Curiosity in Hindsight

Daniel Jarrett
DeepMind

Corentin Tallec
DeepMind

Florent Altché
DeepMind

Thomas Mesnard
DeepMind

Rémi Munos
DeepMind

Michal Valko
DeepMind

Abstract

Consider the problem of exploration in sparse-reward or reward-free environments, such as Montezuma's Revenge. The *curiosity-driven* paradigm dictates an intuitive technique: At each step, the agent is rewarded for how much the realized outcome differs from their predicted outcome. However, using predictive error as intrinsic motivation is prone to fail in *stochastic environments*, as the agent may become hopelessly drawn to high-entropy areas of the state-action space, such as a noisy TV. Therefore it is important to distinguish between aspects of world dynamics that are inherently *predictable* (for which errors reflect epistemic uncertainty) and aspects that are inherently *unpredictable* (for which errors reflect aleatoric uncertainty): The former should constitute a source of intrinsic reward, whereas the latter should not. In this work, we study a natural solution derived from structural causal models of the world: Our key idea is to learn representations of the future that capture precisely the unpredictable aspects of each outcome—not any more, not any less—which we use as additional input for predictions, such that intrinsic rewards do vanish in the limit. First, we propose incorporating such hindsight representations into the agent's model to disentangle "noise" from "novelty", yielding *Curiosity in Hindsight*: a simple and scalable generalization of curiosity that is robust to all types of stochasticity. Second, we implement this framework as a drop-in modification of any prediction-based exploration bonus, and instantiate it for the recently introduced BYOL-Explore algorithm as a prime example, resulting in the noise-robust "BYOL-Hindsight". Third, we illustrate its behavior under various stochasticities in a grid world, and find improvements over BYOL-Explore in hard-exploration Atari games with sticky actions. Importantly, we show state-of-the-art results in exploring Montezuma's Revenge with sticky actions, while preserving performance in the non-sticky setting.

1 Introduction

Learning to understand the world without supervision is a hallmark of intelligent behavior [1], and *exploration* is a key pillar of research in reinforcement learning agents [2]. How might an agent learn meaningful behaviors when external rewards are sparse or absent? A predominant approach is given by the *curiosity-driven* paradigm [3], in which an agent's ability to predict the future is used as a proxy for their "understanding" of the world. Maintaining a learned model of the environment, at each transition the agent receives an intrinsic reward proportional to how much the realized outcome differs from their predicted outcome—which naturally directs them towards new areas that have not been seen.

There are two primary hurdles. The first is *dimensionality*: While outcomes can be predicted directly at the level of observations [4–8], pixel-based losses have generally not been found to work well in conjunction with complex, high-dimensional spaces [9]. Popular solutions have taken to operating on lower-dimensional *latent representations*, such as frame-predictive features [10], inverse dynamics features [11], random features [12], or features that maximize mutual information across time [13]. Most recently, bootstrapped features are employed in BYOL-Explore [14]—which achieves superhuman performance on hard-exploration games in Atari with a much simpler design than comparable agents.

Deep Reinforcement Learning Workshop at NeurIPS 2022.

The second is *stochasticity*: Curiosity-driven exploratory agents are often susceptible to bad behavior in environments with stochastic transitions, since they are often hopelessly distracted by high-entropy elements in the state-action space [9]. A classic example is the problem of a "noisy TV" generating a stream of intrinsic rewards, around which predictive error-based agents become stuck indefinitely [15]. More broadly, this problem manifests with respect to any aspect of environment dynamics that is inherently unpredictable, including noise specific to certain states, or noise actively induced by the agent.

Novelty vs. Noise In the presence of stochasticity, predictive error *per se* is no longer a good measure for an agent's lack of "understanding" of the world. Intuitively, we wish to measure "understanding" by how much *epistemic* knowledge an agent has acquired (viz. necessary truths about how the world works in general), which is distinct from how much *aleatoric* variation each outcome can display (viz. contingent facts about how the world happens to be). Precisely, we want to distinguish between aspects of world dynamics that are inherently predictable—for which (reducible) errors stem from "novelty"—and aspects that are inherently unpredictable—for which (irreducible) errors stem from "noise". Crucially, while the former should constitute a source of intrinsic reward, the latter should not.

Contributions In this work, we operationalize this distinction by deriving a natural solution based on structural causal models of the world: Our key idea is to learn representations of the future that capture precisely the unpredictable aspects of each outcome—not any more, not any less—which we then use as additional input for predictions, such that intrinsic rewards indeed vanish in the limit. First, we propose incorporating such hindsight representations into the agent's model to disentangle "noise" from "novelty", yielding *Curiosity in Hindsight*: a simple and scalable generalization of curiosity that is robust to all types of stochasticity (Section 3). Second, we implement this framework as a drop-in modification of any prediction-based exploration bonus regardless of representation space, and instantiate it for BYOL-Explore, resulting in the noise-robust "BYOL-Hindsight" (Section 4). Third, we illustrate its behavior under various stochasticities in a grid world, and improve over BYOL-Explore in hard-exploration Atari games with sticky actions—a standard protocol for introducing stochasticity in training/evaluation in Atari. Importantly, we show state-of-the-art results in exploring Montezuma's Revenge with sticky actions, while preserving original performance in the non-sticky setting (Section 5).

2 Motivation

2.1 Problem Formalism

Consider the standard MDP setup. We use uppercase for random variables and lowercase for specific values: Let X denote the *state* variable, taking on values $x \in \mathcal{X}$, and A the *action* variable, taking on values $a \in \mathcal{A}$. While we keep our notation simple, we allow X to play the role of "contexts", "features", "beliefs", or "embeddings" depending on environment observability and design of the agent. Denote with $\tau \in \Delta(\mathcal{X})^{\mathcal{X}\times\mathcal{A}}$ the world dynamics such that $X_{t+1} \sim \tau(\cdot|x_t, a_t)$, and $\pi \in \Delta(\mathcal{A})^{\mathcal{X}}$ the agent's policy such that $A_t \sim \pi(\cdot|x_t)$. Finally, let ρ_π denote the distribution of states induced by π.

Definition 1 (Curiosity-driven Exploration) In this work, we focus on *predictive error-based* curiosity, which most popular approaches to curiosity fall under. Denote the intrinsic reward as follows:

$$\mathcal{R}_\eta(x_t, a_t) \coloneqq -\mathbb{E}_{X_{t+1}\sim\tau(\cdot|x_t,a_t)} \log \tau_\eta(X_{t+1}|x_t, a_t) \tag{1}$$

where τ_η denotes the agent's model of the environment parameterized by η, which is trained using the trajectories collected by rolling out a policy that seeks to maximize this same prediction error:

$$\underset{\pi}{\overset{\text{(policy)}}{\text{maximize}}} \quad \underset{\eta}{\overset{\text{(model)}}{\min}} \quad \mathbb{E}_{\substack{X_t\sim\rho_\pi \\ A_t\sim\pi(\cdot|X_t)}} \mathcal{R}_\eta(X_t, A_t) \tag{2}$$

Stochastic Traps In stochastic environments, this reward converges to the entropy $\mathbb{H}[X_{t+1}|x_t, a_t]$, so the agent may become stuck on repeatedly experiencing (intrinsically rewarding) transitions where entropy is high. What we desire is a reward that converges to zero in the limit. The notion of "optimistic" exploration offers a hint of what might be possible—Consider constructing a reward that satisfies:

$$\mathcal{R}_\eta(x_t, a_t) \geq D_{\text{KL}}\big(\tau(X_{t+1}|x_t, a_t)\|\tau_\eta(X_{t+1}|x_t, a_t)\big) \tag{3}$$

upper bounding the distance between the world and the agent's model. On the one hand, Definition 1 verifies this, but the bound fails to tighten even in the limit. On the other hand, it is hard to measure this distance directly, as the entropy term is by construction unknown. As it turns out, we shall later see that our proposed technique effectively gives a reward that verifies the inequality—and is tight in the limit.

Table 1: *Relationship with Curiosity-driven Exploration*. Curiosity in Hindsight is a drop-in modification applicable to any prediction error-based exploration bonus with a dynamics model (using any underlying representation space). Relative to any specific exploration method, Curiosity in Hindsight is uniquely characterized by the properties of being robust to all noise types, being dynamics aware, and being general to any representation space.

Curiosity-driven Exploration Method	*Prediction Inputs*	*Prediction Target*	*Measure of Learning*	Random Noise	X-/A-Dep. Noise	Dynamics Awareness	**Representation Space**
AE [10]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✗	✗	✓	reconstructive
ICM [11]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✓	✗	✓	action predictive
EMI [13]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✗	✗	✓	MI-maximizing
RND [12]	X_t	$f_{\text{random}}(X_t)$	$\mathcal{L}_\eta^{\text{predict}}$	✓	✓	✗	random projection
Dora [16]	X_t, A_t	const. zero	$\mathcal{L}_\eta^{\text{predict}}$	✓	✓	✗	pixel space
AMA [15]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}} - \text{Tr}(\hat{\Sigma}_{t+1})$	✓	✓	✓	pixel space
BYOL-Explore [14]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✗	✗	✓	bootstrapped
Curiosity in Hindsight + any representation	X_t, A_t, Z_{t+1}	X_{t+1}	$\mathcal{L}_{\theta,\eta}^{\text{reconstruct}} + \mathcal{L}_{\theta,\nu}^{\text{invariance}}$	✓	✓	✓	***any representation***

2.2 Related Work

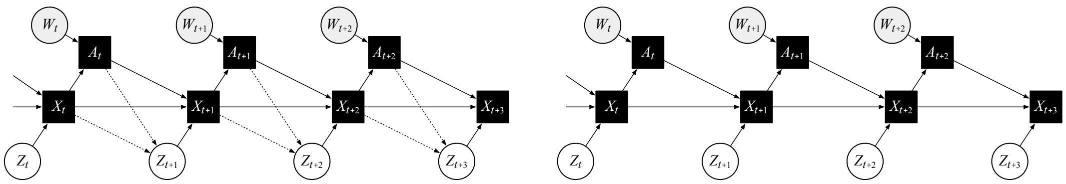
Our work inherits from the curiosity-driven paradigm [3–5, 9–15, 17, 18], among which some methods have been designed with robustness to certain stochasticities in mind (Table 1). However, note that our technique of using hindsight information is uniquely characterized by the following properties:

