WHICH MUTUAL-INFORMATION REPRESENTATION LEARNING OBJECTIVES ARE SUFFICIENT FOR CONTROL?

Anonymous authors
Paper under double-blind review

ABSTRACT

Mutual information maximization provides an appealing formalism for learning representations of data. In the context of reinforcement learning, such representations can accelerate learning by discarding irrelevant and redundant information, while retaining the information necessary for control. Much of the prior work on these methods has addressed the practical difficulties of estimating mutual information from samples of high-dimensional observations, while comparatively less is understood about *which* mutual information objectives are sufficient for reinforcement learning (RL) from a theoretical perspective. In this paper we identify conditions under which representations that maximize specific mutual-information objectives are theoretically sufficient for learning and representing the optimal policy. Somewhat surprisingly, we find that several popular objectives can yield insufficient representations given mild and common assumptions on the structure of the MDP. We corroborate our theoretical results with empirical results experiments on a simulated game environment with visual observations.

1 INTRODUCTION

While deep reinforcement learning (RL) algorithms are capable of learning policies from high-dimensional observations, such as images (Mnih et al., 2013; Lee et al., 2019; Kalashnikov et al., 2018), in practice policy learning faces a bottleneck in acquiring useful representations of the observation space (Shelhamer et al., 2016). State representation learning approaches aim to remedy this issue by learning structured and compact representations on which to perform RL. While a wide range of representation learning objectives have been proposed (Lesort et al., 2018), a particularly appealing class of methods that is amenable to rigorous analysis is based on maximizing mutual information (MI) between variables. In the unsupervised learning setting, this is often realized as the InfoMax principle (Linsker, 1988; Bell & Sejnowski, 1995), which maximizes the mutual information between the input and its latent representation subject to domain-specific constraints. This approach has been widely applied in unsupervised learning in the domains of image, audio, and natural language understanding (Oord et al., 2018; Hjelm et al., 2018; Ravanelli & Bengio, 2019). In RL, the variables of interest for MI maximization are sequential states, actions, and rewards (see Figure 1). As we will discuss, several popular methods for representation learning in RL involve mutual information maximization with different combinations of these variables (Anand et al., 2019; Oord et al., 2018; Pathak et al., 2017; Shelhamer et al., 2016).

A useful representation should retain the factors of variation that are necessary to learn and represent the optimal policy or the optimal value function, and discard irrelevant and redundant information. While much prior work has focused on the problem of how to optimize various mutual information objectives in high dimensions (Song & Ermon, 2019; Belghazi et al., 2018; Oord et al., 2018; Hjelm et al., 2018), we focus instead on whether the representations that maximize these objectives are actually theoretically sufficient for learning and representing the optimal policy or value function. We find that some commonly used objectives are insufficient given relatively mild and common assumptions on the structure of the MDP, and identify other objectives which are sufficient. We show these results theoretically and illustrate the analysis empirically in didactic examples in which MI can be computed exactly. Our results provide some guidance to the deep RL practitioner on when and why objectives may be expected to work well or fail, and also provide a framework to analyze

newly proposed representation learning objectives based on mutual information. To investigate how our theoretical results pertain to deep RL, we compare the performance of RL agents in a simulated game trained with state representations learned by maximizing the MI objective given visual inputs. The experimental results corroborate our theoretical findings, and demonstrate that the sufficiency of a representation can have a substantial impact on the performance of an RL agent that uses that representation.

2 Related Work

In this paper, we analyze several widely used mutual information objectives for control. In this section we first review MI-based unsupervised learning, then the application of these techniques to the RL setting. Finally, we discuss alternative perspectives on representation learning in RL.

Mutual information-based unsupervised learning. Mutual information-based methods are particularly appealing for representation learning as they admit both rigorous analysis and intuitive interpretation. Tracing its roots to the InfoMax principle (Linsker, 1988; Bell & Sejnowski, 1995), a common technique is to maximize the mutual information between the (often high-dimensional) input and its latent representation subject to domain-specific constraints (Becker & Hinton, 1992). This technique has been applied to learn representations for natural language (Devlin et al., 2019), video (Sun et al., 2019), and images (Bachman et al., 2019; Hjelm et al., 2018). A major challenge to using MI maximization methods in practice is the difficulty of estimating MI from samples (McAllester & Statos, 2018) and with high-dimensional inputs (Song & Ermon, 2019). A great deal of recent work has focused on improving MI estimation via variational methods (Song & Ermon, 2019; Poole et al., 2019; Oord et al., 2018; Belghazi et al., 2018). In this work we are concerned with analyzing the MI objectives, and not the estimation method. In our experiments with image observations, we make use of noise contrastive estimation methods (Gutmann & Hyvärinen, 2010), though other choices could also suffice.

Mutual information objectives in RL. Reinforcement learning adds aspects of temporal structure and control to the standard unsupervised learning problem discussed above (see Figure 1). This structure can be leveraged in MI maximization approaches by taking into account mutual information between sequential states, actions, or combinations thereof. Some prior work omits the action, maximizing the MI between current and future states (Anand et al., 2019; Oord et al., 2018; Stooke et al., 2020). Much prior work learns latent forward dynamics models (Watter et al., 2015; Karl et al., 2016; Zhang et al., 2018b; Hafner et al., 2019; Lee et al., 2019), related to Eq. 2. Multi-step inverse models, closely related to Eq. 4 have been used to learn control-centric representations (Yu et al., 2019; Gregor et al., 2016). Single-step inverse models have been deployed as regularization of forward models (Zhang et al., 2018a; Agrawal et al., 2016) and as an auxiliary loss for policy gradient RL Shelhamer et al. (2016); Pathak et al. (2017). The MI objectives that we study have also been used as reward bonuses to improve exploration, without impacting the representation, in the form of empowerment (Klyubin et al., 2008; 2005; Leibfried et al., 2019) and information-theoretic curiosity (Still & Precup, 2012).

