y, Z) = \mathcal{I}(X; Z) + \mathcal{I}(X; Y|Z)$$

Definition 8 (Chain Rule for Mutual Information): *For a random variables X and Z_1, \dots, Z_n the multivariate mutual information decomposes into a sum of conditional mutual information terms:*

$$\mathcal{I}(X; Z_1, \dots, Z_n) = \sum_{i=1}^n \mathcal{I}(X; Z_i|Z_1, \dots, Z_{i-1})$$

A.2 Causal Information Theory

Definition 9 (Causal Conditioning & Entropy [Kramer, 1998]): *Consider two time-synchronized stochastic processes $X^T = (X_1, \dots, X_T)$ and $Y^T = (Y_1, \dots, Y_T)$. Let x^T and y^T denote two realizations of the respective processes. The causally conditioned probability of the sequence y^T given x^T is*

$$\begin{aligned}p(y^T||x^T) &= \prod_{t=1}^T p(y_t|y^{t-1}, x^{t-1}) \\ &= \prod_{t=1}^T p(y_t|y_1, \dots, y_{t-1}, x_1, \dots, x_{t-1})\end{aligned}$$

The causal entropy of Y^T given X^T is

$$\begin{aligned}\mathcal{H}(Y^T||X^T) &= -\mathbb{E}_{p(x^T, y^T)}[\log(p(y^T||x^T))] \\ &= \sum_{t=1}^T \mathcal{H}(Y_t|Y^{t-1}, X^{t-1})\end{aligned}$$

Definition 10 (Directed Information Flow [Massey, 1990, Permuter et al., 2008b]): *Let $X^T = (X_1, \dots, X_T)$ and $Y^T = (Y_1, \dots, Y_T)$ denote two time-synchronized stochastic processes.*

The directed information that flows from X^T to Y^T is given by

$$\begin{aligned} \mathcal{I}(X^T \rightarrow Y^T) &= \sum_{t=1}^T \mathcal{I}(X^t; Y_t | Y^{t-1}) \\ &= \mathcal{H}(Y^T) - \mathcal{H}(Y^T | X^T) \end{aligned}$$

Remark 1. Notice that, in general, $\mathcal{I}(X \rightarrow Y) \neq \mathcal{I}(Y \rightarrow X)$

B Related Work

While there are various methods for circumventing the challenges of sparse rewards, there are relatively fewer works that directly study the credit assignment problem in RL. Perhaps foremost among them are the classic eligibility traces [Sutton, 1985, 1988, Sutton and Barto, 1998, Singh and Sutton, 1996, Precup et al., 2000, van Hasselt et al., 2020] which leverage temporal recency as a heuristic for assigning credit to visited states/state-action pairs during temporal-difference learning. Konidaris et al. [2011], Thomas et al. [2015] offer a formal derivation of standard eligibility traces, underscoring how the underlying theoretical assumptions on the random variables denoting n -step returns are typically not realized in practice. In a similar spirit, our work is also concerned with rectifying the shortcomings of eligibility traces, highlighting how temporal recency acts a poor heuristic in hard credit-assignment problems where informative feedback is scarce and credit must be allocated across a number of timesteps that may far exceed the effective horizon of the trace.

Other recent works offer means of modulating or re-using data collected within the environment to help alleviate the burdens of credit assignment. Arjona-Medina et al. [2019] give a formal characterization of return-equivalent decision-making problems and introduce the idea of reward redistribution for dispersing sparse, informative feedback backwards in time to the facilitating state-action pairs while still preserving the optimal policy. They realize a practical instantiation of this RUDDER idea through recurrent neural networks and examining differences between return predictions conditioned on partial trajectories. Harutyunyan et al. [2019] offer a refactoring of the Bellman equation via importance sampling to introduce a hindsight distribution, $h(a|s, Z(\tau))$, over actions conditioned on state and the return $Z(\tau)$ of a completed trajectory τ . This distribution is then used solely for the purposes of hindsight credit assignment (HCA) in determining the impact of a particular action a on an observe outcome $Z(\tau)$ through the likelihood ratio $\frac{h(a|s, Z(\tau))}{\pi(a|s)}$. In our work, we introduce two quantities to help measure the credit of a single behavior step towards an observed return. One of these measures, similar to RUDDER, prescribes examining the information contained in partial trajectories over next rewards (rather than returns). The other measure naturally yields the same likelihood ratio of HCA purely from information theory.