1. **Stochasticity Types**: First, it is capable of handling all types of stochasticities in generality. Specifically, this includes stochasticity that is *entirely random* (e.g. a viewport polluted by noise sampled according to a distribution independent of states and actions), stochasticity that is *state-dependent* (e.g. a visible object that performs a random walk within the environment), as well as *action-dependent* (e.g. a layer of random pixels that only appears if sampled on demand by specific actions). For instance, previous works have found that inverse dynamics features can learn to filter out random noise [11], but may break down in the presence of action-dependent noise [9, 19].
2. **Dynamics Awareness**: Second, it does not require discarding the curiosity-driven paradigm. By way of contrast, consider purely frequency-oriented exploration strategies, such as learning to predict a random projection of observations [12], or simply to predict the constant zero [16]. Since these are deterministic functions of inputs, they are in principle resilient to stochasticity. However, empirically they can still behave poorly in the presence of action-dependent stochasticities [15]: If the noise is sufficiently *diffuse*, the agent may never learn the function well, so in the absence of any other learning signal—such as the world's dynamics—they may still become stuck [20].
3. **Generality and Scalability**: As a drop-in modification, it is *generally applicable* to any underlying choice of representation space. In contrast, existing techniques capable of handling stochasticity are often tied to specific feature spaces, such as to employ inverse dynamics features [11], random features [12], or pixel-space features [15]—which may limit their flexibility of application. Moreover, unlike ensemble-based or disagreement-based techniques that require training a large number of models [19, 21, 22], we shall see that incorporating hindsight is simpler and more *scalable* by only requiring the addition of an auxiliary component to the usual prediction loss.

Alternative Paradigms Other paradigms have also been studied. Novelty-based methods encourage exploration on the basis of visitation counts [23], hashes [24], density estimates [25–28], and adversarial guidance [29, 30]; further extensions have accounted for the long-term value of exploratory actions [16, 31–33], as well as investigating the benefits of episodic memory [34–36]. Knowledge-based methods encourage exploration on the basis of the agent's uncertainty about the world [19, 37], with most work focusing on estimating the information gain from different actions [21, 22, 38–44], or directly estimating learning progress [45–47]. Finally, diversity-based methods seek to maximize the state entropy [48–51], or to encourage learning diverse skills [52–63] and reaching different goals [64–68].

3 Curiosity in Hindsight

Consider the game of betting on a hidden dice roll: Suppose we take the action $A_t =$ "bet on 6", then observe the outcome $X_{t+1} =$ "lost the bet". Two facts are clear: (1) *a priori*, we could not have predicted this result at all; (2) *a posteriori*, we may deduce the (latent) fact $Z_{t+1} =$ "the die must have rolled 1–5". These are not contradictory. In particular, the former does *not* imply that we lack an



(a) ***X*-/*A*-Dependent Noise** (b) **Reparameterized Model**

Figure 1: *Structural Causal Model.* Independent noise corresponds to latents Z_{t+1} with no incoming edges; state-dependent noise corresponds to directed edges from X_t to Z_{t+1}; and action-dependent noise corresponds to directed edges from A_t to Z_{t+1}. By the reparameterization lemma, there always exists an equivalent graphical representation under which all stochastic variables are effectively exogenous (i.e. with no directed edges into latents).

understanding of how the game works, nor does it suggest that we should engage in further such bets to improve our understanding. Indeed, knowing how the game works, in hindsight (i.e. given what we deduced about Z_{t+1}), the outcome is obvious to us (i.e. we can now deterministically identify X_{t+1}). Conversely, suppose we actually *didn't* know how the game works: Then we couldn't have correctly inferred Z_{t+1}, nor would its knowledge have enabled us to identify X_{t+1} with certainty. If so, engaging in additional bets may indeed allow us to learn and improve our understanding of how it works.

Intuitively, we can thus measure our understanding of each transition based on how much the outcome makes sense *in hindsight*. In other words, instead of asking "How well can we predict X_{t+1} *a priori*?", we actually want to ask "How well can we reconstruct X_{t+1} *a posteriori*—given hindsight Z_{t+1}?". In the sequel, we first formalize this intuition using the language of *posterior inference* when a known model of the world is available (Section 3.1). Subsequently, we generalize this approach to generating learned *hindsight representations* when a model of the world needs to be learned at the same time (Section 3.2). Finally, we derive *Curiosity in Hindsight* on the basis of these ingredients, showing that it approximates optimistic exploration (Inequality 3) while being robust to stochasticities (Section 3.3).

3.1 Structural Causal Model

Let Z denote a *latent* variable, taking on values $z \in \mathcal{Z}$. Specifically, for each observed transition tuple (x_t, a_t, x_{t+1}), we let z_{t+1} encapsulate *all* sources of unobserved stochasticity in the dynamics, such that—by construction—we have that $x_{t+1} = f(x_t, a_t, z_{t+1})$ for some deterministic function f. In this construction, a prior p over the latent Z_{t+1} induces the environment dynamics $\tau(X_{t+1}|x_t, a_t)$.

Figure 1(a) illustrates the structural causal model for this, where solid squares denote deterministic nodes, shaded circles denote observable stochastic nodes, and unshaded circles denote unobservable stochastic nodes (here we use W to capture any randomness in the agent's policy). Note that while stochasticities can be entirely random (i.e. no edges into Z_{t+1}), state-dependent (i.e. $X_t \rightarrow Z_{t+1}$), or action-dependent (i.e. $A_t \rightarrow Z_{t+1}$), by the *reparameterization lemma* it is always possible to represent an environment such that all stochastic variables are effectively exogenous [69, 70]—as in Figure 1(b).

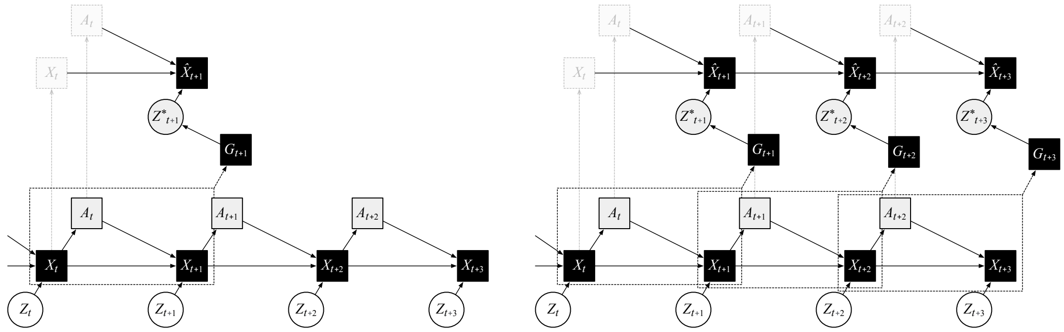
From Prediction to Reconstruction Consider the setting in which we know the model f. Suppose first that we somehow had access to each latent z_{t+1}. Then the outcome of a transition at state x_t and action a_t would be deterministically computable with no uncertainty (i.e. reconstruction error = zero):

$$x_{t+1} \coloneqq f(x_t, a_t, z_{t+1}) \tag{4}$$

In reality, of course, the latent variable z_{t+1} is not observable. Thus it may seem like the best we can accomplish is to compute the *a priori* expectation of the outcome (i.e. prediction error = entropy):

$$\bar{x}_{t+1} \coloneqq \mathbb{E}_{X_{t+1} \sim \tau(\cdot|x_t, a_t)} X_{t+1} = \mathbb{E}_{Z_{t+1} \sim p} f(x_t, a_t, Z_{t+1}) \tag{5}$$

However, while z_{t+1} is not observable, based on the transition (x_t, a_t, x_{t+1}) we can infer *a posteriori* what its values could have been. Importantly, by the consistency property of counterfactuals we know $f(x_t, a_t, Z_{t+1}) = x_{t+1}$ for any $Z_{t+1} \sim p(\cdot|x_t, a_t, x_{t+1})$ [71]. That is to say, conditioned on hindsight information, the reconstruction error of the true model is zero. This suggests that when f is unknown and learned by the agent, *the reconstruction error is an attractive candidate for an intrinsic reward*. Of course, now the missing piece is how to sample Z_{t+1} from the posterior—which we discuss next.



(a) **Single-Step Reconstruction** (b) **Multi-Step Reconstruction**

Figure 2: *Hindsight Representations*. For each transition, a learned generator $G_{t+1} := p_\theta(\cdot|x_t, a_t, x_{t+1})$ generates hindsight vectors in lieu of exact posterior inference. (In this figure, we use asterisks to distinguish them from unobserved "ground-truth" latents). For both (a) single-step and (b) multi-step horizons, hindsight vectors should be reconstructive of outcomes x_{t+1} when combined with x_t, a_t, as well as being independent of x_t, a_t.

3.2 Hindsight Representations

In the practical setting where the model $f(X_t, A_t, Z_{t+1})$ is unknown, we learn to approximate it using a *reconstructor* f_η, parameterized by η. Since exact posterior inference $p_\eta(Z_{t+1}|X_t, A_t, X_{t+1})$ is intractable, we simultaneously learn to approximate it using a *generator* p_θ, parameterized by θ. Two objectives are key. First, as noted above, representations Z_{t+1} should be *reconstructive* of outcomes X_{t+1}; here we use a squared loss, but in principle any kind of reconstruction loss can be selected:

Objective 1 (Reconstruction) Let the *reconstruction loss* for a given transition $(x_t, a_t, z_{t+1}, x_{t+1})$—including the generated hindsight representation z_{t+1} drawn from $p_\theta(\cdot|x_t, a_t, x_{t+1})$—be defined as:

$$\mathcal{L}_\eta(x_t, a_t, z_{t+1}, x_{t+1}) := \Big\| x_{t+1} - f_\eta(x_t, a_t, z_{t+1}) \Big\|_2^2 \tag{6}$$

and the (state-action) *reconstruction bonus* for the agent's policy:

$$\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x_t, a_t) := \mathbb{E}_{\substack{X_{t+1}\sim\tau(\cdot|x_t,a_t)\\ Z_{t+1}\sim p_\theta(\cdot|x_t,a_t,X_{t+1})}} \mathcal{L}_\eta(x_t, a_t, Z_{t+1}, X_{t+1}) \tag{7}$$