Representation learning for reinforcement learning. In RL, the problem of finding a compact state space has been studied as state aggregation or abstraction (Bean et al., 1987; Li et al., 2006). Abstraction schemes include bisimulation (Givan et al., 2003), homomorphism (Ravindran & Barto, 2003), utile distinction (McCallum, 1996), and policy irrelevance (Jong & Stone, 2005). While efficient algorithms exist for MDPs with known transition models for some abstraction schemes such as bisimulation (Ferns et al., 2006; Givan et al., 2003), in general obtaining error-free abstractions is highly impractical for most problems of interest. For approximate abstractions prior work has bounded the sub-optimality of the policy (Bertsekas et al., 1988; Dean & Givan, 1997; Abel et al., 2016) as well as the sample efficiency (Lattimore & Szepesvari, 2019; Van Roy & Dong, 2019; Du et al., 2019), with some results extending to the deep learning setting (Gelada et al., 2019; Nachum et al., 2018). In this paper, we focus on whether a representation can be used to learn the optimal policy, and not the tractability of learning. Alternatively, priors based on the structure of the physical world can be used to guide representation learning (Jonschkowski & Brock, 2015). In deep RL, many auxiliary objectives distinct from the objectives that we study have been proposed, including meta-learning general value functions (Veeriah et al., 2019), predicting multiple value functions (Bellemare et al., 2019; Fedus et al., 2019; Jaderberg et al., 2016) and predicting domain-

specific measurements (Mirowski, 2019; Dosovitskiy & Koltun, 2016). We restrict our analysis to objectives that can be expressed as MI-maximization.

3 REPRESENTATION LEARNING FOR RL

The goal of representation learning for RL is to learn a compact representation of the state space that discards irrelevant and redundant information. In this section we formalize each part of this statement, starting with defining the RL problem and representation learning in the context of RL. We then propose and define the metric of sufficiency to evaluate the usefulness of a representation.

3.1 PRELIMINARIES

We begin with brief preliminaries of reinforcement learning and mutual information.

Reinforcement learning. A Markov decision process (MDP) is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ the set of actions, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ the state transition distribution, and $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ the reward function. We will use capital letters to refer to random variables and lower case letters to refer to values of those variables (e.g., S is the random variable for the state and $\mathbf{s}$ is a specific state). Throughout our analysis we will often be interested in multiple reward functions, and denote a set of reward functions as $\mathcal{R}$. The objective of RL is to find a policy that maximizes the sum of discounted returns $\bar{R}$ for a given reward function r, and we denote this optimal policy as $\pi^*_r = \arg\max_\pi \mathbb{E}_\pi[\sum_t r(S_t, A_t)]$. We also define the optimal Q-function as $Q^*_r(\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}_{\pi^*}[\sum_{t=1}^{\infty} r(S_t, A_t)|\mathbf{s}_t, \mathbf{a}_t]$. The optimal Q-function satisfies the recursive Bellman equation, $Q^*_r(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \mathbb{E}_{p(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t)} \arg\max_{\mathbf{a}_{t+1}} Q^*_r(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})$. The optimal policy and the optimal Q-function are related according to $\pi^*(\mathbf{s}) = \arg\max_{\mathbf{a}} Q^*(\mathbf{s}, \mathbf{a})$.

Mutual information. In information theory, the mutual information (MI) between two random variables, X and Y, is defined as (Cover & Thomas, 1999):

$$I(X;Y) = \mathbb{E}_{p(x,y)} \log \frac{p(x,y)}{p(x)p(y)} = H(X) - H(X|Y). \tag{1}$$

The first definition indicates that MI can be understood as a relative entropy (or KL-divergence), while the second underscores the intuitive notion that MI measures the reduction in the uncertainty of one random variable from observing the value of the other.

Representation learning for RL. The goal of representation learning for RL is to find a compact representation of the state space that discards details in the state that are not relevant for representing the policy or value function, while preserving task-relevant information (see Figure 1). While state aggregation methods typically define deterministic rules to group states in the representation (Bean et al., 1987; Li et al., 2006), MI-based representation learning methods used for deep RL treat the representation as a random variable (Nachum et al., 2018; Oord et al., 2018; Pathak et al., 2017). Accordingly, we formalize a representation as a stochastic mapping between original state space and representation space.

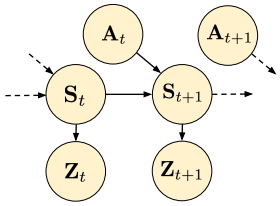


Figure 1: Probabilistic graphical model illustrating the state representation learning problem, estimating state representation Z from original state S.

Definition 1. *A **stochastic representation** $\phi_{\mathcal{Z}}(\mathbf{s})$ is a mapping from states $\mathbf{s} \in \mathcal{S}$ to a probability distribution $p(Z|S = \mathbf{s})$ over elements of a new representation space $z \in \mathcal{Z}$.*

In this work we consider learning state representations from data by maximizing an objective $\mathbb{J}$. Given an objective $\mathbb{J}$, we define the set of representations that maximize this objective as $\Phi_{\mathbb{J}} = \{\phi_{\mathcal{Z}}\}$ s.t. $\phi_{\mathcal{Z}} \in \arg\max \mathbb{J}(\phi)$.

Unlike problem formulations for partially observed settings (Watter et al., 2015; Hafner et al., 2019; Lee et al., 2019), we assume that S is a Markovian state; therefore the representation for a given

state is conditionally independent of the past states, a common assumption in the state aggregation literature (Bean et al., 1987; Li et al., 2006). See Figure 1 for a depiction of the graphical model.

3.2 Sufficient Representations for Reinforcement Learning

We now turn to the problem of evaluating stochastic representations for RL. Intuitively, we expect a useful state representation to be capable of representing the optimal policy in the original state space.

Definition 2. *A representation $\phi_{\mathcal{Z}}$ is π^***-sufficient** with respect to a set of reward functions $\mathcal{R}$ if* $\forall r \in \mathcal{R}$, $\phi_{\mathcal{Z}}(\mathbf{s}_1) = \phi_{\mathcal{Z}}(\mathbf{s}_2) \implies \pi^*_r(A|\mathbf{s}_1) = \pi^*_r(A|\mathbf{s}_2)$.