For many years, work at the intersection of information theory and reinforcement learning has been a topic of great interest. Perhaps most recently, there has been a resurgence of interest in the control-as-inference framework [Todorov, 2007, Toussaint, 2009, Kappen et al., 2012, Levine, 2018] leading to maximum-entropy (deep) RL approaches [Ziebart, 2010, Fox et al., 2016, Haarnoja et al., 2017, 2018]. These methods take the uniform distribution as an uninformative prior over actions and can be more broadly categorized as KL-regularized RL [Todorov, 2007, Galashov et al., 2019, Tirumala et al., 2019]. Curiously, these information-theoretic RL objectives, derived from the perspective of probabilistic inference, can also be derived from information-theoretic characterizations of bounded rationality [Tishby and Polani, 2011, Ortega and Braun, 2011, Rubin et al., 2012, Ortega and Braun, 2013]. In this setting, a policy is viewed as a channel in the information-theoretic sense and, as with any channel, there is a cost to channel communication (mapping individual states to actions) that goes unaccounted for in the standard RL objective of reward maximization. By modeling this communication cost explicitly, these works also arrive at the KL-regularized RL objective. Examining the communication rate over time, rather than the instantaneous cost, yields a natural analog to these approaches [Tiomkin and Tishby, 2017] articulated in terms of causal information theory [Kramer, 1998]. Orthogonally, Russo and Van Roy [2016, 2018] study the role of information theory in analyzing efficient exploration; they introduce the information ratio to characterize the balance between taking regret-minimizing actions (exploitation) and actions that have high information gain (exploration), leading to a general regret bound for the multi-armed bandit setting. Empirical

work studying effective heuristics for guiding exploration in deep RL have also made great use of information-theoretic quantities [Houthoof et al., 2016, Kim et al., 2019]. Tackling the issue of generalization in RL, [Abel et al., 2019] use rate-distortion theory [Shannon, 1959, Berger, 1971] to reformulate the learning of state abstractions [Li et al., 2006] as an optimization problem in the space of lossy-compression schemes.

A common thread among all the aforementioned work on information theory and RL is that the quantities leveraged are exclusively concerned with only the states and/or actions taken by an agent. In contrast, the information-theoretic quantities explored in this work also incorporate a focus on returns. Given the recent successes of distributional RL [Sobel, 1982, Chung and Sobel, 1987, Bellemare et al., 2017, Hessel et al., 2018] that explicitly draw attention to the return random variable (and its underlying distribution) as a quantity of interest, it seems natural to re-examine the role that information theory might play in RL with this critical random variable involved. Moreover, a natural place to explore such a connection is the credit assignment problem which directly asks about the dependence (or information) that individual steps of behavior carry about returns (outcomes).

Coincidentally, there are numerous works in the information-theory community which study MDPs in the context of computing and optimizing multivariate mutual information and directed information [Tatikonda, 2005, Tatikonda and Mitter, 2008, Permuter et al., 2008a, Li et al., 2018]. These approaches formulate a specific MDP and applying dynamic programming to recover channel capacity as the corresponding optimal value function. There have also been works examining the success of neural networks for computing and optimizing directed information [Aharoni et al., 2020b,a]. While our work moves in the opposite direction to ask how information theory can help address a core challenge of RL, we may turn to these approaches for inspiration on how to employ our information-theoretic quantities for credit assignment in practice.

C Proofs: Information-Theoretic Credit Assignment

Here we present the full version of all theoretical results presented in the main paper.

