

Informed Asymmetric Actor-Critic: Leveraging Privileged Signals Beyond Full-State Access

Daniel Ebi¹, Gaspard Lambrechts², Damien Ernst², Klemens Böhm¹

¹Karlsruhe Institute of Technology, Karlsruhe, Germany

{daniel.ebi, klemens.boehm}@kit.edu

²University of Liège, Liège, Belgium

{gaspard.lambrechts, dernst}@uliege.be

Abstract

Reinforcement learning in partially observable environments requires agents to act under uncertainty from noisy, incomplete observations. Asymmetric actor-critic methods leverage privileged information during training to improve learning under these conditions. However, existing approaches typically assume full-state access during training. In this work, we challenge this assumption by proposing a novel actor-critic framework, called informed asymmetric actor-critic, that enables conditioning the critic on arbitrary privileged signals without requiring access to the full state. We show that policy gradients remain unbiased under this formulation, extending the theoretical foundation of asymmetric methods to the more general case of privileged partial information. To quantify the impact of such signals, we propose informativeness measures based on kernel methods and return prediction error, providing practical tools for evaluating training-time signals. We validate our approach empirically on benchmark navigation tasks and synthetic partially observable environments, showing that our informed asymmetric method improves learning efficiency and value estimation when informative privileged inputs are available. Our findings challenge the necessity of full-state access and open new directions for designing asymmetric reinforcement learning methods that are both practical and theoretically sound.

1 INTRODUCTION

Reinforcement learning (RL) has emerged as a powerful tool for optimizing control policies in various domains, including the control of heating, ventilation, and air conditioning systems [1], energy systems [2, 3], autonomous driving [4], and robotics [5]. However, when deploying RL in such real-world applications, agents must often operate under partial observability, relying on incomplete and noisy observations to make decisions. This setting is formalized by partially observable Markov decision processes (POMDPs) [6], where optimal actions depend on the history of past observations and actions.

To address this, RL methods for fully observable settings have been adapted by learning history-dependent policies, typically using recurrent neural networks (RNNs) to encode observation-action sequences [7, 8]. Although these methods are, in principle, capable of learning optimal history-dependent policies, they assume the same level of observability during training and execution, constraining policy learning to the limited information available at deployment. Yet, in practice, this assumption is unnecessarily restrictive and possibly suboptimal. Many training environments provide privileged information unavailable at execution, such as diagnostic sensors or simulators

exposing internal variables, without necessarily providing full access to the true state. Leveraging such asymmetric observability motivates the paradigm of asymmetric learning, which aims to exploit additional information at training while ensuring deployable history-dependent policies.

Asymmetric actor-critic methods provide a framework for this setting, as they condition the actor on observable histories while allowing the critic access to privileged information during training [9, 10]. However, existing approaches often assume either full-state access or no additional information, leaving the more general case of privileged partial information underexplored.

In this work, we introduce the informed asymmetric actor-critic framework, which generalizes prior asymmetric approaches by allowing the critic to condition on arbitrary state-dependent privileged inputs, without requiring full-state access. We show that this formulation yields unbiased estimators of value functions and policy gradients, extending the theoretical foundation of asymmetric learning to a broader class of training-time signals. To guide the selection and evaluation of privileged inputs, we propose two informativeness criteria: (i) a pre-training measure based on the Hilbert-Schmidt Conditional Independence Criterion (HSCIC), and (ii) a post-training metric based on return prediction error. These tools allow practitioners to quantify the utility of privileged signals before or after policy optimization. We empirically validate our informed asymmetric method on benchmark navigation tasks and in synthetic informed POMDP environments, demonstrating improved policy learning when informative privileged signals are used. Our findings highlight the importance of informativeness and challenge the assumption that full-state access is essential for asymmetric reinforcement learning.

2 RELATED WORK

Traditional RL methods have been adapted to partially observable settings by learning history-dependent policies that process sequences of past observations and actions using RNNs [11, 8, 7, 12, 13]. Since directly optimizing from raw histories is challenging, many approaches compress these sequences into compact latent representations, often by introducing auxiliary learning objectives [14, 15, 16].