When a stochastic representation $\phi_{\mathcal{Z}}$ produces the same distribution over the representation space for two different states $\mathbf{s}_1$ and $\mathbf{s}_2$ we say it *aliases* these states. Unfortunately, as already proven in Theorem 4 of Li et al. (2006) for the more restrictive case of deterministic representations, being able to represent the optimal policy does not guarantee that it can be learned via RL in the representation space. Accordingly, we define a stricter notion of sufficiency that *does* guarantee the convergence of Q-learning to the optimal policy in the original state space (refer to Theorem 4 of Li et al. (2006) for the proof of this).

Definition 3. *A representation $\phi_{\mathcal{Z}}$ is Q^***-sufficient** with respect to a set of reward functions $\mathcal{R}$ if* $\forall r \in \mathcal{R}$, $\phi_{\mathcal{Z}}(S_1) = \phi_{\mathcal{Z}}(S_2) \implies \forall \mathbf{a}, Q^*_r(\mathbf{a}, \mathbf{s}_1) = Q^*_r(\mathbf{a}, \mathbf{s}_2)$.

Note that Q^*-sufficiency implies π^*-sufficiency since the optimal policy and the optimal Q-function are directly related via $\pi^*_r(s) = \arg\max_a Q^*_r(s, a)$ (Sutton & Barto, 2018); however the converse is not true. We emphasize that while Q^*-sufficiency guarantees convergence, it does not guarantee tractability, which has been explored in prior work (Lattimore & Szepesvari, 2019; Du et al., 2019).

We will further say that an *objective* $\mathbb{J}$ is sufficient with respect to some set of reward functions $\mathcal{R}$ if all the representations that maximize that objective $\Phi_{\mathbb{J}}$ are sufficient with respect to every element of $\mathcal{R}$ according to the definition above. Surprisingly, we will demonstrate that not all commonly used objectives satisfy this basic qualification even when $\mathcal{R}$ contains a single known reward function.

4 Mutual Information for Representation Learning in RL

In our study, we consider several MI objectives proposed in the literature.

Forward information: A commonly sought characteristic of a state representation is to ensure it retains maximum predictive power over future state representations. This property is satisfied by representations maximizing the following MI objective,

$$\mathbb{J}_{fwd} = I(Z_{t+k}; Z_t, A_t) = H(Z_{t+k}) - H(Z_{t+k}|Z_t, A_t). \tag{2}$$

We suggestively name this objective "forward information" due to the second term, which is the entropy of the forward dynamics distribution. This objective is related to that proposed in Nachum et al. (2018), where they consider a sequence of actions.

State-only transition information: Several popular methods (Oord et al., 2018; Anand et al., 2019; Stooke et al., 2020) optimize a similar objective, but do not include the action [1]:

$$\mathbb{J}_{state} = I(Z_{t+k}; Z_t) = H(Z_{t+k}) - H(Z_{t+k}|Z_t). \tag{3}$$

As we will show, the exclusion of the action can have a profound effect on the characteristics of the resulting representations.

[1] It is common to use a history of states, optimizing an objective like $I(Z_{t+k}; Z_1, ..., Z_t)$. In our setting we assume the given state is Markovian, in which case the objectives are equivalent.

Inverse information: Another commonly sought characteristic of state representations is to retain maximum predictive power of the action distribution that could have generated an observed transition from $\mathbf{s}_t$ to $\mathbf{s}_{t+1}$. Such representations can be learned by maximizing the following information theoretic objective:

$$\mathbb{J}_{inv} = I(A_t; Z_{t+k}|Z_t) = H(A_t|Z_t) - H(A_t|Z_t, Z_{t+k}) \tag{4}$$

We suggestively name this objective "inverse information" due to the second term, which is the entropy of the inverse dynamics. A wide range of prior work learns representations by optimizing closely related objectives (Gregor et al., 2016; Shelhamer et al., 2016; Agrawal et al., 2016; Pathak et al., 2017; Yu et al., 2019; Zhang et al., 2018a). Intuitively, inverse models allow the representation to capture only the elements of the state that are necessary to predict the action, allowing the discard of potentially irrelevant information.

5 Sufficiency Analysis

In this section we analyze the sufficiency for control of representations obtained by maximizing each objective presented in Section 4. To focus on the representation learning problem, we decouple it from RL by assuming access to a dataset of transitions collected with a uniform random policy that reaches all states, which can then be used to learn the desired representation. We also assume that distributions, such as the dynamics or inverse dynamics, can be modeled with arbitrary accuracy, and that the maximizing set of representations for a given objective can be computed. While these assumptions might be relaxed in any practical RL algorithm, and exploration plays a confounding role, studying these objectives under such simplifying assumptions allows us to compare them in terms of sufficiency on an equal playing field, isolating the role of representation learning from other confounding components of a complete RL algorithm.

5.1 Forward Information

In this section we show that a representation that maximizes $\mathbb{J}_{fwd}$ is sufficient for optimal control under any reward function. This result aligns with intuition that a representation that captures forward dynamics can represent everything predictable in the state space, and can thus be used to learn the optimal policy for any task. Note that this strength can also be a weakness if there are many predictable elements that are irrelevant for downstream tasks, since the representation retains more information than is needed for the task.

Proposition 1. *$\mathbb{J}_{fwd}$ is sufficient for all reward functions.*

Proof. (Sketch) We first show that if Z_t, A_t are maximally informative of Z_{t+k}, they are also maximally informative of the return $\bar{R}_t$. Due to the Markov structure, $\mathbb{E}_{p(Z_t|S_t=\mathbf{s})} p(\bar{R}_t|Z_t, A_t) = p(\bar{R}_t|S_t = \mathbf{s}, A_t)$. In other words, given $\phi_{\mathcal{Z}}$, additionally knowing S doesn't change our belief about the future return. The Q-value is the expectation of the return, so Z has as much information about the Q-value as S does. The full proof can be found in Appendix 8.1. □

5.2 State-only Transition Information

While $\mathbb{J}_{state}$ is closely related to $\mathbb{J}_{fwd}$, we now show that it is not sufficient.