Driven to zero, this requires hindsight representations to encapsulate *at least* all aspects of the world's dynamics that are unpredictable (so we *don't* reward the agent for irreducible error). However, we also don't want Z_{t+1} to simply leak information about the outcome that is actually predictable to begin with (so we *do* reward the agent for reducible error). Thus our second objective requires it to be *independent* of X_t, A_t; here we use a contrastive loss, but in principle any kind of invariance loss can be selected:

Objective 2 (Invariance) Let the *invariance loss* for a tuple (x_t, a_t, z_{t+1}) be defined with respect to a batch of $K-1$ "negative" samples $Z^1_{t+1}, ..., Z^{K-1}_{t+1}$, using an auxiliary *critic* g_ν parameterized by ν:

$$\mathcal{L}^K_{\theta,\nu}(x_t, a_t, z_{t+1}) := \mathbb{E}_{\substack{(X^1_t,\ldots,X^{K-1}_t)\sim\prod_{i=1}^{K-1}\rho_\pi\\ (A^1_t,\ldots,A^{K-1}_t)\sim\prod_{i=1}^{K-1}\pi(\cdot|X^i_t)\\ (X^1_{t+1},\ldots,X^{K-1}_{t+1})\sim\prod_{i=1}^{K-1}\tau(\cdot|X^i_t,A^i_t)\\ (Z^1_{t+1},\ldots,Z^{K-1}_{t+1})\sim\prod_{i=1}^{K-1}p_\theta(\cdot|X^i_t,A^i_t,X^i_{t+1})}} \log \frac{e^{g_\nu(x_t,a_t,z_{t+1})}}{\frac{1}{K}\left(e^{g_\nu(x_t,a_t,z_{t+1})} + \sum_{i=1}^{K-1} e^{g_\nu(x_t,a_t,Z^i_{t+1})}\right)} \tag{8}$$

and the (state-action) *invariance bonus* for the agent's policy:

$$\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x_t, a_t) := \mathbb{E}_{Z_{t+1}\sim p_\theta(\cdot|x_t,a_t)} \mathcal{L}^K_{\theta,\nu}(x_t, a_t, Z_{t+1}) \tag{9}$$

Driven to minimax optimality over the state-action space—between the critic (i.e. maximizer) and generator (i.e. minimizer)—this requires hindsight representations to encapsulate *at most* the aspects of the world's dynamics that are unpredictable. How does it accomplish this? We can be more precise:

Proposition 1 (Optimal Invariance) Denote the pointwise mutual information between state-action x_t, a_t and hindsight z_t by $\mathrm{PMI}_\theta(x_t, a_t; z_{t+1}) := \log \frac{p_\theta(z_{t+1}|x_t,a_t)}{p_\theta(z_{t+1})}$. Then the invariance bonus satisfies:

$$\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x_t, a_t) \leq \mathbb{E}_{Z_{t+1}\sim p_\theta(\cdot|x_t,a_t)} \mathrm{PMI}_\theta(x_t, a_t; Z_{t+1}) \tag{10}$$

Moreover, denote the optimal critic parameter:

$$\nu^* := \arg\max_\nu \; \mathbb{E}_{\substack{X_t\sim\rho_\pi \\ A_t\sim\pi(\cdot|X_t) \\ X_{t+1}\sim\tau(\cdot|X_t,A_t) \\ Z_{t+1}\sim p_\theta(\cdot|X_t,A_t,X_{t+1})}} \mathcal{L}^K_{\theta,\nu}(X_t, A_t, Z_{t+1}) \tag{11}$$

Then the bound is asymptotically tight:

$$\lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(x_t, a_t) = \mathbb{E}_{Z_{t+1}\sim p_\theta(\cdot|x_t,a_t)} \mathrm{PMI}_\theta(x_t, a_t; Z_{t+1}) \tag{12}$$

Proof. Appendix A. □

In other words, in the limit of large batch sizes $K \to \infty$, at minimax optimality we have that Z_{t+1} is invariant to the values of x_t, a_t, and the invariance objective is equal to zero. One question remains: Can reconstruction be simultaneously driven to zero in the limit of infinite experience? The answer is yes:

Proposition 2 (Optimal Reconstruction) Denote with $\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x_t, a_t)$ the reconstruction bonus as in Objective 1, $\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x_t, a_t)$ the invariance bonus as in Objective 2, and their weighted sum for any λ:

$$J(\theta, \eta, \nu; \lambda) := \mathbb{E}_{\substack{X_t\sim\rho_\pi \\ A_t\sim\pi(\cdot|X_t)}} \left[\frac{1}{\lambda} \mathcal{R}^{\text{rec.}}_{\theta,\eta}(X_t, A_t) + \lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(X_t, A_t) \right] \tag{13}$$

Then its minimax optimal value is zero:

$$\min_{\theta,\eta} \max_\nu J(\theta, \eta, \nu; \lambda) = 0 \tag{14}$$

Proof. Appendix A. □

This suggests that such a weighted combination—of reconstruction loss (of a learned dynamics model) plus invariance loss (of a learned hindsight model)—may serve as a good intrinsic reward. We now have all the ingredients for Curiosity in Hindsight, which it is instructive to contrast with Definition 1:

3.3 Optimistic Exploration

Definition 2 (Curiosity in Hindsight) Denote the *hindsight intrinsic reward function* $\mathcal{R}_{\theta,\eta,\nu^*}$ with the following (for now, this is idealized in that the critic is assumed optimal and batch sizes are infinite):

$$\mathcal{R}_{\theta,\eta,\nu^*}(x_t, a_t) := \frac{1}{\lambda} \mathcal{R}^{\text{rec.}}_{\theta,\eta}(x_t, a_t) + \lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(x_t, a_t) \tag{15}$$

Similar to before, the agent maintains an internal dynamics model trained to minimize this quantity over the trajectories it collects, while rolling out a policy that seeks to maximize this same quantity:

$$\underset{\pi}{\overset{\text{(policy)}}{\text{maximize}}} \; \underset{\theta,\eta}{\overset{\text{(model)}}{\min}} \; \mathbb{E}_{\substack{X_t\sim\rho_\pi \\ A_t\sim\pi(\cdot|X_t)}} \mathcal{R}_{\theta,\eta,\nu^*}(X_t, A_t) \tag{16}$$

Recall that in the presence of stochasticity, standard curiosity-driven exploration can be seen as a poor approximation to "optimistic" exploration (Inequality 3)—because the bound is never tight even in the limit. The following observation shows that exploration using Curiosity in Hindsight can resolve this:

Theorem 3 (Optimistic Exploration) Let coefficient λ satisfy $\frac{1}{2}\log(\lambda\pi) \leq \mathbb{H}_\theta[X_{t+1}|x_t, a_t, Z_{t+1}] + D_{\text{KL}}\big(p_\theta(Z_{t+1}|x_t, a_t) \| p_\theta(Z_{t+1})\big)$, with π the mathematical constant (not the agent's policy). Then:

$$\mathcal{R}_{\theta,\eta,\nu^*}(x_t, a_t) \geq D_{\text{KL}}\big(\tau(X_{t+1}|x_t, a_t) \| \tau_{\theta,\eta}(X_{t+1}|x_t, a_t)\big) \tag{17}$$

where $\tau_{\theta,\eta}(X_{t+1}|x_t, a_t) := \mathbb{E}_{Z_{t+1}\sim p_\theta} p_\eta(X_{t+1}|x_t, a_t, Z_{t+1})$ denotes the learned environment model. Furthermore, for optimal model parameters θ^*, η^* we have that $\mathcal{R}_{\theta^*,\eta^*,\nu^*}(x_t, a_t) = 0$ for all x_t, a_t.

Proof. Appendix A. □

In other words, by choosing a small enough λ term, the hindsight intrinsic reward (Equation 15) is an upper bound on the KL-term we care about (Inequality 3). Since this intrinsic reward can be driven to zero in the limit (Proposition 2), so is the KL-term, thus the reward-maximizing exploration policy (Equation 16) is approximating precisely the sort of "optimistic" exploration that we desired.

4 Practical Framework

In practice, $K < \infty$, ν is not fully optimized, and λ is a hyperparameter. The intrinsic reward is now:

$$\mathcal{R}^{K}_{\theta,\eta,\nu}(x_t, a_t) \coloneqq \frac{1}{\lambda}\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x_t, a_t) + \mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x_t, a_t) \tag{18}$$

and the agent performs:

$$\underset{\pi}{\overset{\text{(policy)}}{\text{maximize}}} \; \underset{\theta,\eta}{\overset{\text{(model)}}{\min}} \max_{\nu} \; \mathbb{E}_{\substack{X_t \sim \rho_\pi \\ A_t \sim \pi(\cdot|X_t)}} \mathcal{R}^{K}_{\theta,\eta,\nu}(X_t, A_t) \tag{19}$$

Overall, our framework involves a simple drop-in modification on top of any standard curiosity-driven exploration: Instead of learning a *predictive model* that specifies $X_{t+1} \sim \tau_\eta(\cdot|X_t, A_t)$, we now learn a (hindsight-augmented) *reconstructive model* that specifies $X_{t+1} = f_\eta(X_t, A_t, Z_{t+1})$. The main ingredients include the reconstructor f_η, the generator $p_\theta(Z_{t+1}|X_t, A_t, X_{t+1})$, and the critic $g_\nu(X_t, A_t, Z_{t+1})$; the main hyperparameters are the contrastive batch size K and the coefficient λ in the intrinsic reward. Finally, note that while for simplicity we have focused our exposition on modeling single-step outcomes, generalizing to the case of multi-step horizons is straightforward (see Figure 2).

4.1 Example: BYOL-Hindsight

BYOL-Explore [14] is a recent technique for curiosity-driven exploration that learns a bootstrapped representation space, an environment dynamics model, as well as an exploration policy simultaneously by optimizing a prediction loss in latent space. First, an *online network* ω encodes observations o_t into representations $w_t = \omega(o_t)$. A *closed-loop* recurrent network then computes representations b_t of histories up until each time t. This is used to initialize an *open-loop* recurrent network that computes representations $b_{t,i}$ for horizon steps indexed as i. Finally, these representations are fed to a predictor ψ to output predictions $\hat{w}_{t,i}$, with the key novelty being that the *target network* $\tilde{\omega}$ is an EMA of the online network ω.