Some works address partial observability by training privileged expert policies conditioned on true states and imitating them [17]. However, these methods often lack theoretical guarantees and may lead to suboptimal policies in POMDPs [18]. To mitigate this, Warrington et al. [18] propose constraining the expert policy to yield an optimal policy under partial observability. Another line of work exploits privileged information in model-based RL by constructing world models that summarize histories or integrate additional state signals. Examples include the Informed Dreamer [19], the Wasserstein Believer [20], and the Scaffolder [21].

Asymmetric actor-critic methods have emerged as a simple yet powerful framework for leveraging privileged information during training. By conditioning the critic on the full state and the actor on the history, these methods aim to guide policy updates more effectively. Pinto et al. [9] introduce an early asymmetric actor-critic approach that achieves strong empirical performance but suffers from biased gradient estimates [10]. Baisero and Amato [10] address this issue by introducing the history-state value function, explicitly modeling the relationship between histories and latent states to ensure unbiased gradients.

Recent theoretical work has established convergence guarantees for policy gradient and actor-critic methods in both fully and partially observable settings, including symmetric recurrent natural actor-critic methods using RNNs [14, 22] and asymmetric settings with linear function approximators [23].

Despite these advances, existing actor-critic methods typically assume either full access to the true state during training or no additional information at all. However, many real-world

settings fall between these extremes: some internal variables may be observable during training, while others remain hidden or only partially measurable. Methods that exploit privileged partial information, without requiring full-state access, are vastly unstudied, and it will be the main focus of this article.

3 BACKGROUND

In this section, we introduce the formal notion of partially observable Markov decision processes (POMDPs) and the informed POMDP framework, which motivates our informed asymmetric actor-critic framework.

3.1 Partially Observable Markov Decision Processes

A partially observable Markov decision process (POMDP) [6] models sequential decision-making under uncertainty as tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, P, \gamma)$, where \mathcal{S} , \mathcal{A} , and \mathcal{O} denote the state, action, and observation spaces. The transition probabilities $T(s' | s, a)$ describe the process dynamics. The agent emits observations via $O(o | s)$ and selects actions a_t based on the observable history h , defined as the sequence of past observations and actions. It receives a reward according to $R(s, a)$. We define the set of observable histories as $\mathcal{H} = \bigcup_{t=0}^{\infty} \mathcal{H}_t$, where $\mathcal{H}_t \subseteq \mathcal{O} \times (\mathcal{A} \times \mathcal{O})^t$ is the set of histories of size t . The objective is to maximize the expected return $J(\pi) = \mathbb{E}^{\pi} [\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$, where $\pi(a | h)$ denotes a history-dependent policy and $\gamma \in [0, 1)$ is a discount factor. P specifies the initial state distribution. The history-based Q-function is defined as

$$Q^{\pi}(h, a) = \mathbb{E}_{s_{0:\infty}, a_{0:\infty}}^{\pi} \left[\sum_{j=0}^{\infty} \gamma^j R(s_j, a_j) \middle| H_0 = h, A_0 = a \right],$$

and the corresponding value function as

$$V^{\pi}(h) = \sum_{a \in \mathcal{A}} \pi(a | h) Q^{\pi}(h, a).$$

In the following, we write conditional expectations by placing the conditioning in the subscript, e.g., $\mathbb{E}_{s_{0:\infty}, a_{0:\infty} | h, a}^{\pi} [\cdot]$.

3.2 Informed POMDPs

The informed POMDP [19] augments the POMDP definition by a so-called information space \mathcal{I} and a corresponding information function $I : \mathcal{S} \rightarrow \Delta(\mathcal{I})$, which gives the probability to obtain information $i_t \in \mathcal{I}$ in the true state $s_t \in \mathcal{S}$. Hence, the informed POMDP is defined by the 10-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{O}, T, I, \tilde{O}, R, P, \gamma)$.