Proposition 2. *$\mathbb{J}_{state}$ is not sufficient for all reward functions.*

Proof. Consider the counter-example in Figure 2. Here, the agent starts in $\mathbf{s}_0$ and can choose between $\mathbf{a}_0$ leading to $\mathbf{s}_1$ with a positive reward and $\mathbf{a}_1$ leading to $\mathbf{s}_2$ with negative reward. The current state does not contain any information about which of the next two states is more likely, so a representation maximizing $\mathbb{J}_{state}$ is free to alias those states. An alternative view is that such a representation can maximize $\mathbb{J}_{state} = H(Z_{t+k}) - H(Z_{t+k}|Z_t)$ by reducing both terms in equal amounts - aliasing $\mathbf{s}_1$ and $\mathbf{s}_2$ decreases the marginal entropy as well as the entropy of predicting the next state starting from $\mathbf{s}_0$. However, this aliased representation is not capable of representing the optimal policy which must distinguish between $\mathbf{s}_1$ and $\mathbf{s}_2$ in order to choose the optimal action. □

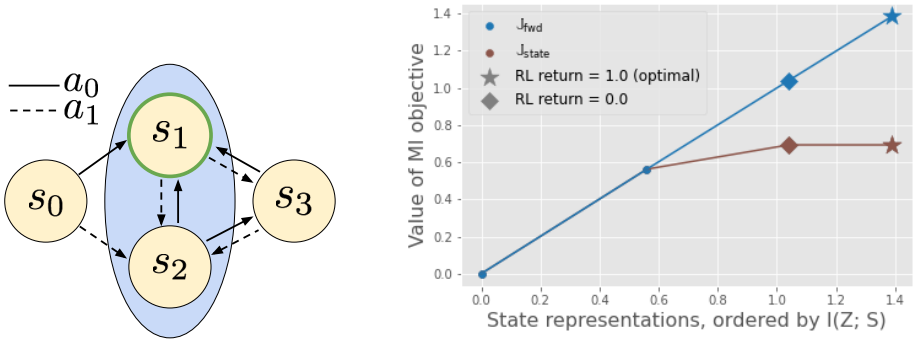


Figure 2: (left) A representation that aliases the states $\mathbf{s}_1$ and $\mathbf{s}_2$ into a single state maximizes $\mathbb{J}_{state}$ but is not sufficient to represent the optimal policy which must distinguish $\mathbf{s}_1$ as the high-reward state (outlined in green) from the low-reward state $\mathbf{s}_2$. (right) Values of $\mathbb{J}_{state}$ and $\mathbb{J}_{fwd}$ for a few representative state representations, ordered by increasing $I(Z;S)$. The representation that aliases $\mathbf{s}_1$ and $\mathbf{s}_2$ (plotted with a diamond) maximizes $\mathbb{J}_{inv}$, but the policy learned with this representation may not be optimal (as shown here). The original state representation (plotted with a star) is sufficient.

5.3 Inverse Information

Here we show that representations that maximize $\mathbb{J}_{inv}$ are not sufficient for control in all MDPs. Intuitively, one way that the representation can be insufficient is by retaining only controllable state elements, while the reward function depends on state elements outside the agent's control. We then proceed to show that additionally representing the immediate reward is not enough to resolve this issue.

Proposition 3. *$\mathbb{J}_{inv}$ is not sufficient for all reward functions. Additionally, adding $I(R_t; Z_t)$ to the objective does not make it sufficient.*

Proof. Consider the MDP illustrated in Figure 3, and the representation that aliases the starting states $\mathbf{s}_0$ and $\mathbf{s}_1$. The same actions taken from these states lead to different next states which may have different rewards ($\mathbf{a}_0$ leads to the reward from $\mathbf{s}_0$ while $\mathbf{a}_1$ leads to the reward from $\mathbf{s}_1$). However, this representation maximizes $\mathbb{J}_{inv}$ because given each pair of states, the action is identifiable. Interestingly, this problem cannot be remedied by simply requiring that the representation also be capable of predicting immediate rewards. The same counterexample in Figure 3 holds since we assumed $\mathbf{s}_0$ and $\mathbf{s}_1$ have the same reward. □

6 Experiments

In this section, we present experiments studying MI-based representation learning with image observations, to analyze whether the conclusions of our theoretical analysis hold in practice. Our goal is not to show that any particular method is necessarily better or worse, but rather to illustrate that the sufficiency arguments that we presented translate into quantifiable performance differences in the deep RL setting.

6.1 Experimental Setup

To separate representation learning from RL, we first optimize each representation learning objective on a dataset of offline data. We then freeze the weights of the state encoder learned in the first phase and train RL agents with the representation as state input. To clearly illustrate the characteristics of each objective, we use the simple pygame (Shinners, 2011) video game Catcher, in which the agent controls a paddle that it can move back and forth to catch fruit that falls from the top of the screen (see Figure 5). A positive reward is given when the fruit is caught and a negative reward when the fruit is not caught. The episode terminates after one piece of fruit falls. We optimize $\mathbb{J}_{fwd}$ and $\mathbb{J}_{state}$ with