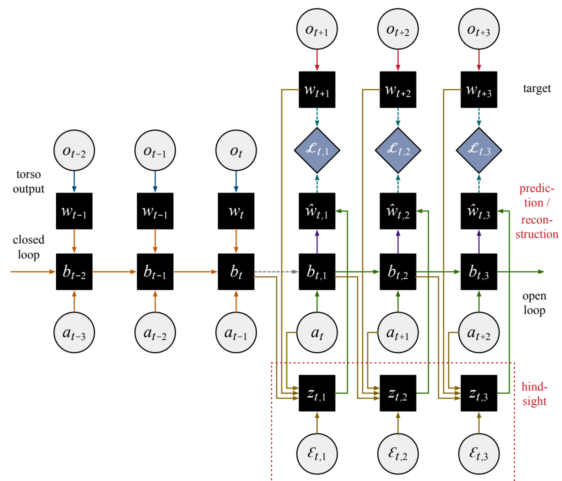


Figure 3: *BYOL-Explore to BYOL-Hindsight*.

Figure 3 shows the BYOL-Explore setup: The predictive error $\mathcal{L}_{t,i}$ at each open-loop step is computed, and the intrinsic reward associated to each transition o_s, a_s, o_{s+1} is the sum of its prediction errors $\sum_{t+i=s+1} \mathcal{L}_{t,i}$. This can be straightforwardly extended to use Curiosity in Hindsight (see dotted red region), yielding "BYOL-Hindsight": At each open-loop step, a hindsight vector is first sampled as $Z_{t,i} \sim p_\theta(\cdot|b_{t,i-1}, a_{t+i-1}, w_{t+i})$. Reconstructions $\hat{W}_{t,i} = f_\eta(b_{t,i-1}, a_{t+i-1}, Z_{t,i})$ are then computed, with the critic g_ν simultaneously encouraging $Z_{t,i}$ to be independent of $B_{t,i-1}, A_{t+i-1}$. Importantly, intrinsic rewards now come from $\mathcal{L}_{t,i}$ = reconstruction + invariance losses, and not prediction losses.

5 Experiments

So far, we proposed an intuitive framework for equipping curiosity-driven exploration with hindsight, and described an implementation on top of BYOL-Explore. Three questions deserve empirical investigation: **(a) Effectiveness**: In stochastic environments, predictive error-based methods—such as BYOL-Explore—may fail. Does BYOL-Hindsight circumvent the problem? **(b) Robustness**: Is BYOL-Hindsight robust to all of the types of stochasticities, including independent noise, state-dependent noise, and action-dependent noise? **(c) Nonspecificity**: In environments with no stochasticity, hindsight should confer no benefit. Does BYOL-Hindsight match the performance of BYOL-Explore?

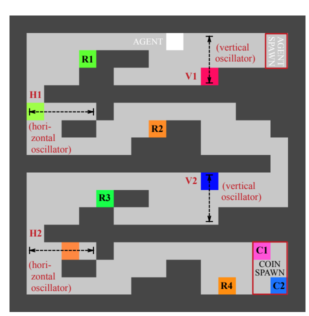
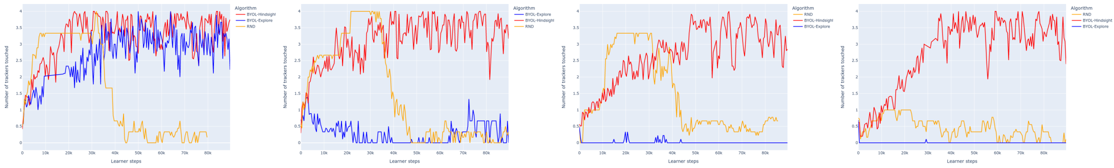


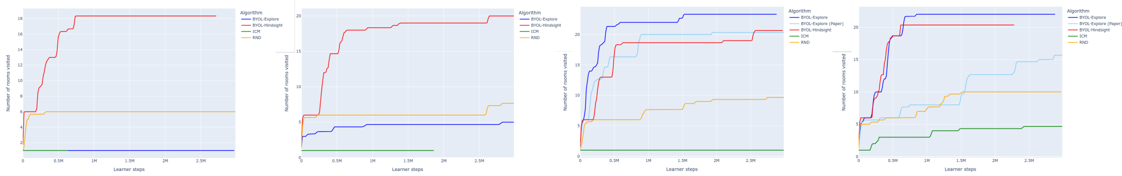
Figure 4: *Pycolab Maze Layout*.

Environments We employ two environments for experiments. First, we use a *Pycolab* [72] maze environment to experiment with different



(a) **Baseline: No Noise** (b) **Brownian Oscillators** (c) **Random Pixel Noise** (d) **On-Demand Pixel Noise**

Figure 5: *Pycolab Results*. BYOL-Hindsight manages to explore similarly to BYOL-Explore and RND in completing the full maze in the deterministic baseline, but is otherwise much more robust to all forms of stochasticities. Exploration performance is measured by the number of trackers touched during evaluation, in a 500-step episode.



(a) **Sticky (Intrinsic Only)** (b) **Sticky (Ext. + Intrinsic)** (c) **Non-Sticky (Intrinsic Only)** (d) **Non-Sticky (Ext. + Intrinsic)**

Figure 6: *Montezuma Results*. BYOL-Hindsight is extremely robust to the stochasticity due to sticky actions in both regimes, and manages to preserve most of the original performance of BYOL-Explore in the non-sticky baseline. Exploration performance is measured by the number of rooms reached during training, in the episodic setting.

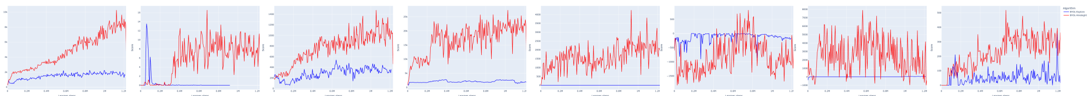
types of stochasticities in a controlled manner. Figure 4 shows the map: The agent spawns in the top right corner, and needs to explore its way past four (possibly stochastically oscillating) block elements (V1/2, H1/2), into the lower right corner where a pair of coins are randomly spawned. The agent is purely intrinsically motivated, and progress is measured by trackers located beyond each of the block elements (R1–4). The world is partially observable, and the agent only has access to a 5×5 frame (i.e. square radius 2) of its immediate surroundings as observations. Second, we use the popular RL benchmark of *Atari* games [73], with preprocessed grayscale 84×84-pixel images as observations. We consider some of the most commonly used hard-exploration games, including Montezuma's Revenge. Here we experiment with both pure-intrinsic and mixed (intrinsic plus extrinsic) exploration regimes.

Stochasticities In the *Pycolab* mazes, we use four different settings: "Baseline" (no noise), "Brownian Oscillators" (a form of state-dependent noise, where oscillators perform random walks along their axes of movements), "Random Pixel Noise" (a form of independent noise, which adds an extra layer of randomly sampled pixels with independent probability 0.25), and "On-Demand Pixel Noise" (a form of action-dependent noise, which does so if the no-op action is taken). In the *Atari* environments, we use "sticky actions" [74] with stickiness 0.1 as the source of noise (a form of action-dependent noise).

Implementation In all experiments, we use the exact same implementation for BYOL-Explore as given in [14] for Atari, including all hyperparameters such as target network EMA 0.99, open-loop horizon 8, intrinsic reward normalization and prioritization, weight sharing between exploration and RL closed-loop representations, and using VMPO [75] as the underlying algorithm. BYOL-Hindsight starts from the same setup but includes hindsight as given in Figure 3. There are two differences from the published version, which we use on both BYOL-Explore and BYOL-Hindsight for fair comparison: The predictor MLP uses three hidden layers of 512 instead of one of 256, and the mixing coefficient (in the mixed reward regime) is 0.2 instead of 0.1. We explicitly indicate the "paper" version in our results. Specifically for BYOL-Hindsight, the generator, reconstructor, and critic are all MLPs with three hidden layers of 512, the dimension of the generator noise ϵ and hindsight vector is 256, and $\lambda=1$. Finally, where shown for reference, RND and ICM are also implemented exactly as described in [14].

5.1 Pycolab Results

Figure 5 reports results (100k learner steps, averaged over 3 seeds). First, the "Baseline" setting tests nonspecificity: Since there is no noise except for the coins spawned at the end of the maze, we expect predictive error-based exploration to perform similarly with or without hindsight. For reference, we also show the performance of RND, which is in principle resilient to all types of stochasticity, because it explores by simply learning to predict the output of a deterministic function. All three algorithms manage to reach all four trackers (with RND eventually losing interest due to vanished rewards, since the environment is small). Second, in the "Brownian Oscillators" setting, BYOL-Explore fails to explore much beyond the first two trackers, since it simply hangs around and reaps the stream of intrinsic



(a) **Alien (I)** (b) **Freeway (I)** (c) **Gravitar (I)** (d) **Hero (I)** (e) **Montez. (I)** (f) **Pitfall (I)** (g) **Priv. Eye (I)** (h) **Venture (I)**

(i) **Alien (E+I)** (j) **Freeway (E+I)** (k) **Gravitar (E+I)** (l) **Hero (E+I)** (m) **Montez. (E+I)** (n) **Pitfall (E+I)** (o) **Priv. Eye (E+I)** (p) **Venture (E+I)**

Figure 7: *Atari Results (Sticky Actions)*. BYOL-Hindsight (red) improves on the performance of BYOL-Explore (blue) in the vast majority of scenarios. In line with the literature, as a proxy for "interesting behavior" in the environment, exploration performance is measured in terms of the extrinsic reward obtained during evaluation.

rewards from the unpredictable motion of the oscillators. In contrast, BYOL-Hindsight (and RND) both still manage to explore the entire maze. Third, in the "Random Pixel Noise" setting the results are similar, except both BYOL-Explore and RND perform even worse due to the fact that the noise is an entire layer of random pixels (i.e. extremely diffuse), which outcompetes all other dynamics of the world in magnitude. Interestingly, while BYOL-Hindsight requires ever so slightly longer to adapt, it manages to perform similarly to before. Even in the presence of high-magnitude, diffuse noise, the use of hindsight to capture the noise quickly allows it to stop bothering to predict it. Fourth, the "On-Demand Pixel Noise" setting is perhaps the most telling. BYOL-Explore is immediately trapped by the noise-inducing action, which it selects endlessly to generate a stream of intrinsic rewards. Differently to before, even RND suffers greatly, which makes sense because the agent is no longer guaranteed a 0.75 probability of observing the world's unpolluted dynamics. In contrast, BYOL-Hindsight still performs as nicely as in the noise-free setting, underscoring its robustness to all forms of stochasticity.