In contrast to the POMDP, the observation function is defined as $\tilde{O} : \mathcal{I} \rightarrow \Delta(\mathcal{O})$ and denotes the probability to obtain $o_t \in \mathcal{O}$ given $i_t \in \mathcal{I}$. The main assumption in an informed POMDP is that the observation o_t is conditionally independent of the true state s_t given the information i_t , i.e., $p(o_t | i_t, s_t) = \tilde{O}(o_t | i_t)$. Each informed POMDP induces an underlying execution POMDP defined as $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, P, \gamma)$, where the observation function is given by

$$O(o_t | s_t) = \sum_{i \in \mathcal{I}} \tilde{O}(o_t | i) I(i | s_t).$$

3.3 Actor-Critic Paradigm

Actor-critic methods combine a policy model (actor), parameterized by θ , with a value estimator (critic), parameterized by ϑ . Under partial observability, both components typically condition on the observation-action history $h_t \in \mathcal{H}$ and are trained via sample-based gradients. The actor selects actions using $\pi_\theta(a_t | h_t)$, while the critic guides learning via value estimates.

In the symmetric setting, both actor and critic share the same input, i.e., h_t . The policy gradient is given by

$$\nabla_\theta J(\pi_\theta) = \mathbb{E} \left[\sum_t \gamma^t Q^\pi(h_t, a_t) \nabla_\theta \log \pi_\theta(a_t | h_t) \right]. \quad (1)$$

In practice, the critic estimates $Q^\pi(h_t, a_t)$ via the temporal-difference (TD) error

$$\delta_t = r_t + \gamma \hat{V}(h_{t+1}; \vartheta) - \hat{V}(h_t; \vartheta),$$

computed from value estimates $\hat{V}(\cdot; \vartheta)$.

Asymmetric actor-critic methods allow the critic to access additional information during training, unavailable to the actor and at execution. Prior approaches often rely on state-based critics $V^\pi(s)$ [9], which are generally ill-defined in POMDPs [10].

4 INFORMED ASYMMETRIC ACTOR-CRITIC

We introduce an asymmetric actor-critic framework that leverages arbitrary privileged information during training. Based on the informed POMDP paradigm, we define informed history-based value functions and derive an unbiased policy gradient for theoretically grounded asymmetric learning.

4.1 Informed History-based Value Functions

Given an informed POMDP \mathcal{P} , we first define the time-invariant informed history-based reward function $R(h, i, a)$, which incorporates additional state-conditioned information $i \sim I(i | s)$.

Definition 4.1 (Informed history-based reward function). *The informed history-based reward function $R(h, i, a)$ is the expected state-based reward $R(s, a)$ given the belief $p(s | h, i)$ about the true state $s \in \mathcal{S}$, i.e.,*

$$R(h, i, a) = \mathbb{E}_{s|h, i} [R(s, a)]. \quad (2)$$

Lemma 4.1 (Unbiasedness of the informed history-based reward). *In an informed POMDP, the informed history-based reward function $R(h, i, a)$ satisfies*

$$\mathbb{E}_{i|h} [R(h, i, a)] = R(h, a),$$

for all $h \in \mathcal{H}$ and $a \in \mathcal{A}$, where the expectation is taken under the belief $p(i | h)$.

Proof. Using the definition of the standard history-based reward function, i.e.,

$$R(h, a) = \mathbb{E}_{s|h} [R(s, a)] = \sum_{s \in \mathcal{S}} R(s, a) p(s | h), \quad (3)$$

and applying the law of total probability, we obtain:

$$\begin{aligned}
R(h, a) &= \sum_{s \in \mathcal{S}} p(s \mid h) R(s, a) \\
&= \sum_{s \in \mathcal{S}} \left(\sum_{i \in \mathcal{I}} p(s \mid h, i) p(i \mid h) \right) R(s, a) \\
&= \sum_{i \in \mathcal{I}} \left(\sum_{s \in \mathcal{S}} R(s, a) p(s \mid h, i) \right) p(i \mid h) \\
&= \mathbb{E}_{i|h} \left[\mathbb{E}_{s|h, i} [R(s, a)] \right] = \mathbb{E}_{i|h} [R(h, i, a)].
\end{aligned}$$

This concludes the proof. \square

By Lemma 4.1, $R(h, i, a)$ defines an unbiased estimator of the standard history-based reward $R(h, a)$. We assume that the reward function $R(s, a)$ is uniformly bounded by a constant $r_{\max} > 0$. This bound also applies to the standard and informed history-based rewards.

Next, we introduce the informed history Q -function, which conditions on h , i , and a .