C.1 Measuring Credit

Fact 2. Let X, Y be two discrete random variables and define $S = X + Y$. Then

$$\mathcal{H}(S|X) = \mathcal{H}(Y|X)$$

Proof.

$$\begin{aligned} \mathcal{H}(S|X) &= -\mathbb{E}_x \left[\sum_s p(S = s|X = x) \log(p(S = s|X = x)) \right] \\ &= -\mathbb{E}_x \left[\sum_s p(Y = s - x|X = x) \log(p(Y = s - x|X = x)) \right] \\ &= -\mathbb{E}_x \left[\sum_y p(Y = y|X = x) \log(p(Y = y|X = x)) \right] \\ &= \mathcal{H}(Y|X) \end{aligned}$$

where we perform a change of variables $y = s - x$ in the third line

□

Proposition 5. Let π be a fixed behavior policy such that $\tau \sim \rho^\pi$. Let R_h be a random variable denoting the reward observed at timestep h (where the randomness of the deterministic reward follows from the randomness of the state-action pair at h , τ_h). It follows that:

$$\mathcal{I}(Z(\tau); \tau_h | \tau^{-h}) = \mathcal{H}(R_h | \tau^{h-1})$$

Proof.

$$\begin{aligned}
\mathcal{I}(Z(\tau); \tau_h | \tau^{-h}) &\stackrel{(a)}{=} \mathcal{I}(Z(\tau); \tau_h, \tau^{-h}) - \mathcal{I}(Z(\tau); \tau^{-h}) \\
&= \mathcal{I}(Z(\tau); \tau) - \mathcal{I}(Z(\tau); \tau^{-h}) \\
&\stackrel{(b)}{=} \mathcal{H}(Z(\tau)) - \mathcal{H}(Z(\tau) | \tau) - \mathcal{H}(Z(\tau)) + \mathcal{H}(Z(\tau) | \tau^{-h}) \\
&= \mathcal{H}(Z(\tau) | \tau^{-h}) - \mathcal{H}(Z(\tau) | \tau) \\
&\stackrel{(c)}{=} \mathcal{H}(Z(\tau) | \tau^{-h}) \\
&\stackrel{(d)}{=} \mathcal{H}(R_h | \tau^{-h}) \\
&= \mathcal{H}(R_h | \tau^{h-1}, \tau_{h+1}^H) \\
&\stackrel{(e)}{=} \mathcal{H}(R_h | \tau^{h-1})
\end{aligned}$$

where the steps follow from: (a) the chain rule of mutual information, (b) the definition of mutual information, (c) $Z(\tau)$ is a deterministic function of τ under a deterministic reward function, (d) applies Fact 2 on $Z(\tau) = \sum_{h=1}^H \gamma^{h-1} R_h$, and (e) R_h is independent of the future trajectory τ_{h+1}^H .

□

C.2 Hindsight

Proposition 6. *Let π be a fixed behavior policy such that $\tau \sim \rho^\pi$ and let $h(a|s, Z(\tau))$ be the hindsight distribution as defined above. We have that*

$$\mathcal{I}(Z(\tau); \tau_h | \tau^{h-1}) = \mathbb{E}_{\tau^h} \left[\mathbb{E}_{Z(\tau) | \tau^h} \left[\log \left(\frac{h(a_h | s_h, Z(\tau))}{\pi(a_h | s_h)} \right) \right] \right]$$

Moreover,

$$\mathcal{I}(Z(\tau); \tau) = \sum_{h=1}^H \mathbb{E}_{\tau^h} \left[\mathbb{E}_{Z(\tau) | \tau^h} \left[\log \left(\frac{h(a_h | s_h, Z(\tau))}{\pi(a_h | s_h)} \right) \right] \right]$$

Proof.