5.2 Atari Results

Figure 6 reports results for Montezuma's Revenge (3M learner steps, averaged over 3 seeds). We consider both intrinsic-only (using no extrinsic signal) and mixed (using extrinsic + intrinsic rewards) regimes. Exploration performance is measured by the number of different rooms of the dungeon the agent manages to discover over its lifetime—which requires understanding complex dynamics including navigating around various timed traps and moving enemies, and collecting keys to open doors in sequence. In the sticky actions setting, BYOL-Explore completely flatlines in the intrinsic-only regime, and only does marginally better in the mixed regime. In contrast, BYOL-Hindsight manages to explore most of the rooms in both regimes, verifying the fact that the learned hindsight representations are able to disentangle the (unpredictable) stickiness from the rest of the (predictable) dynamics of the world. Next, to test nonspecificity we run the same algorithms on the non-sticky setting: In both regimes, we observe that BYOL-Hindsight manages to preserve most of the original exploratory performance of BYOL-Explore. Finally, Figure 7 reports broad-based results for *Atari* hard-exploration games (1.2M learner steps, 1 seed), again for both intrinsic-only "(I)" and mixed "(E+I)" regimes. Following existing literature, we use the extrinsic reward obtained during evaluation as a proxy for an agent's ability to display "interesting behavior" in the environment. We observe that in the vast majority of cases BYOL-Hindsight improves on the performance of BYOL-Explore (especially when the latter flatlines). Overall, these results verify that Curiosity in Hindsight consistently bestows resilience to stickiness.

6 Conclusion

In this work, we studied the problem that stochasticity posed for predictive error-based exploration. Theoretically, we refined our notion of curiosity to separate (learnable) epistemic knowledge from (unlearnable) aleatoric variation during exploration. Algorithmically, we proposed a method for learning (future-summarizing) representations of hindsight disentangled from (history-summarizing) representations of context. Practically, we arrived at a simple and scalable framework for generating (reducible) intrinsic rewards even in the presence of (irreducible) stochastic traps—without having to estimate the problematic entropy term at all. Our perspective has tight connections with the study of counterfactuals in policy evaluation [69, 70], credit assignment [76, 77], and fairness [78, 79]. Future work may investigate the use of explicitly generative world models to map stochastic latents to outcomes, which may have potential use beyond generating intrinsic rewards—in the RL algorithm itself.

References

[1] Nikolaus Kriegeskorte and Pamela K Douglas. Cognitive computational neuroscience. *Nature neuroscience*, 21(9):1148–1160, 2018.

[2] Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Jianye Hao, Zhaopeng Meng, and Peng Liu. Exploration in deep reinforcement learning: a comprehensive survey. *arXiv preprint arXiv:2109.06668*, 2021.

[3] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

[4] Sebastian Thrun. Exploration in active learning. *Handbook of Brain Science and Neural Networks*, pages 381–384, 1995.

[5] Andrew G Barto, Satinder Singh, Nuttapong Chentanez, et al. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning*, pages 112–19. Piscataway, NJ, 2004.

[6] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. *Advances in neural information processing systems*, 28, 2015.

[7] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016.

[8] Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aaron van den Oord. Shaping belief states with generative environment models for rl. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *International Conference on Learning Representations*, 2019.

[10] Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *International Conference on Learning Representations*, 2016.

[11] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[12] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *International Conference on Learning Representations*, 2019.

[13] Hyoungseok 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. PMLR, 2019.

[14] Zhaohan Daniel Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems*, 35, 2022.

[15] Augustine Mavor-Parker, Kimberly Young, Caswell Barry, and Lewis Griffin. How to stay curious while avoiding noisy tvs using aleatoric uncertainty estimation. In *International Conference on Machine Learning*, pages 15220–15240. PMLR, 2022.

[16] Leshem Choshen, Lior Fox, and Yonatan Loewenstein. Dora the explorer: Directed outreaching reinforcement action-selection. *International Conference on Learning Representations*, 2018.

[17] Laurent Orseau, Tor Lattimore, and Marcus Hutter. Universal knowledge-seeking agents for stochastic environments. In *International conference on algorithmic learning theory*, pages 158–172. Springer, 2013.

[18] Zhang-Wei Hong, Tsu-Jui Fu, Tzu-Yun Shann, and Chun-Yi Lee. Adversarial active exploration for inverse dynamics model learning. In *Conference on Robot Learning*, pages 552–565. PMLR, 2020.

[19] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pages 5062–5071. PMLR, 2019.

[20] Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity. In *International conference on machine learning*, pages 5306–5315. PMLR, 2020.

[21] Mikael Henaff. Explicit explore-exploit algorithms in continuous state spaces. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *International conference on machine learning*, pages 5779–5788. PMLR, 2019.

[23] Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74, 2008.

[24] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[25] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

[26] Georg Ostrovski, Marc G Bellemare, Aäron Oord, and Rémi Munos. Count-based exploration with neural density models. In *International conference on machine learning*, pages 2721–2730. PMLR, 2017.

[27] Rui Zhao and Volker Tresp. Curiosity-driven experience prioritization via density estimation. *Advances in neural information processing systems*, 31, 2018.

[28] Omar Darwiche Domingues, Corentin Tallec, Remi Munos, and Michal Valko. Density-based bonuses on learned representations for reward-free exploration in deep reinforcement learning. In *ICML 2021 Workshop on Unsupervised Reinforcement Learning*, 2021.

[29] Justin Fu, John Co-Reyes, and Sergey Levine. Ex2: Exploration with exemplar models for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[30] Yannis Flet-Berliac, Johan Ferret, Olivier Pietquin, Philippe Preux, and Matthieu Geist. Adversarially guided actor-critic. *International Conference on Learning Representations*, 2021.

[31] Marlos C Machado, Marc G Bellemare, and Michael Bowling. Count-based exploration with the successor representation. In *ICML Workshop on Exploration in Reinforcement Learning*, 2018.

[32] Min-hwan Oh and Garud Iyengar. Directed exploration in pac model-free reinforcement learning. In *ICML Workshop on Exploration in Reinforcement Learning*, 2018.

[33] Marlos C Machado, Marc G Bellemare, and Michael Bowling. Count-based exploration with the successor representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[34] Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. *International Conference on Learning Representations*, 2019.

[35] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt, et al. Never give up: Learning directed exploration strategies. *International Conference on Learning Representations*, 2020.

[36] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *International Conference on Machine Learning*, pages 507–517. PMLR, 2020.

[37] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[38] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[39] Mauricio Araya, Olivier Buffet, Vincent Thomas, and Françcois Charpillet. A pomdp extension with belief-dependent rewards. *Advances in neural information processing systems*, 23, 2010.

[40] Yi Sun, Faustino Gomez, and Jürgen Schmidhuber. Planning to be surprised: Optimal bayesian exploration in dynamic environments. In *International conference on artificial general intelligence*, pages 41–51. Springer, 2011.

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

[42] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

[43] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pages 8583–8592. PMLR, 2020.

[44] Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and achieving goals via world models. *Advances in Neural Information Processing Systems*, 34:24379–24391, 2021.

[45] Jürgen Schmidhuber. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pages 1458–1463, 1991.

[46] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286, 2007.

[47] Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo Avila Pires, Jean-Bastien Grill, Florent Altché, and Rémi Munos. World discovery models. *arXiv preprint arXiv:1902.07685*, 2019.

[48] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

[49] Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. *Advances in Neural Information Processing Systems*, 34:18459–18473, 2021.

[50] Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Alaa Saade, Shantanu Thakoor, Bilal Piot, Bernardo Avila Pires, Michal Valko, Thomas Mesnard, Tor Lattimore, and Rémi Munos. Geometric entropic exploration. *arXiv preprint arXiv:2101.02055*, 2021.

[51] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pages 11920–11931. PMLR, 2021.

[52] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *International Conference on Learning Representations*, 2017.

[53] Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299*, 2018.

[54] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *International Conference on Learning Representations*, 2019.

[55] Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

[56] Víctor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giró-i Nieto, and Jordi Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. In *International Conference on Machine Learning*, pages 1317–1327. PMLR, 2020.

[57] Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *International Conference on Learning Representations*, 2020.

[58] Kate Baumli, David Warde-Farley, Steven Hansen, and Volodymyr Mnih. Relative variational intrinsic control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6732–6740, 2021.

[59] Oliver Groth, Markus Wulfmeier, Giulia Vezzani, Vibhavari Dasagi, Tim Hertweck, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Is curiosity all you need? on the utility of emergent behaviours from curious exploration. *arXiv preprint arXiv:2109.08603*, 2021.

[60] Taehwan Kwon. Variational intrinsic control revisited. *International Conference on Learning Representations*, 2022.

[61] Hao Liu and Pieter Abbeel. Aps: Active pretraining with successor features. In *International Conference on Machine Learning*, pages 6736–6747. PMLR, 2021.

[62] Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. The information geometry of unsupervised reinforcement learning. *International Conference on Learning Representations*, 2022.

[63] Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Cic: Contrastive intrinsic control for unsupervised skill discovery. *arXiv preprint arXiv:2202.00161*, 2022.

[64] 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. *Advances in neural information processing systems*, 30, 2017.

[65] Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pages 1515–1528. PMLR, 2018.

[66] Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *Advances in neural information processing systems*, 31, 2018.

[67] Meng Fang, Tianyi Zhou, Yali Du, Lei Han, and Zhengyou Zhang. Curriculum-guided hindsight experience replay. *Advances in neural information processing systems*, 32, 2019.

[68] Yunzhi Zhang, Pieter Abbeel, and Lerrel Pinto. Automatic curriculum learning through value disagreement. *Advances in Neural Information Processing Systems*, 33:7648–7659, 2020.

[69] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

[70] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

[71] Judea Pearl. *Causality*. Cambridge university press, 2009.

[72] Thomas Stepleton. The pycolab game engine, 2017. *URL https://github. com/deepmind/pycolab*, 2017.

[73] 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.