Definition 4.2 (Informed history Q -function). *The informed history Q -function $Q^\pi(h, i, a)$ denotes the expected discounted return when starting from history $h \in \mathcal{H}$, privileged information $i \in \mathcal{I}$, and action $a \in \mathcal{A}$, and then following policy π :*

$$Q^\pi(h, i, a) = \mathbb{E}_{s_{0:\infty}, a_{0:\infty} | h, i, a} \left[\sum_{j=0}^{\infty} \gamma^j R(s_j, a_j) \right]. \quad (4)$$

Lemma 4.2 (Unbiasedness of the informed Q -function). *In an informed POMDP, the informed history Q -function satisfies*

$$\mathbb{E}_{i|h} [Q^\pi(h, i, a)] = Q^\pi(h, a),$$

for all $h \in \mathcal{H}$ and $a \in \mathcal{A}$.

Proof. Starting with the definition of the history Q -function and using the law of total expectation, we have:

$$\begin{aligned}
Q^\pi(h, a) &= \mathbb{E}_{s_{0:\infty}, a_{0:\infty} | h, a} \left[\sum_{j=0}^{\infty} \gamma^j R(s_j, a_j) \right] \\
&= \mathbb{E}_{i|h} \left[\mathbb{E}_{s_{0:\infty}, a_{0:\infty} | h, i, a} \left[\sum_{j=0}^{\infty} \gamma^j R(s_j, a_j) \right] \right] \\
&= \mathbb{E}_{i|h} [Q^\pi(h, i, a)].
\end{aligned}$$

This concludes the proof. \square

Hence, by Lemma 4.2, $Q^\pi(h, i, a)$ defines an unbiased estimator of the standard history-based value function $Q^\pi(h, a)$.

Based on the proposed $Q^\pi(h, i, a)$, we can define the time-invariant informed asymmetric value function that evaluates a history h of past observations and actions in conjunction with state-conditioned information i .

Definition 4.3 (Informed history value function). *The informed value function $V^\pi(h, i)$ denotes the expected return starting from history $h \in \mathcal{H}$ and additional information $i \in \mathcal{I}$:*

$$V^\pi(h, i) = \mathbb{E}_{s_{0:\infty}, a_{0:\infty} | h, i}^\pi \left[\sum_{j=0}^{\infty} \gamma^j R(s_j, a_j) \right]. \quad (5)$$

It satisfies the recursive form:

$$V^\pi(h, i) = \sum_{a \in \mathcal{A}} \pi(a | h) Q^\pi(h, i, a), \quad (6)$$

where the informed Q -function satisfies the Bellman equation:

$$Q^\pi(h, i, a) = R(h, i, a) + \gamma \mathbb{E}_{o', i' | i, a} [V^\pi(h', i')], \quad (7)$$

with $h' = hao'$.

In contrast to the history value $V^\pi(h)$, the informed history value $V^\pi(h, i)$ leverages additional state-conditioned context, potentially enabling a more comprehensive understanding of the environment's current state and its reward structure. Similarly, unlike the state value $V^\pi(s)$, the history-information pair $(h, i) \in \mathcal{H} \times \mathcal{I}$ provides a richer observable basis for forecasting agent behavior.

Lemma 4.3 (Unbiasedness of the informed value function). *In an informed POMDP, the informed value function satisfies for all $h \in \mathcal{H}$:*

$$\mathbb{E}_{i|h} [V^\pi(h, i)] = V^\pi(h).$$

Proof. Given the definition of the history value function, i.e.,

$$V^\pi(h) = \mathbb{E}_{s_{0:\infty}, a_{0:\infty} | h}^\pi \left[\sum_{j=0}^{\infty} \gamma^j R(s_j, a_j) \right],$$

and using the law of total expectation, we have:

$$\begin{aligned} V^\pi(h) &= \mathbb{E}_{s_{0:\infty}, a_{0:\infty} | h}^\pi \left[\sum_j \gamma^j R(s_j, a_j) \right] \\ &= \mathbb{E}_{i|h} \left[\mathbb{E}_{s_{0:\infty}, a_{0:\infty} | h, i}^\pi \left[\sum_j \gamma^j R(s_j, a_j) \right] \right] \\ &= \mathbb{E}_{i|h} [V^\pi(h, i)]. \end{aligned}$$

This concludes the proof. □

Hence, by Lemma 4.3, the informed history value function $V^\pi(h, i)$ is an unbiased estimator of the standard history value $V^\pi(h)$. Put differently, $V^\pi(h, i)$ provides, in expectation, the same signal as the standard history value function.