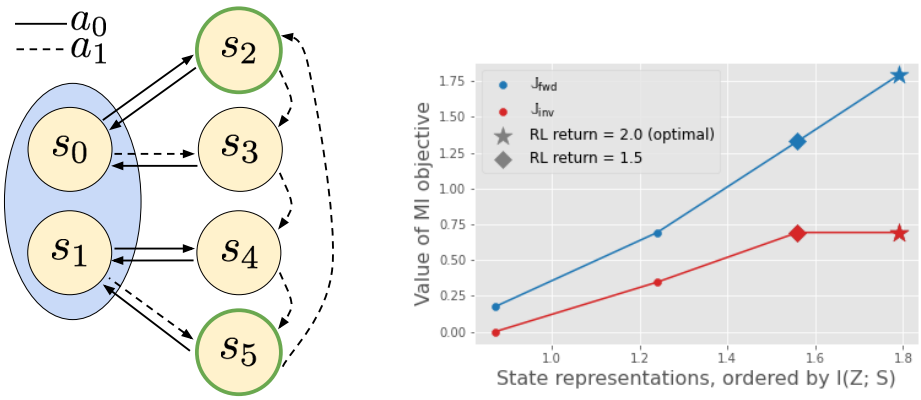


Figure 3: (left) In this MDP, a representation that aliases the states $\mathbf{s}_0$ and $\mathbf{s}_1$ into a single state maximizes $\mathbb{J}_{inv}$, yet is not sufficient to represent the optimal policy, which must distinguish between $\mathbf{s}_0$ and $\mathbf{s}_1$ in order to take a different action (towards the high-reward states outlined in green). (right) Values of $\mathbb{J}_{inv}$ and $\mathbb{J}_{fwd}$ for a few selected state representations, ordered by increasing $I(Z;S)$. The representation that aliases $\mathbf{s}_0$ and $\mathbf{s}_1$ (plotted with a diamond) maximizes $\mathbb{J}_{inv}$, but is not sufficient to learn the optimal policy. Note that this counterexample holds also for $\mathbb{J}_{inv} + I(R;Z)$ since $I(R;Z)$ does not decrease when $\mathbf{s}_0$ and $\mathbf{s}_1$ are aliased.

noise contrastive estimation (Gutmann & Hyvärinen, 2010), and $\mathbb{J}_{inv}$ by training an inverse model via maximum likelihood. We train all methods on an offline dataset of 50k transitions collected from a uniform random policy.

6.2 Computational Results

In principle, we expect that a representation learned with $\mathbb{J}_{inv}$ may not sufficient to solve the *catcher* game. Because the agent does not control the fruit, a representation maximizing $\mathbb{J}_{inv}$ might discard that information, thereby making it impossible to represent the optimal policy. We observe in Figure 5 (top) that indeed representations trained to maximize $\mathbb{J}_{inv}$ result in RL agents that converge slower and to a lower asymptotic expected return. Further, attempting to learn a decoder from the learned representation to the position of the falling fruit incurs a high error (Figure 5, bottom), indicating that the fruit is not precisely captured by the representation. We argue that this type of problem setting is not contrived, and is representative of many situations in realistic tasks. Consider, for instance, an autonomous vehicle that is stopped at a stoplight. Because the agent does not control the color of the stoplight, it may not be captured in the representation learned by $\mathbb{J}_{inv}$ and the resulting RL policy may choose to run the light.

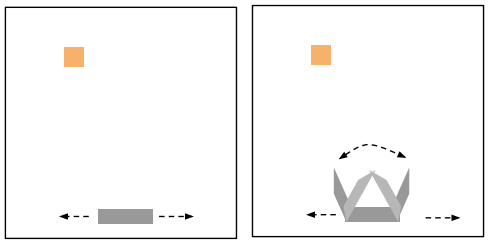
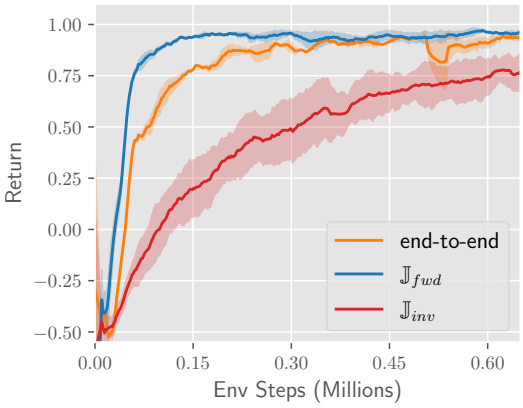


Figure 4: (left) Original *catcher* game in which the agent (grey paddle) moves left or right to catch fruit (yellow square) that falls from the top of the screen. (right) Variation *catcher-grip* in which the agent is instantiated as a gripper, and must open the gripper to catch fruit.

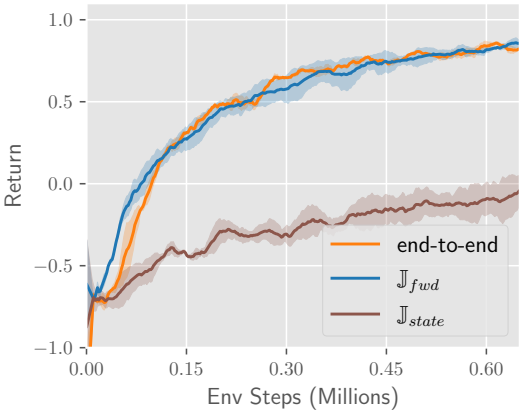
In the second experiment, we consider a failure mode of $\mathbb{J}_{state}$. We augment the paddle with a gripper that the agent controls and must be open in order to properly catch the fruit. Since the change in the gripper is completely controlled by a single action, the current state contains no information about the state of the gripper in the future. Therefore, a representation maximizing $\mathbb{J}_{state}$ might alias states where the gripper is open with states where the gripper is closed. In our experiment, we see that the error in predicting the state of the gripper from the representation learned via $\mathbb{J}_{state}$ is chance (Figure 6, bottom). This degrades the performance of an RL agent trained with this state representation since the best the agent can do is move under the fruit and randomly open or close the gripper (Figure 6, top). In the driving example, suppose turning on the headlights incurs positive

reward if it's raining but negative reward if it's sunny. The representation could fail to distinguish the state of the headlights, making it impossible to learn when to properly use the headlights.



Objective	Agent Err.	Fruit Err.
$\mathbb{J}_{fwd}$	0.1	0.1
$\mathbb{J}_{inv}$	0.15	0.47

Figure 5: *catcher* environment: Representations trained to maximize $\mathbb{J}_{inv}$ need not represent the fruit, leading to poor performance when used to train an RL agent to play the game (top). Attempting to decode the position of the fruit incurs high error (bottom).



Obj.	Agent Err.	Fruit Err.	Grip Err.
$\mathbb{J}_{fwd}$	0.1	0.1	0.3
$\mathbb{J}_{inv}$	0.1	0.1	0.5

Figure 6: *gripper* environment: Representations trained to maximize $\mathbb{J}_{state}$ need not represent the gripper, leading to poor performance when used as a state representation for RL (top). Attempting to decode the state of the gripper incurs chance accuracy (bottom).