[74] Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

[75] H Francis Song, Abbas Abdolmaleki, Jost Tobias Springenberg, Aidan Clark, Hubert Soyer, Jack W Rae, Seb Noury, Arun Ahuja, Siqi Liu, Dhruva Tirumala, et al. V-mpo: On-policy maximum a posteriori policy optimization for discrete and continuous control. *arXiv preprint arXiv:1909.12238*, 2019.

[76] Thomas Mesnard, Théophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Tom Stepleton, Nicolas Heess, Arthur Guez, et al. Counterfactual credit assignment in model-free reinforcement learning. In *International Conference on Machine Learning*, 2021.

[77] Chris Nota, Philip Thomas, and Bruno C Da Silva. Posterior value functions: Hindsight baselines for policy gradient methods. In *International Conference on Machine Learning*, pages 8238–8247. PMLR, 2021.

[78] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *International Conference on Learning Representations*, 2016.

[79] Adam Foster, Árpi Vezér, Craig A Glastonbury, Páidí Creed, Samer Abujudeh, and Aaron Sim. Contrastive mixture of posteriors for counterfactual inference, data integration and fairness. In *International Conference on Machine Learning*, pages 6578–6621. PMLR, 2022.

A Proofs of Propositions

To simplify our notation, we remove subscripts such that X, A, Y denotes the transition X_t, A_t, X_{t+1}, and Z denotes the latent Z_{t+1}. Then the environment's dynamics is given by $\tau(Y|x, a)$, the agent's policy is given by $\pi(A|x)$, and the induced state visitation given by $\rho_\pi(X)$. The generator is denoted $p_\theta(Z|x, a, y)$, the reconstructor $f_\eta(x, a, z)$, and the critic $g_\nu(x, a, z)$. We start with several lemmas that will be useful, the first being a pointwise version of Barber and Agakov's variational lower bound:

Lemma 4 (Pointwise Barber-Agakov) Denote the pointwise mutual information:

$$\mathrm{PMI}_\theta(x, a; z) := \log \frac{p_\theta(z|x, a)}{p_\theta(z)} \tag{20}$$

Then for any variational distribution q:

$$\mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathrm{PMI}_\theta(x, a; Z) \ \geq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{q(Z|x, a)}{p_\theta(Z)} \tag{21}$$

Proof. Starting from the left hand side:

$$\begin{aligned}
\mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathrm{PMI}_\theta(x, a; Z) \ &= \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{p_\theta(Z|x, a)}{p_\theta(Z)} &(22)\\
&= \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{p_\theta(Z|x, a)}{p_\theta(Z)} + \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{q(Z|x, a)}{q(Z|x, a)} &(23)\\
&= \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{q(Z|x, a)}{p_\theta(Z)} + D_{\mathrm{KL}}\big(p_\theta(Z|x, a)\|q(Z|x, a)\big) &(24)\\
&\geq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{q(Z|x, a)}{p_\theta(Z)} &(25)
\end{aligned}$$

which completes the proof. □

Next, we define a generic contrastive expression with $K-1$ "negative" samples of Z, and show that taking its expectation with respect to those samples yields a valid (i.e. normalized) probability density:

Lemma 5 (Normalized Variational) Given independent samples $z_{1:K-1}$ from p_θ, define:

$$q(z|x, a, z_{1:K-1}) := \frac{p_\theta(z) \cdot e^{g_\nu(x,a,z)}}{\frac{1}{K}\Big(e^{g_\nu(x,a,z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,z_i)}\Big)} \tag{26}$$

then the following defines a normalized density:

$$q(Z|x, a) := \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} q(Z|x, a, Z_{1:K-1}) \tag{27}$$

Proof. The expectation integrates to one:

$$\begin{aligned}
\int_{\mathcal{Z}} q(z|x, a)dz \ &= \int_{\mathcal{Z}} \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \frac{p_\theta(z) \cdot e^{g_\nu(x,a,z)}}{\frac{1}{K}\Big(e^{g_\nu(x,a,z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\Big)} dz &(28)\\
&= \mathbb{E}_{\substack{Z\sim p_\theta\\ Z_{1:K-1}\sim p_\theta^{K-1}}} \frac{e^{g_\nu(x,a,Z)}}{\frac{1}{K}\Big(e^{g_\nu(x,a,Z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\Big)} &(29)\\
&= K \cdot \mathbb{E}_{Z_{1:K}\sim p_\theta^K} \frac{e^{g_\nu(x,a,Z_1)}}{\sum_{i=1}^{K} e^{g_\nu(x,a,Z_i)}} &(30)\\
&= \mathbb{E}_{Z_{1:K}\sim p_\theta^K} \frac{\sum_{j=1}^{K} e^{g_\nu(x,a,Z_j)}}{\sum_{i=1}^{K} e^{g_\nu(x,a,Z_i)}} = 1 &(31)
\end{aligned}$$

which completes the proof. □

These two results allow us to show that the information Z contains on a tuple x, a—with respect to the generator parameterized as θ—is lower-bounded by the x, a-conditioned contrastive loss between "positive" samples $Z \sim p_\theta(\cdot|x,a)$ from the posterior and "negative" samples $Z \sim p_\theta$ from the prior:

Lemma 6 (State-Action Lower Bound) The x, a-wise mutual information satisfies:

$$\begin{aligned}
&\mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathrm{PMI}_\theta(x,a;Z) \\
&\qquad \geq \mathbb{E}_{\substack{Z\sim p_\theta(\cdot|x,a)\\ Z_{1:K-1}\sim p_\theta^{K-1}}} \log \frac{e^{g_\nu(x,a,Z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,Z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\right)}
\end{aligned} \tag{32}$$

Proof. Use Lemmas 4 and 5, then Jensen's inequality:

$$\begin{aligned}
&\mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathrm{PMI}_\theta(x,a;Z) \\
&\qquad \geq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \frac{q(Z|x,a)}{p_\theta(Z)} && (33) \\
&\qquad = \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \log \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \frac{q(Z|x,a,Z_{1:K-1})}{p_\theta(Z)} && (34) \\
&\qquad \geq \mathbb{E}_{\substack{Z\sim p_\theta(\cdot|x,a)\\ Z_{1:K-1}\sim p_\theta^{K-1}}} \log \frac{q(Z|x,a,Z_{1:K-1})}{p_\theta(Z)} && (35) \\
&\qquad = \mathbb{E}_{\substack{Z\sim p_\theta(\cdot|x,a)\\ Z_{1:K-1}\sim p_\theta^{K-1}}} \log \frac{e^{g_\nu(x,a,Z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,Z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\right)} && (36)
\end{aligned}$$

which completes the proof. □

Next, we show that our invariance loss (Objective 2) for a tuple x, a, z is equal to the pointwise mutual information in the limit of infinitely large negative batches, assuming an optimal critic parameter:

Lemma 7 (Pointwise Asymptotic Equality) Define the pointwise invariance loss:

$$\mathcal{L}_{\theta,\nu}^K(x,a,z) \coloneqq \mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \log \frac{e^{g_\nu(x,a,z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\right)} \tag{37}$$

and the optimal critic parameter:

$$\nu^* \coloneqq \operatorname*{arg\,max}_\nu \mathbb{E}_{\substack{X\sim\rho_\pi\\ A\sim\pi(\cdot|X)\\ Y\sim\tau(\cdot|X,A)\\ Z\sim p_\theta(\cdot|X,A,Y)}} \mathcal{L}_{\theta,\nu}^K(X,A,Z) \tag{38}$$

Then $\lim_{K\to\infty}\mathcal{L}_{\theta,\nu^*}^K(x,a,z) = \mathrm{PMI}_\theta(x,a;z)$.

Proof. The $\mathbb{E}[\mathcal{L}_{\theta,\nu}^K(X,A,Z)]$ term is just the InfoNCE loss between variables Z and X, A, so we know that ν^* satisfies $g_{\nu^*}(x,a,z) = \log\frac{p_\theta(z|x,a)}{p_\theta(z)} + c(x,a)$. Substituting this back into $\mathcal{L}_{\theta,\nu}^K(x,a,z)$:

$$\begin{aligned}
&\lim_{K\to\infty}\mathcal{L}_{\theta,\nu^*}^K(x,a,z) && (39) \\
&\quad = \lim_{K\to\infty}\mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \log \frac{e^{g_{\nu^*}(x,a,z)}}{\frac{1}{K}\left(e^{g_{\nu^*}(x,a,z)} + \sum_{i=1}^{K-1} e^{g_{\nu^*}(x,a,Z_i)}\right)} && (40) \\
&\quad = \lim_{K\to\infty}\mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \log \frac{\frac{p_\theta(z|x,a)}{p_\theta(z)}}{\frac{1}{K}\left(\frac{p_\theta(z|x,a)}{p_\theta(z)} + \sum_{i=1}^{K-1}\frac{p_\theta(Z_i|x,a)}{p_\theta(Z_i)}\right)} && (41) \\
&\quad = \lim_{K\to\infty}\mathbb{E}_{Z_{1:K-1}\sim p_\theta^{K-1}} \left[\log\frac{p_\theta(z|x,a)}{p_\theta(z)} - \log\frac{\frac{p_\theta(z|x,a)}{p_\theta(z)} + \sum_{i=1}^{K-1}\frac{p_\theta(Z_i|x,a)}{p_\theta(Z_i)}}{K}\right] && (42)
\end{aligned}$$

$$= \log \frac{p_\theta(z|x,a)}{p_\theta(z)} - \lim_{K\to\infty} \log \frac{\frac{p_\theta(z|x,a)}{p_\theta(z)} + K - 1}{K} = \text{PMI}_\theta(x,a;z) \tag{43}$$

which completes the proof. □

This gives us what we need to derive Proposition 1, which we restate using our subscript-less notation:

Proposition 8 (Optimal Invariance) The (state-action) invariance bonus satisfies:

$$\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x,a) \leq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \text{PMI}_\theta(x,a;Z) \tag{44}$$

and for the optimal ν^* the bound is asymptotically tight as $K \to \infty$:

$$\lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(x,a) = \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \text{PMI}_\theta(x,a;Z) \tag{45}$$