In the special case where the privileged information i corresponds to the full state $s \in \mathcal{S}$, i.e., $i = s$, the informed history value function reduces to the history-state value function of Baisero and Amato [10] (cf. Corollary A.1).

4.2 Informed Asymmetric Policy Gradient

Based on the informed history-based Q-function, we define the informed asymmetric policy gradient as:

$$\nabla_{\theta}^{\text{IAAC}} J(\pi_{\theta}) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi}(h_t, i_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t) \right],$$

where the actor conditions on observable histories, and the critic may condition on additional privileged inputs. We show that this gradient remains unbiased:

Theorem 4.1 (Informed asymmetric policy gradient). *Given an informed POMDP, the informed asymmetric policy gradient is equivalent to the standard policy gradient:*

$$\nabla_{\theta}^{\text{IAAC}} J(\pi_{\theta}) = \nabla_{\theta} J(\pi_{\theta}).$$

Proof. Given Equation 1 and following the Lemmas 4.2-4.3, we have

$$\begin{aligned} \nabla_{\theta} J(\pi_{\theta}) &= \mathbb{E} \left[\sum_t \gamma^t Q^{\pi}(h_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t) \right] \\ &\stackrel{(a)}{=} \sum_t \gamma^t \mathbb{E}_{h_t, a_t} [Q^{\pi}(h_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t)] \\ &\stackrel{(b)}{=} \sum_t \gamma^t \mathbb{E}_{h_t, a_t} [\mathbb{E}_{i_t | h_t} [Q^{\pi}(h_t, i_t, a_t)] \nabla_{\theta} \log \pi_{\theta}(a_t | h_t)] \\ &\stackrel{(c)}{=} \sum_t \gamma^t \mathbb{E}_{h_t, i_t, a_t} [Q^{\pi}(h_t, i_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t)] \\ &\stackrel{(d)}{=} \mathbb{E} \left[\sum_t \gamma^t Q^{\pi}(h_t, i_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t) \right] \\ &= \nabla_{\theta}^{\text{IAAC}} J(\pi_{\theta}). \end{aligned}$$

In (a) and (d), we use the linearity of the expectation to decompose or combine the summation over t and the expectation over (h_t, i_t, a_t) , respectively. In (b), using Lemma 4.2, we substitute $Q^{\pi}(h, a)$ with $\mathbb{E}_{i|h} [Q^{\pi}(h, i, a)]$, as the informed history-action value function is an unbiased estimate of $Q^{\pi}(h, a)$. By applying the law of total expectation in (c), i.e., $\mathbb{E}_{h_t, a_t, i_t} [\cdot] = \mathbb{E}_{h_t, a_t} [\mathbb{E}_{i_t | h_t} [\cdot]]$, we can rewrite the expression. This concludes the proof. \square

Thus, the critic can incorporate additional training-time signals without biasing policy updates. This result generalizes prior work and recovers the asymmetric policy gradient presented by Baisero and Amato [10] for $i_t = s_t$ (cf. Corollary A.2).

Based on Theorem 4.1, we introduce the informed history critic $\hat{V} : \mathcal{H} \times \mathcal{I} \rightarrow \mathbb{R}$, which estimates the informed history value $V^{\pi}(h_t, i_t)$ given the history h_t and additional information i_t . Combined with a history-dependent policy model $\pi_{\theta}(a_t | h_t)$, this yields an asymmetric actor-critic method, which we refer to as informed asymmetric actor-critic (IAAC). The informed asymmetric policy gradient is approximated by sampling $\hat{\nabla}_{\theta}^{\text{IAAC}} J(\pi_{\theta}) = \mathbb{E} [\sum_t \gamma^t \delta_t \nabla_{\theta} \log \pi_{\theta}(a_t | h_t)]$, where the TD errors $\delta_t = r_t + \gamma \hat{V}(h_{t+1}, i_{t+1}; \vartheta) - \hat{V}(h_t, i_t; \vartheta)$ are computed using the critic's informed value estimates.