$\mathbb{J}_{fwd}$ produces useful representations in all cases, and is equally or more effective than learning representations purely from the RL objective alone (as in Figure 5). In environments with more visual complexity, we expect representations learned with $\mathbb{J}_{fwd}$ to yield larger gains over learning representations end-to-end via RL.

7 Discussion

In this work, we aimed to analyze mutual information representation learning objectives for control from a theoretical perspective. In contrast to much prior work that studies how these objectives can be effectively optimized given high-dimensional observations, we analyze which objectives are guaranteed to yield representations that are actually sufficient for learning the optimal policy. Surprisingly, we show that two common objectives yield representations that are theoretically insufficient, and provide a proof of sufficiency for a third. We validate our theoretical results with an empirical investigation on a simple video game environment, and show that the insufficiency of these objectives can degrade the performance of deep RL agents.

We view this investigation as a step forward in understanding the theoretical characteristics of representation learning techniques commonly used in deep RL. We see many exciting avenues for future work. First, identifying more restrictive MDP classes in which insufficient objectives are in fact sufficient, and relating these to realistic applications. Second, investigating if sample complexity bounds can be established in the case of a sufficient objective. Third, extending our analysis to the partially observed setting, which is more reflective of practical applications. We see these directions as fruitful in providing a deeper understanding of the learning dynamics of deep RL, and potentially yielding novel algorithms for provably accelerating RL with representation learning.

REFERENCES

David Abel, D Ellis Hershkowitz, and Michael L Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pp. 2915–2923, 2016.

Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. In *Advances in neural information processing systems*, pp. 5074–5082, 2016.

Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. Unsupervised state representation learning in atari. In *Advances in Neural Information Processing Systems*, pp. 8766–8779, 2019.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15509–15519, 2019.

James C Bean, John R Birge, and Robert L Smith. Aggregation in dynamic programming. *Operations Research*, 35(2):215–220, 1987.

Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Marc Bellemare, Will Dabney, Robert Dadashi, Adrien Ali Taiga, Pablo Samuel Castro, Nicolas Le Roux, Dale Schuurmans, Tor Lattimore, and Clare Lyle. A geometric perspective on optimal representations for reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 4360–4371, 2019.

Dimitri P Bertsekas, David A Castanon, et al. Adaptive aggregation methods for infinite horizon dynamic programming. 1988.

Thomas M Cover and A Joy Thomas. *Elements of information theory*. John Wiley & Sons, 1999.

Thomas Dean and Robert Givan. Model minimization in markov decision processes. In *AAAI/IAAI*, pp. 106–111, 1997.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy and Vladlen Koltun. Learning to act by predicting the future. In *International Conference on Learning Representations*, 2016.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.

Norm Ferns, Pablo Samuel Castro, Doina Precup, and Prakash Panangaden. Methods for computing state similarity in markov decision processes. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 174–181, 2006.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, pp. 2170–2179, 2019.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pp. 2555–2565, 2019.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

Nicholas K Jong and Peter Stone. State abstraction discovery from irrelevant state variables. In *IJCAI*, volume 8, pp. 752–757, 2005.

Rico Jonschkowski and Oliver Brock. Learning state representations with robotic priors. *Autonomous Robots*, 39(3):407–428, 2015.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pp. 651–673, 2018.

Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. In *International Conference on Learning Representations*, 2016.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135. IEEE, 2005.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Keep your options open: An information-based driving principle for sensorimotor systems. *PloS one*, 3(12), 2008.

Tor Lattimore and Csaba Szepesvari. Learning with good feature representations in bandits and in rl with a generative model. *arXiv preprint arXiv:1911.07676*, 2019.

Alex X. Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *arXiv preprint arXiv:1907.00953*, 2019.

Felix Leibfried, Sergio Pascual-Díaz, and Jordi Grau-Moya. A unified bellman optimality principle combining reward maximization and empowerment. In *Advances in Neural Information Processing Systems*, pp. 7867–7878, 2019.

Timothée Lesort, Natalia Díaz-Rodríguez, Jean-Franois Goudou, and David Filliat. State representation learning for control: An overview. *Neural Networks*, 108:379–392, 2018.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. In *ISAIM*, 2006.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

David McAllester and Karl Statos. Formal limitations on the measurement of mutual information. *arXiv preprint arXiv:1811.04251*, 2018.

Andrew Kachites McCallum. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, University of Rochester, 1996.

Piotr Mirowski. Learning to navigate. In *1st International Workshop on Multimodal Understanding and Learning for Embodied Applications*, pp. 25–25, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Near-optimal representation learning for hierarchical reinforcement learning. 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2778–2787. JMLR. org, 2017.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

Mirco Ravanelli and Yoshua Bengio. Learning speaker representations with mutual information, 2019.

Balaraman Ravindran and Andrew G Barto. Smdp homomorphisms: an algebraic approach to abstraction in semi-markov decision processes. In *Proceedings of the 18th international joint conference on Artificial intelligence*, pp. 1011–1016, 2003.

Evan Shelhamer, Parsa Mahmoudieh, Max Argus, and Trevor Darrell. Loss is its own reward: Self-supervision for reinforcement learning. *arXiv preprint arXiv:1612.07307*, 2016.

Pete Shinners. Pygame. `http://pygame.org/`, 2011.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. *arXiv preprint arXiv:2009.08319*, 2020.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Benjamin Van Roy and Shi Dong. Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

Vivek Veeriah, Matteo Hessel, Zhongwen Xu, Janarthanan Rajendran, Richard L Lewis, Junhyuk Oh, Hado P van Hasselt, David Silver, and Satinder Singh. Discovery of useful questions as auxiliary tasks. In *Advances in Neural Information Processing Systems*, pp. 9306–9317, 2019.

Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in neural information processing systems*, pp. 2746–2754, 2015.

Tianhe Yu, Gleb Shevchuk, Dorsa Sadigh, and Chelsea Finn. Unsupervised visuomotor control through distributional planning networks. In *Robotics Science and Systems*, 2019.

Amy Zhang, Harsh Satija, and Joelle Pineau. Decoupling dynamics and reward for transfer learning. *arXiv preprint arXiv:1804.10689*, 2018a.