Notice that by the definition of conditional mutual information:

$$\begin{aligned}
\mathcal{I}(Z(\tau); \tau_h | \tau^{h-1}) &= \mathbb{E}_{\tau^{h-1}} \left[\mathbb{E}_{\tau_h | \tau^{h-1}} \left[D_{KL}(p(Z(\tau) | \tau_h, \tau^{h-1}) || p(Z(\tau) | \tau^{h-1})) \right] \right] \\
&= \mathbb{E}_{\tau^h} \left[D_{KL}(p(Z(\tau) | \tau^h) || p(Z(\tau) | \tau^{h-1})) \right] \\
&= \mathbb{E}_{\tau^h} \left[\mathbb{E}_{Z(\tau) | \tau^h} \left[\log \left(\frac{p(Z(\tau) | \tau^h)}{p(Z(\tau) | \tau^{h-1})} \right) \right] \right]
\end{aligned}$$

By applying Bayes' rule twice, we have:

$$\begin{aligned}
\mathcal{I}(Z(\tau); \tau_h | \tau^{h-1}) &= \mathbb{E}_{\tau^h} \left[\mathbb{E}_{Z(\tau) | \tau^h} \left[\log \left(\frac{p(Z(\tau) | \tau^h)}{p(Z(\tau) | \tau^{h-1})} \right) \right] \right] \\
&= \mathbb{E}_{\tau^h} \left[\mathbb{E}_{Z(\tau) | \tau^h} \left[\log \left(\frac{p(\tau^h | Z(\tau)) p(\tau^{h-1})}{p(\tau^{h-1} | Z(\tau)) p(\tau^h)} \right) \right] \right] \\
&= \mathbb{E}_{\tau^h} \left[\mathbb{E}_{Z(\tau) | \tau^h} \left[\log \left(\frac{h(a_h | s_h, Z(\tau))}{\pi(a_h | s_h)} \right) \right] \right]
\end{aligned}$$

where the final equation follows from the fact that

$$p(\tau^h) = \beta(s_1)\pi(a_1|s_1) \prod_{h'=2}^h \pi(a_{h'}|s_{h'})\mathcal{T}(s_{h'}|s_{h'-1}, a_{h'-1})$$

$$p(\tau^h|Z(\tau)) = \beta(s_1)h(a_1|s_1, Z(\tau)) \prod_{h'=2}^h h(a_{h'}|s_{h'}, Z(\tau))\mathcal{T}(s_{h'}|s_{h'-1}, a_{h'-1})$$

with analogous distributions for $p(\tau^{h-1})$ and $p(\tau^{h-1}|Z(\tau))$.

To show the second claim, we simply apply the chain rule of mutual information:

$$\begin{aligned} \mathcal{I}(Z(\tau); \tau) &= \mathcal{I}(Z(\tau); \tau_1, \tau_2, \dots, \tau_H) \\ &= \sum_{h=1}^H \mathcal{I}(Z(\tau); \tau_h | \tau^{h-1}) \\ &= \sum_{h=1}^H \mathbb{E}_{\tau^h} \left[\mathbb{E}_{Z(\tau) | \tau^h} \left[\log \left(\frac{h(a_h | s_h, Z(\tau))}{\pi(a_h | s_h)} \right) \right] \right] \end{aligned}$$

□

C.3 Causal Information Theory & Hindsight

Fact 3. Let X, Y, Z be three random variables.

$$\mathcal{I}(X; Y, Z) = \mathcal{I}(X; Z, Y)$$

Proof.

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

□

Proposition 7. Let $\tau = (\tau_1, \dots, \tau_H)$ be a H -step trajectory and let $Z^H = (Z_H, Z_{H-1}, \dots, Z_1)$ be the associated time-synchronized sequence of return random variables. Then, we have that

$$\mathcal{I}(\tau; Z_1, \dots, Z_H) = \mathcal{I}(\tau^H \rightarrow Z^H)$$

Proof.