5 INFORMATIVENESS OF PRIVILEGED SIGNALS

While the informed asymmetric actor-critic framework guarantees unbiased policy gradients for arbitrary additional information (cf. Theorem 4.1), not all privileged signals are equally beneficial for learning. In practice, some signals may accelerate policy optimization, while others may degrade value estimation or introduce instability when poorly correlated with the environment’s true state.

This motivates the need to quantify the informativeness of additional information $i_t \in \mathcal{I}$ with respect to the underlying control task. Specifically, we seek criteria that assess whether i_t provides useful structure for value estimation and can improve the policy learning of actor-critic methods.

In the following, we formalize two criteria that enable the comparison of different forms of privileged signals in terms of their informativeness about true returns.

5.1 Informativeness via Kernel-based Independence Criterion

The relevance of a privileged signal in the critic depends on whether it captures information about the return that is not already encoded in the history of past observations and actions. To quantify this, we evaluate the conditional dependence between the return $G_t := \sum_{j=0}^{T-t-1} \gamma^j R_{t+j}$ and , conditioned on h_t and a_t , using a non-parametric kernel-based approach.

Kernel methods embed probability distributions into a reproducing kernel Hilbert space (RKHS), where statistical relationships can be expressed as distances between mean elements. For instance, a distribution \mathbb{P} over X is represented as the kernel mean embedding $\mu_{\mathbb{P}} := \mathbb{E}_{x \sim \mathbb{P}}[k(x, \cdot)] \in \mathcal{H}_X$, where $k(\cdot)$ denotes a positive-definite kernel function.

Kernel-based measures such as Maximum Mean Discrepancy (MMD) [24] and the Hilbert-Schmidt Independence Criterion (HSIC) [25] enable non-parametric tests for differences in distributions and (in)dependence, respectively. In this work, we adopt the Hilbert-Schmidt Conditional Independence Criterion (HSCIC) [26], which extends HSIC to the conditional setting.

Given random variables X, Y, Z , HSCIC measures the RKHS norm between the conditional joint embedding and the product of conditional marginals:

$$\mathcal{J}(X, Y \mid Z) := \left\| \mu_{\mathbb{P}_{X,Y|Z}} - \mu_{\mathbb{P}_{X|Z}} \otimes \mu_{\mathbb{P}_{Y|Z}} \right\|_{\mathcal{H}_X \otimes \mathcal{H}_Y}.$$

Under standard conditions, such as a characteristic kernel on $X \times Y$ and that the conditional probability distributions given Z admit a regular version, HSCIC equals zero almost surely if and only if $X \perp\!\!\!\perp Y \mid Z$ (cf. Theorem C.1).

We leverage HSCIC to test whether i_t is conditionally dependent on the return G_t given h_t and a_t . This analysis does not require a trained critic and can be applied to trajectories collected under a random or exploratory policy. As a result, informativeness can be estimated prior to training, enabling informed decisions regarding which auxiliary signals to include in the critic.

Let $X_t = G_t$, $Y_t = i_t$, and $Z_t = (h_t, a_t)$. Given samples $\{(x_i, y_i, z_i)\}_{i=1}^n$, the empirical plug-in estimator of the squared HSCIC is given by

$$\begin{aligned} \hat{\mathcal{J}}_{X,Y|Z}^2(\cdot) &= k_Z^\top W (K_X \odot K_Y) W^\top k_Z - 2k_Z^\top W ((K_X W^\top k_Z) \odot (K_Y W^\top k_Z)) \\ &\quad + (k_Z^\top W K_X W^\top k_Z)(k_Z^\top W K_Y W^\top k_Z), \end{aligned}$$

where k_Z denotes the kernel function $k_Z(\cdot)$ on Z evaluated at the estimator input z_i , $[K_Y]_{ij} = k_Y(y_i, y_j)$, $W = (K_Z + \lambda \mathbf{I})^{-1}$, and \odot denotes the element-wise product.