Marvin Zhang, Sharad Vikram, Laura Smith, Pieter Abbeel, Matthew J Johnson, and Sergey Levine. Solar: Deep structured latent representations for model-based reinforcement learning. *arXiv preprint arXiv:1808.09105*, 2018b.

8 APPENDIX

We describe the proofs for the sufficiency results from Section 5 here. We begin by providing a set of lemmas, before proving the sufficiency of $\mathbb{J}_{fwd}$.

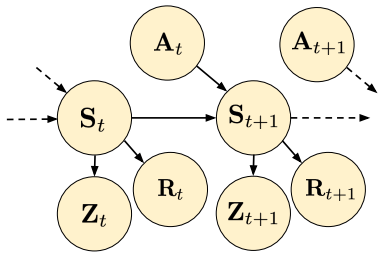


Figure 7: Graphical model showing random variables true states S, states in the representation Z, actions A, and rewards R.

Lemma 1. *If* $I(Z_{t+k}; Z_t, A_t) = I(S_{t+k}; S_t, A_t) \forall k$, *then* $I(R_{t+k}; Z_t, A_t) = I(R_{t+k}; S_t, A_t) \forall k$.

Proof. First note that $I(Z_{t+k}; Z_t, A_t) \leq I(S_{t+k}; S_t, A_t)$ for all representations $\phi_{\mathcal{Z}}$. This follows from the graphical model in Figure 1 because S is a Markov state and Z is a function of S. For proof by contradiction, assume there is some $\phi_{\mathcal{Z}}$ and some r such that $I(Z_t, A_t; R_{t+k}) < I(S_t, A_t; R_{t+k})$ and that $I(Z_{t+k}; Z_t, A_t) = I(S_{t+k}; S_t, A_t)$. Now we know by the data processing inequality (DPI) that $I(Z_t, A_t; Z_{t+k}) \leq I(Z_t, A_t; S_{t+k}) \leq I(S_t, A_t; S_{t+k})$. We will proceed by showing that that $I(Z_t, A_t; R_{t+k}) < I(S_t, A_t; R_{t+k}) \implies I(Z_t, A_t; S_{t+k}) < I(S_t, A_t; S_{t+k})$, which gives the needed contradiction.

Using chain rule, we can expand the following expression in two different ways.

$$\begin{aligned} I(Z_t, S_t, A_t; R_{t+k}) =& I(Z_t; R_{t+k}|S_t, A_t) + I(S_t, A_t; R_{t+k}) \\ =& I(S_t; R_{t+k}|Z_t, A_t) + I(Z_t, A_t; R_{t+k}) \end{aligned}$$

Note that the first term in the first expansion is zero by conditional independence. To satisfy $I(Z_t, A_t; R_{t+k}) < I(S_t, A_t; R_{t+k})$ we must have that $I(S_t; R_{t+k}|Z_t, A_t) > 0$. Again by chain rule,

$$\begin{aligned} I(Z_t, S_t, A_t; S_{t+k}) =& I(Z_t; S_{t+k}|S_t, A_t) + I(S_t, A_t; S_{t+k}) \\ =& I(S_t; S_{t+k}|Z_t, A_t) + I(Z_t, A_t; S_{t+k}) \end{aligned}$$

Again the first term in the first expansion is zero by conditional independence. Comparing the first term in the second expansion with the first term in the second expansion in Equation 8, we see by DPI that $I(S_t; S_{t+k}|Z_t, A_t) \geq I(S_t; R_{t+k}|Z_t, A_t)$. Therefore we must have $I(S_t; S_{t+k}|Z_t, A_t) > 0$. Plugging this fact back into Equation 8 above, this implies that $I(Z_t, A_t; S_{t+k}) < I(S_t, A_t; S_{t+k})$, which is exactly what we wanted to show. □

Lemma 2. *If* $I(X; A) = I(Y; A)$ *and* $I(X; B) = I(Y; B)$ *then* $I(X; A + B) = I(Y; A + B)$

Proof.

$$\begin{aligned} I(X; A+B) &= H(A+B) - H(A+B|X) \\ &\geq H(A+B) - H(A|X) - H(B|X) \\ &= H(A+B) - H(A|Y) - H(B|Y) \\ &= I(Y; A+B) \\ I(Y; A+B) &= H(A+B) - H(A+B|Y) \\ &\geq H(A+B) - H(A|Y) - H(B|Y) \\ &= H(A+B) - H(A|X) - H(B|X) \\ &= I(X; A+B) \end{aligned}$$

Since $I(X; A+B) \geq I(Y; A+B)$ and $I(Y; A+B) \geq I(X; A+B)$, we must have $I(X; A+B) = I(Y; A+B)$. □

Lemma 3. *If $I(Y;Z) = I(Y;X)$ and $Y \perp Z|X$, then $\exists p(Z|X)$ s.t. $\forall x, p(Y|X = x) = \int p(Y|Z)p(Z|X = x)dz$.*

Proof. First note that the statement is not trivially true. Without any assumption regarding MI, we can write,

$$p(Y|X = x) = \int p(Y, Z|X = x)dz = \int p(Y|Z, X = x)p(Z|X = x)dz$$

Comparing this with the statement we'd like to prove, we can see that the key idea is to show that the MI equivalence implies that $p(Y|Z, X = x) = p(Y|Z)$. To begin, consider $I(Y; Z) = I(Y; X)$. We can re-write this equality using the entropy definition of MI.

$$H(Y) - H(Y|Z) = H(Y) - H(Y|X)$$

Note that the $H(Y)$ cancel and substituting the definition of entropy we have,

$$\mathbb{E}_{p(Y,Z)}[\log p(Y|Z)] = \mathbb{E}_{p(Y,X)}[\log p(Y|X)]$$

Note that on the right-hand side, we can use the Tower property to re-write the expectation as,

$$\begin{aligned}\mathbb{E}_{p(Y,X)}[\log p(Y|X)] &= \mathbb{E}_{p(Z)}\mathbb{E}_{p(Y,X|Z)}[\log p(Y|X)] \\ &= \mathbb{E}_{p(Y|X)p(X,Z)}[\log p(Y|X)]\end{aligned}$$