Notice that, for any timestep h , we can group the state-action pairs contained in a trajectory by h into a past $\tau^{h-1} = (\tau_1, \dots, \tau_{h-1})$, present τ_h , and future $\tau_{h+1}^H = (\tau_{h+1}, \dots, \tau_H)$.

$$\begin{aligned} \mathcal{I}(\tau; Z_1, \dots, Z_H) &= \sum_{h=1}^H \mathcal{I}(Z_h; \tau_1, \dots, \tau_H | Z_{h+1}^H) \\ &= \sum_{h=1}^H \mathcal{I}(Z_h; \tau^{h-1}, \tau_h, \tau_{h+1}^H | Z_{h+1}^H) \\ &= \sum_{h=1}^H \mathcal{I}(Z_h; \tau^{h-1}, | Z_{h+1}^H) + \mathcal{I}(Z_h; \tau_h, | \tau^{h-1}, Z_{h+1}^H) + \mathcal{I}(Z_h; \tau_{h+1}^H, | \tau^{h-1}, \tau_h, Z_{h+1}^H) \end{aligned}$$

Recall that we take the reward function of our MDP to be deterministic. Consequently, we have that Z_h is a deterministic function of τ_h and Z_{h+1} . Thus, $\mathcal{I}(Z_h; \tau_{h+1}^H | \tau^{h-1}, \tau_h, Z_{h+1}^H) = 0$ and we're left with

$$\begin{aligned} \mathcal{I}(\tau; Z_1, \dots, Z_H) &= \sum_{h=1}^H \mathcal{I}(Z_h; \tau^{h-1}, | Z_{h+1}^H) + \mathcal{I}(Z_h; \tau_h, | \tau^{h-1}, Z_{h+1}^H) \\ &= \sum_{h=1}^H \mathcal{I}(Z_h; \tau^{h-1}, \tau_h | Z_{h+1}^H) \\ &= \sum_{h=1}^H \mathcal{I}(Z_h; \tau^h | Z_{h+1}^H) \\ &= \mathcal{I}(\tau^H \rightarrow Z^H) \end{aligned}$$

□

Fact 4. Let X, Y, Z be three discrete random variables and let $S = X + Y$. Then,

$$\mathcal{I}(S; Y | Z) = \mathcal{I}(X; Y | Z)$$

Proof.

$$\begin{aligned} \mathcal{I}(S; Y | Z) &= \mathcal{I}(X + Z; Y | Z) \\ &= \mathcal{H}(X + Z | Z) - \mathcal{H}(X + Z | Y, Z) \\ &= \mathcal{H}(X | Z) - \mathcal{H}(X | Y, Z) \\ &= \mathcal{I}(X; Y | Z) \end{aligned}$$

□

where the third line applies Fact 2.

Proposition 8.

$$\mathcal{I}(\tau; Z_1, \dots, Z_H) = \mathcal{I}(\tau^H \rightarrow Z^H) = \sum_{h=1}^H \mathcal{H}(R_h | Z_{h+1}^H)$$

Proof.

Using τ^{h-1} to denote the past trajectory from a given timestep as above

$$\begin{aligned}
\mathcal{I}(\tau; Z^H) &= \sum_{h=1}^H \mathcal{I}(Z_h; \tau^{h-1}, |Z_{h+1}^H) + \mathcal{I}(Z_h; \tau_h, |\tau^{h-1}, Z_{h+1}^H) \\
&= \sum_{h=1}^H \mathcal{I}(R_h; \tau^{h-1} | Z_{h+1}^H) + \mathcal{I}(R_h; \tau_h | \tau^{h-1}, Z_{h+1}^H) \\
&= \sum_{h=1}^H \mathcal{I}(R_h; \tau^{h-1}, \tau_h | Z_{h+1}^H) \\
&= \sum_{h=1}^H \mathcal{I}(R_h; \tau^h | Z_{h+1}^H) \\
&= \sum_{h=1}^H \mathcal{H}(R_h | Z_{h+1}^H) - \mathcal{H}(R_h | \tau^h, Z_{h+1}^H) \\
&= \sum_{h=1}^H \mathcal{H}(R_h | Z_{h+1}^H)
\end{aligned}$$

where the second line recognizes that $Z_h = R_h + \gamma Z_{h+1}$ and employs Fact 4. The last line follows from the fact that R_h is a deterministic function of τ_h .

□