Hypothesis test. We can construct a hypothesis test based on permutation testing to assess statistical significance. Let

$$\mathbb{H}_0 : G_t \perp\!\!\!\perp i_t \mid h_t, a_t, \quad \text{and} \quad \mathbb{H}_1 : G_t \not\perp\!\!\!\perp i_t \mid h_t, a_t.$$

To obtain a scalar test statistic, we average the pointwise HSCIC values over all sample realizations z_i of Z :

$$\bar{\mathcal{J}}_{X,Y|Z} = \frac{1}{n} \sum_{i=1}^n \hat{\mathcal{J}}_{X,Y|Z}(z_i).$$

Under \mathbb{H}_0 , the null distribution is generated by independently permuting $\{y_i\}$ while keeping $\{x_i, z_i\}$ fixed. For each of B permutations, we compute $\bar{\mathcal{J}}^{(b)}$, and evaluate the empirical p -value as $p = \frac{1}{B} \sum_{b=1}^B \mathbb{I}[\bar{\mathcal{J}}^{(b)} \geq \bar{\mathcal{J}}]$, where \mathbb{I} denotes the indicator function. We reject \mathbb{H}_0 at significance level $\alpha \geq 0$ if $p < \alpha$, indicating that i_t is informative for predicting G_t given h_t and a_t .

5.2 Informativeness via Return Prediction Error

Beyond pre-training evaluation of privileged signals using HSCIC, we propose a complementary post-hoc criterion to quantify the utility of a privileged signal i_t in improving return prediction after training. Importantly, this evaluation focuses exclusively on the value prediction accuracy and does not involve any policy-related metrics.

Consider two critics trained on the same task: a symmetric critic $Q(h_t, a_t) = \mathbb{E}[G_t \mid h_t, a_t]$, and an asymmetric critic $Q(h_t, i_t, a_t) = \mathbb{E}[G_t \mid h_t, i_t, a_t]$.

Let T be the length of a trajectory. We define the pointwise reduction in squared error from conditioning on i_t as

$$E_t := (Q(h_t, a_t) - G_t)^2 - (Q(h_t, i_t, a_t) - G_t)^2, \quad (8)$$

where a positive E_t indicates improved return prediction at time t . To summarize the benefit over a trajectory, we compute the empirical mean and variance of $\{E_t\}_{t=0}^T$:

$$\hat{x} = \frac{1}{T} \sum_{t=0}^{T-1} E_t, \quad \hat{\sigma}^2 = \frac{1}{T} \sum_{t=0}^{T-1} (E_t - \hat{x})^2.$$

Definition 5.1 ((ϵ, δ) -Informativeness). A privileged signal i_t is said to be (ϵ, δ) -informative if, with probability at least $1 - \delta$, the expected gain satisfies

$$\mathbb{E}[E_t] \geq \epsilon := \hat{x} - \sqrt{\frac{2\hat{\sigma}^2 \log(2/\delta)}{T}} - \frac{2C \log(2/\delta)}{3T}, \quad (9)$$

where $C := \max_{0 \leq t < T} |E_t|$.

This bound, derived via Bernstein's inequality, provides a lower confidence bound on the expected gain from including i_t , serving as both a quantitative metric and a test statistic for a hypothesis test.

Hypothesis test. We test the hypotheses

$$\mathbb{H}_0 : \epsilon \leq 0, \quad \text{and} \quad \mathbb{H}_1 : \epsilon > 0,$$

where ϵ is the lower bound defined in Definition 5.1. We reject \mathbb{H}_0 if $\epsilon > 0$, concluding that the privileged signal i_t is (ϵ, δ) -informative with confidence $1 - \delta$. Otherwise, we fail to reject \mathbb{H}_0 , indicating insufficient evidence that i_t improves value prediction. Hence, this test offers a statistically grounded post-training criterion for evaluating privileged signals based on their empirical effect on return prediction accuracy.