Now we can use the Tower property again to re-write the expectation on both sides,

$$\begin{aligned}\mathbb{E}_{p(X)}[\mathbb{E}_{p(Y,Z|X)}[\log p(Y|Z)]] &= \mathbb{E}_{p(X)}[\mathbb{E}_{p(Y|X)p(X,Z|X)}[\log p(Y|X)]] \\ \mathbb{E}_{p(X)}[\mathbb{E}_{p(Y|X)p(Z|X)}[\log p(Y|Z)]] &= \mathbb{E}_{p(X)}[\mathbb{E}_{p(Y|X)p(Z|X)}[\log p(Y|X)]] \\ \mathbb{E}_{p(X)}[\mathbb{E}_{p(Y|X)p(Z|X)}[\log p(Y|Z)]] - \log p(Y|X)] &= 0 \\ \mathbb{E}_{p(X)}[\mathbb{E}_{p(Y|X)}[\mathbb{E}_{p(Z|X)}[\log p(Y|Z)] - \log p(Y|X)]] &= 0\end{aligned}$$

Log probabilities are always ≤ 0, therefore for the sum to equal zero, each term must be zero.

$$\log p(Y|X) = \mathbb{E}_{p(Z|X)}[\log p(Y|Z)]$$

By Jensen's inequality,

$$\log p(Y|X) \leq \log \mathbb{E}_{p(Z|X)}[p(Y|Z)]$$

By the monotonicity of the logarithm,

$$p(Y|X) \leq \mathbb{E}_{p(Z|X)}[p(Y|Z)]$$

If there exists some x and some y such that $p(Y = y|X = x) < \mathbb{E}_{p(Z|X=x)}[p(Y = y|Z)]$, then there must be some other y' for the same x where $p(Y = y'|X = x) > \mathbb{E}_{p(Z|X=x)}[p(Y = y'|Z)]$ because p$(Y|X = x)$ must sum to 1. Therefore,

$$p(Y|X) = \mathbb{E}_{p(Z|X)}[p(Y|Z)] = \int p(Y|Z)p(Z|X = x)dz = \int p(Y, Z|X = x)dz$$

Where the last equality follows by conditional independence of Y and Z given X. □

Given the lemmas stated above, we can then use them to prove the sufficiency of $\mathbb{J}_{fwd}$.

8.1 PROOF OF SUFFICIENCY OF $\mathbb{J}_{fwd}$

Proposition 4. *(Sufficiency of $\mathbb{J}_{fwd}$) For a representation $\phi_{\mathcal{Z}}$ and set of reward functions $\mathcal{R}$, if $I(Z_t, A_t; Z_{t+k})$ is maximized $\forall k > 0, t > 0$ then $\forall r \in \mathcal{R}$, $\phi_{\mathcal{Z}}(S_1) = \phi_{\mathcal{Z}}(S_2) \implies \forall \mathbf{a}, Q^*_r(\mathbf{a}, \mathbf{s}_1) = Q^*_r(\mathbf{a}, \mathbf{s}_2)$.*

Proof. See Figure 7 for the relationship of the variables of interest. We assume that the policy $p(\mathbf{a}|\mathbf{s})$ is uniform random in S space, and that $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ is the true environment dynamics in S space. Note that each state $\mathbf{z}_t$ in the representation is dependent only on $\mathbf{s}_t$, via $p(\mathbf{z}_t|\mathbf{s}_t)$. While typically rewards are a deterministic function of states and actions, here we cast them as random variables R_t for simplicity. First, the MI objective is maximized if the representation is taken to be the identity, in other words $\max_\phi I(Z_t, A_t; Z_{t+k}) = I(S_t, A_t; S_{t+k})$. From here, the result will follow by combining the preceding lemmas. We know from Lemma 1 that

$$I(Z_t, A_t; Z_{t+k}) = I(S_t, A_t; S_{t+k}) \qquad \implies \qquad I(R_{t+k}; Z_t, A_t) = I(R_{t+k}; S_t, A_t)$$

Define the random variable $\bar{R}_t$ to be the discounted return starting from state $\mathbf{s}_t$ and taking action $\mathbf{a}_t$.

$$\bar{R}_t = \sum_{k=1}^{H-t} \gamma^k R_{t+k}$$

By Lemma 2,

$$I(R_{t+k}; Z_t, A_t) = I(R_{t+k}; S_t, A_t) \forall k, t \qquad \implies \qquad I(\bar{R}_t; Z_t, A_t) = I(\bar{R}_t; S_t, A_t) \forall t$$

Now let $X = [S_t, A_t]$, $Y = \bar{R}_t$, and $Z = Z_t$, and note that by the structure of the graphical model in Figure 7, $Y \perp Z|X$. Plugging into Lemma 3 we then have that

$$\mathbb{E}_{p(\mathbf{z}_t|S_t=\mathbf{s})} p(\bar{R}_t|Z_t, A_t) = p(\bar{R}_t|S_t = \mathbf{s}, A_t)$$

Now the Q-function for a state-action pair $(\mathbf{s}, \mathbf{a})$ can be written as an expectation of this random variable $\bar{R}_t$ over the dynamics and policy distributions, assuming that $S_t = \mathbf{s}$ and $A = \mathbf{a}$.

$$Q(\mathbf{s}, \mathbf{a}) = \mathbb{E}_{\mathbf{s}_{t+1:H} \sim p(\mathbf{s}'|\mathbf{s}, \mathbf{a}), \mathbf{a}_{t+1:H} \sim p(\mathbf{a}|\mathbf{s})} [\bar{R}_t | S_t = \mathbf{s}, A_t = \mathbf{a}]$$

Since $\phi_{\mathcal{Z}}(\mathbf{s}_1) = \phi_{\mathcal{Z}}(\mathbf{s}_2)$, $p(\mathbf{z}_t|S_t = \mathbf{s}_1) = p(\mathbf{z}_t|S_t = \mathbf{s}_2)$. Therefore by the equation above, $Q^*_r(\mathbf{a}, \mathbf{s}_1) = Q^*_r(\mathbf{a}, \mathbf{s}_2)$.

□