Proof. Use Lemma 6 for the first part:

$$\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x,a) \coloneqq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \mathcal{L}^K_{\theta,\nu}(x,a,Z) \tag{46}$$

$$= \mathbb{E}_{\substack{Z\sim p_\theta(\cdot|x,a) \\ Z_{1:K-1}\sim p_\theta^{K-1}}} \log \frac{e^{g_\nu(x,a,Z)}}{\frac{1}{K}\left(e^{g_\nu(x,a,Z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\right)} \tag{47}$$

$$\leq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \text{PMI}_\theta(x,a;Z) \tag{48}$$

and use Lemma 7 for the second part:

$$\lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(x,a) = \lim_{K\to\infty} \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \mathcal{L}^K_{\theta,\nu^*}(x,a,Z) \tag{49}$$

$$= \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \lim_{K\to\infty} \mathcal{L}^K_{\theta,\nu^*}(x,a,Z) \tag{50}$$

$$= \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)} \text{PMI}_\theta(x,a;Z) \tag{51}$$

which completes the proof. □

Next, we show that Proposition 2 is true, which we similarly restate using our subscript-less notation:

Proposition 9 (Optimal Reconstruction) Denote with $\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x,a)$ the reconstruction bonus as in Objective 1, $\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x,a)$ the invariance bonus as in Objective 2, and their weighted sum for any λ:

$$J(\theta,\eta,\nu;\lambda) \coloneqq \mathbb{E}_{\substack{X\sim\rho_\pi \\ A\sim\pi(\cdot|X)}} \left[\frac{1}{\lambda}\mathcal{R}^{\text{rec.}}_{\theta,\eta}(X,A) + \lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(X,A)\right] \tag{52}$$

Then its minimax optimal value is zero:

$$\min_{\theta,\eta} \max_{\nu} J(\theta,\eta,\nu;\lambda) = 0 \tag{53}$$

Proof. Take any MDP, as in Figure 1(a). By reparameterization, we know there exists an equivalent graphical representation under which Z is exogenous, as in Figure 1(b). Assuming realizability, let η^* be such that $f_{\eta^*} = f$, and let θ^* be such that $p_{\theta^*}(Z|x,a,y) = p_{\eta^*}(Z|x,a,y)$ for any x,a,y. First, by construction we have that $Z \perp X, A$, so the mutual information between Z and X, A must be zero:

$$\mathbb{E}_{\substack{X\sim\rho_\pi \\ A\sim\pi(\cdot|X)}} \left[\lim_{K\to\infty} \mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(X,A)\right] = \mathbb{E}_{\substack{X\sim\rho_\pi \\ A\sim\pi(\cdot|X) \\ Z\sim p_\theta(\cdot|X,A)}} \text{PMI}_\theta(X,A;Z) \tag{54}$$

$$= \mathbb{I}_\theta[X,A;Z] = 0 \tag{55}$$

for optimal critic parameter ν^*, where the first equality uses Proposition 1. Second, by consistency of counterfactuals $f_{\eta^*}(x,a,Z) = y$ for any $Z \sim p_{\theta^*}(\cdot|x,a,y)$, so the reconstruction term is also zero. It is easy to verify the optimal critic is a maximizer, and the optimal generator/reconstructor minimizers, which completes the proof. □

Finally, we recall the following basic relationship:

Lemma 10 (Conditional Mutual Information) Conditioned on any x, a, we have that:

$$\mathbb{I}_\theta[Y;Z|x,a] = \mathbb{H}[Y|x,a] + \mathbb{H}_\theta[Y|x,a,Z] \tag{56}$$

Proof. Starting from the left hand side:

$$\mathbb{I}_\theta[Y;Z|x,a] \coloneqq \mathbb{E}_{Z\sim p_\theta} D_{\mathrm{KL}}\big(p_\theta(Y|x,a,Z)\|\tau(Y|x,a)\big) \tag{57}$$

$$= \mathbb{E}_{\substack{Z\sim p_\theta \\ Y\sim p_\theta(\cdot|x,a,Z)}} \log p_\theta(Y|x,a,Z) - \mathbb{E}_{\substack{Z\sim p_\theta \\ Y\sim p_\theta(\cdot|x,a,Z)}} \tau(Y|x,a) \tag{58}$$

$$= -\textstyle\int_{\mathcal{Z}} p_\theta(z)\mathbb{H}_\theta[Y|x,a,z]dz - \mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a) \\ Z\sim p_\theta(\cdot|x,a,Y)}} \tau(Y|x,a) \tag{59}$$

$$= \mathbb{H}[Y|x,a] - \mathbb{H}_\theta[Y|x,a,Z] \tag{60}$$

which completes the proof. □

Now, in our structural causal model, by construction Z captures all sources of noise—that is, there is no residual noise in each outcome Y. However, for the purposes of optimization, while learning η we let the residual error be captured by a Gaussian "log-likelihood" (note that λ plays the role of "$2\sigma^2$"):

$$\log p_\eta(Y|x,a,z) \coloneqq -\frac{1}{2}\log(\lambda\pi) - \frac{1}{\lambda}\big(Y - f_\eta(x,a,z)\big)^2 \tag{61}$$

and note that θ also induces a log-likelihood of the "ground-truth" conditional:

$$\log p_\theta(Y|x,a,z) \coloneqq \frac{p_\theta(z|x,a,Y)\tau(Y|x,a)\pi(a,x)\rho_\pi(x)}{\int_{\mathcal{Y}} p_\theta(z|x,a,y)\tau(y|x,a)\pi(a|x)\rho_\pi(x)dy} \tag{62}$$

Now, recall the reconstruction loss and (state-action) reconstruction bonus:

$$\mathcal{L}_\eta(x,a,z,y) \coloneqq \Big\|y - f_\eta(x,a,z)\Big\|_2^2 \tag{63}$$

$$\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x,a) \coloneqq \mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a) \\ Z\sim p_\theta(\cdot|x,a,Y)}} \mathcal{L}_\eta(x,a,Z,Y) \tag{64}$$

as well as the invariance loss and (state-action) invariance bonus:

$$\mathcal{L}^K_{\theta,\nu}(x,a,z) \coloneqq \mathbb{E}_{\substack{(X_1,\dots,X_{K-1})\sim\prod_{i=1}^{K-1}\rho_\pi \\ (A_1,\dots,A_{K-1})\sim\prod_{i=1}^{K-1}\pi(\cdot|X_i) \\ (Y_1,\dots,Y_{K-1})\sim\prod_{i=1}^{K-1}\tau(\cdot|X_i,A_i) \\ (Z_1,\dots,Z_{K-1})\sim\prod_{i=1}^{K-1}p_\theta(\cdot|X_i,A_i,Y_i)}} \log \frac{e^{g_\nu(x,a,z)}}{\frac{1}{K}\Big(e^{g_\nu(x,a,z)} + \sum_{i=1}^{K-1} e^{g_\nu(x,a,Z_i)}\Big)} \tag{65}$$

$$\mathcal{R}^{K,\text{inv.}}_{\theta,\nu}(x,a) \coloneqq \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathcal{L}^K_{\theta,\nu}(x,a,Z) \tag{66}$$

Moreover, recall the hindsight intrinsic reward function:

$$\mathcal{R}_{\theta,\eta,\nu^*}(x,a) \coloneqq \frac{1}{\lambda}\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x,a) + \lim_{K\to\infty}\mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(x,a) \tag{67}$$

We can now show that Theorem 3 is true, which we similarly restate using our subscript-less notation:

Theorem 11 (Optimistic Exploration) Let λ satisfy the inequality $\frac{1}{2}\log(\lambda\pi) \leq \mathbb{H}_\theta[Y|x,a,Z] + D_{\mathrm{KL}}\big(p_\theta(Z|x,a)\|p_\theta(Z)\big)$, with π here being the mathematical constant (not the agent's policy). Then:

$$\mathcal{R}_{\theta,\eta,\nu^*}(x,a) \geq D_{\mathrm{KL}}\big(\tau(Y|x,a)\|\tau_{\theta,\eta}(Y|x,a)\big) \tag{68}$$

where $\tau_{\theta,\eta}(Y|x,a) \coloneqq \mathbb{E}_{Z\sim p_\theta} p_\eta(Y|x,a,Z)$ denotes the learned environment model. Furthermore, for optimal model parameters θ^*, η^* we have that the intrinsic reward $\mathcal{R}_{\theta^*,\eta^*,\nu^*}(x,a) = 0$ for all x, a.

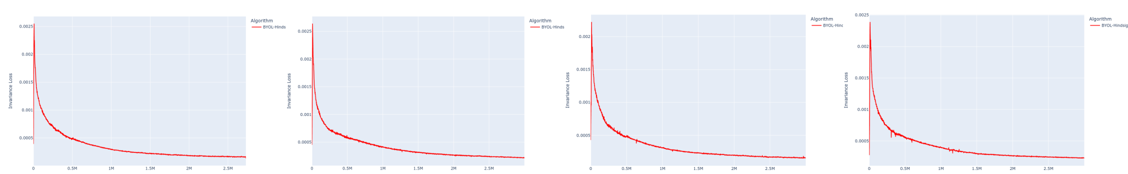
Proof. Use Proposition 8, then the constraint on λ, then Lemma 10:

$$
\begin{aligned}
\mathcal{R}_{\theta,\eta,\nu^*}(x,a) &\coloneqq \tfrac{1}{\lambda}\mathcal{R}^{\text{rec.}}_{\theta,\eta}(x,a) + \lim_{K\to\infty}\mathcal{R}^{K,\text{inv.}}_{\theta,\nu^*}(x,a) && (69)\\
&= \mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a)\\ Z\sim p_\theta(\cdot|x,a,Y)}} \tfrac{1}{\lambda}\big(Y - f_\eta(x,a,Z)\big)^2 + \mathbb{E}_{Z\sim p_\theta(\cdot|x,a)}\mathrm{PMI}_\theta(x,a;Z) && (70)\\
&= \mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a)\\ Z\sim p_\theta(\cdot|x,a,Y)}} \tfrac{1}{\lambda}\big(Y - f_\eta(x,a,Z)\big)^2 + D_{\mathrm{KL}}\big(p_\theta(Z|x,a)\|p_\theta(Z)\big) && (71)\\
&\geq -\mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a)\\ Z\sim p_\theta(\cdot|x,a,Y)}} \log p_\eta(Y|x,a,Z) - \mathbb{H}_\theta[Y|x,a,Z] && (72)\\
&= -\mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a)\\ Z\sim p_\theta(\cdot|x,a,Y)}} \log p_\eta(Y|x,a,Z) + \mathbb{I}_\theta[Y;Z|x,a] - \mathbb{H}[Y|x,a] && (73)\\
&= -\mathbb{E}_{\substack{Y\sim\tau(\cdot|x,a)\\ Z\sim p_\theta(\cdot|x,a,Y)}} \log p_\eta(Y|x,a,Z) \leftarrow \textbf{remaining}\text{ stochasticity}\\
&\quad + \mathbb{E}_{Y\sim\tau(\cdot|x,a)} D_{\mathrm{KL}}\big(p_\theta(Z|x,a,Y)\|p_\theta(Z|x,a)\big) \leftarrow \textbf{hindsight}\text{ information}\\
&\quad - \mathbb{E}_{Y\sim\tau(\cdot|x,a)}\big[-\log\tau(Y|x,a)\big] \leftarrow \textbf{total}\text{ stochasticity} && (74)\\
&\geq -\mathbb{E}_{Y\sim\tau(\cdot|x,a)}\big[\mathbb{E}_{Z\sim p_\theta(\cdot|x,a,Y)}\log p_\eta(Y|x,a,Z)\\
&\quad - D_{\mathrm{KL}}\big(p_\theta(Z|x,a,Y)\|p_\theta(Z|x,a)\big) + D_{\mathrm{KL}}\big(p_\theta(Z|x,a,Y)\|p_\eta(Z|x,a,Y)\big)\big]\\
&\quad + \mathbb{E}_{Y\sim\tau(\cdot|x,a)}\log\tau(Y|x,a) && (75)\\
&= -\mathbb{E}_{Y\sim\tau(\cdot|x,a)}\log\mathbb{E}_{Z\sim p_\theta}p_\eta(Y|x,a,Z) + \mathbb{E}_{Y\sim\tau(\cdot|x,a)}\log\tau(Y|x,a) && (76)\\
&= -\mathbb{E}_{Y\sim\tau(\cdot|x,a)}\log\tau_{\theta,\eta}(Y|x,a) + \mathbb{E}_{Y\sim\tau(\cdot|x,a)}\log\tau(Y|x,a) && (77)\\
&= D_{\mathrm{KL}}\big(\tau(Y|x,a)\|\tau_{\theta,\eta}(Y|x,a)\big) && (78)
\end{aligned}
$$

which completes the proof. □

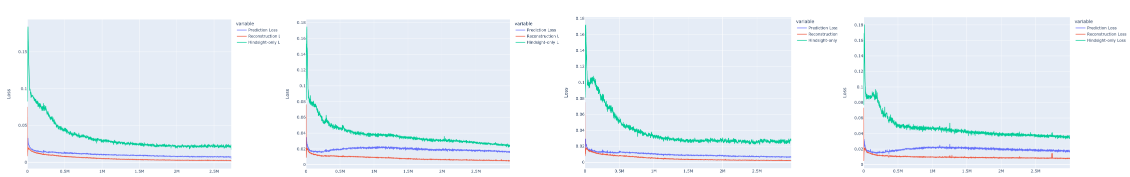
The intuition is as follows: Assuming realizability, at convergence "hindsight information" and "total stochasticity" cancel (i.e. neither more nor less), and the "remaining stochasticity" term goes to zero.

B Further Experiment Detail



(a) Sticky (Intrinsic Only) (b) Sticky (Ext. + Intrinsic) (c) Non-Sticky (Intrinsic Only) (d) Non-Sticky (Ext. + Intrinsic)

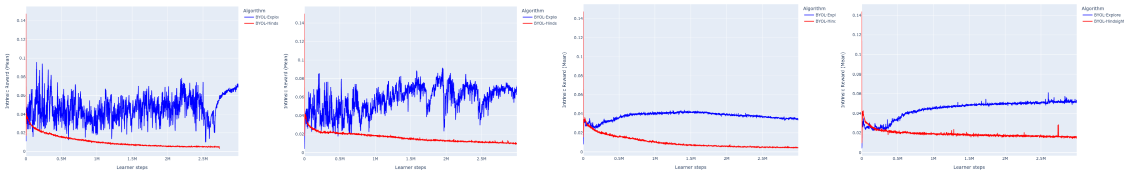
Figure 8: **Invariance Loss**. Invariance loss curves during training in Montezuma's Revenge for BYOL-Hindsight.



(a) Sticky (Intrinsic Only) (b) Sticky (Ext. + Intrinsic) (c) Non-Sticky (Intrinsic Only) (d) Non-Sticky (Ext. + Intrinsic)

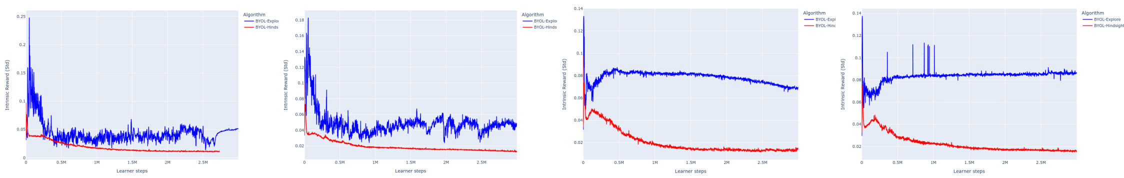
Figure 9: **Loss Comparison**. Prediction error, reconstruction error, and hindsight-only error for BYOL-Explore.

The derivation above makes it clear that "hindsight information" and "total stochasticity" can be mis-aligned in two ways: First, hindsight may capture *less* information than in the "true" latent (which means the reconstruction error is not driven to zero). Second, hindsight may capture *more* information than in the "true" latent (which means it leaks some information about the current state and action).



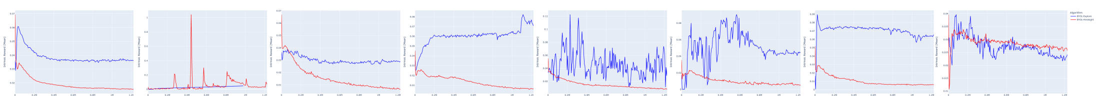
(a) Sticky (Intrinsic Only) (b) Sticky (Ext. + Intrinsic) (c) Non-Sticky (Intrinsic Only) (d) Non-Sticky (Ext. + Intrinsic)

Figure 10: **Intrinsic Reward**. Rewards during training in Montezuma's Revenge, for BYOL-Explore/Hindsight.

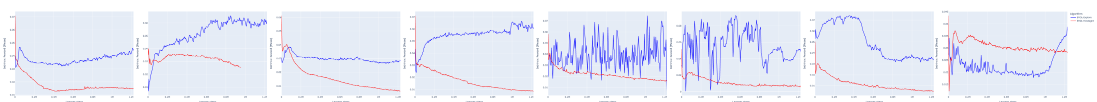


(a) Sticky (Intrinsic Only) (b) Sticky (Ext. + Intrinsic) (c) Non-Sticky (Intrinsic Only) (d) Non-Sticky (Ext. + Intrinsic)

Figure 11: **Reward Variance**. Std. dev. during training in Montezuma's Revenge, for BYOL-Explore/Hindsight.



(a) **Alien (I)** (b) **Freeway (I)** (c) **Gravitar (I)** (d) **Hero (I)** (e) **Montez. (I)** (f) **Pitfall (I)** (g) **Priv. Eye (I)** (h) **Venture (I)**



(i) **Alien (E+I)** (j) **Freeway (E+I)** (k) **Gravitar (E+I)** (l) **Hero (E+I)** (m) **Montez. (E+I)** (n) **Pitfall (E+I)** (o) **Priv. Eye (E+I)** (p) **Venture (E+I)**

Figure 12: **Intrinsic Reward**. Rewards during training in *Atari* games, with (red) and without (blue) hindsight.

Regarding the former, Figure 9 shows that the reconstruction loss indeed converges to very small values, but not exactly zero, which is consistent with the fact that Z does not perfectly capture the entirety of what is unpredictable. (This may be due to a variety of the usual factors, such as non-realizability and errors in optimization and estimation of expectations). Regarding the latter, to assess how well the invariance constraint between Z and X, A is enforced, Figure 8 shows the invariance loss over time. We observe that—to the best of the critic's ability to tell—the constraint is relatively well-enforced.

As another view into this, we can also gauge the amount of informational overlap between Z and X, A as follows: In addition to learning $f_\eta(X, A, Z)$, consider training additional predictors to predict Y, but only using states and actions as input (as in the usual forward prediction), or only using hindsight as input. From Figure 9 we observe that this hindsight-only error (i.e. using only Z as input) is the highest; prediction error (i.e. using only X, A) is lower, but not as low as reconstruction error (i.e. using X, A, Z as input). For reference, the variance of target vectors is around 0.4. These observations are consistent with the fact that leakage indeed occurs, but much of it is regularized away by the invariance constraint. Precisely, from the loss comparison we see that Z alone contains strictly less information than in X, A for predicting Y, and that their union X, A, Z contains the most information for predicting Y. (Note that much information about each room is statically determined by the room number, thus even a tiny amount if leakage may be sufficient to determine large portions of the future).

Finally, we stress that care is required when interpreting these curves due to the *bootstrapped* nature of the latent space in BYOL-Explore/Hindsight, as well as the fact that the dataset on which the predictor/reconstructor, generator, and critic are trained is *endogenous* to the rollout policy trained on the basis of the intrinsic reward. For reference, Figure 10 additionally shows the intrinsic rewards arising from BYOL-Explore and BYOL-Hindsight, and Figure 11 shows the variance of those rewards.

For the other hard-exploration games, Figure 12 shows the intrinsic rewards from BYOL-Explore and BYOL-Hindsight during training, corresponding to the performance results in Figure 7. Lastly, Table 2 compares BYOL-Explore and BYOL-Hindsight based on maximum scores obtained during training.

Table 2: *Atari Results (Sticky Actions)*. Max score over training for BYOL-Explore/Hindsight.

Games	*Intrinsic Only*		*Extrinsic + Intrinsic*	
	BYOL-Explore	BYOL-Hindsight	BYOL-Explore	BYOL-Hindsight
alien	2764	**14713**	9078	**95386**
freeway	17	**23**	26	**28**
gravitar	2893	**33719**	14734	**176847**
hero	685	**2510**	8810	**12890**
montezuma	10	**4920**	0	**6860**
pitfall	0	**2054**	0	**6927**
private eye	536	**19241**	1485	**38337**
venture	400	**910**	1920	**2060**