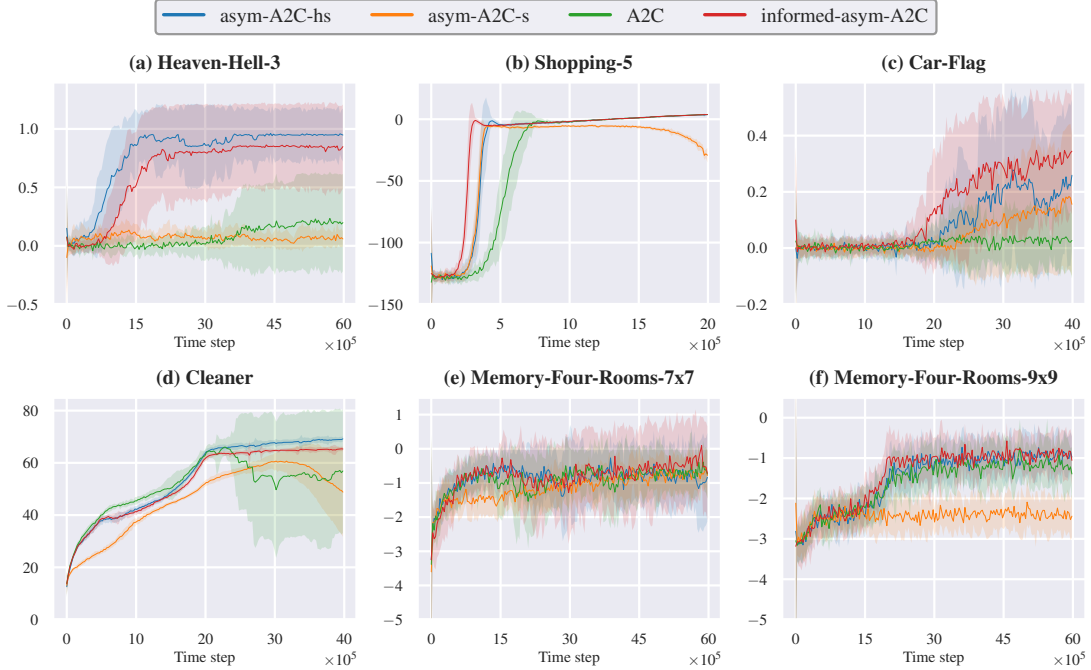


Figure 1: Learning performance on six benchmark navigation tasks. Curves show episodic returns averaged over the last 100 episodes, with means and standard deviations computed across 20 independent runs.

6 EXPERIMENTS

We evaluate the proposed informed asymmetric actor-critic framework on POMDP benchmarks against three baseline actor-critic variants. Moreover, we empirically validate the effectiveness of the presented informativeness criteria on synthetic informed POMDP instances.

6.1 Environments

We use six benchmark navigation tasks from the work of Baisero and Amato [10] to evaluate the learning performance of our method: *Heaven-Hell-3*, *Shopping-5*, *Car-Flag*, *Cleaner*, *Memory-Four-Rooms-7x7*, and *Memory-Four-Rooms-9x9*. Each environment is formulated as a POMDP, and we define task-specific privileged partial information accessible only to the critic. For the first two tasks, the privileged input corresponds to the Earth Mover’s Distance between the agent’s position and the target. In *Car-Flag*, the agent’s velocity is provided as additional information. For the remaining three environments, the privileged signal consists of an expanded spatial observation of the agent’s surroundings. See Appendix B for details.

To assess the expressiveness of our informativeness criteria, we generate synthetic informed POMDP instances with a finite state space ($|\mathcal{S}| = 10$), a discrete action space ($|\mathcal{A}| = 4$), and continuous observation and information spaces. Following the methodology of François-Lavet et al. [27], transition probabilities are randomly assigned by setting each (s, a, s') -entry to zero with probability 0.75, and sampling uniformly from $[0, 1]$ otherwise. To ensure valid transitions,

we assign a non-zero probability to a randomly chosen next state whenever all transitions from a given state-action pair are initially zero. We then normalize the probabilities to ensure they sum to one. Rewards are sampled uniformly from $[-1, 1]$ at initialization. Privileged information is generated by sampling from a Gaussian distribution centered on a state-specific embedding, with variance controlled by a noise parameter $\varsigma \in \mathbb{R}_{\geq 0}$. Observations are then obtained by applying a noisy linear transformation to the privileged information, with another noise parameter controlling the observation uncertainty.

6.2 Baselines

We compare our informed asymmetric actor-critic method against three advantage actor-critic (A2C) variants: (1) *A2C*, a symmetric approach using a history-based critic $\hat{V}(h)$; (2) *asym-A2C-s*, an asymmetric variant with a state-based critic $V(s)$; and (3) *asym-A2C-hs*, an asymmetric variant with a history-state critic $\hat{V}(h, s)$. For the benchmark environments and all baselines, we adopt the model architectures and hyperparameters recommended by Baisero and Amato [10].

6.3 Results and Discussion

We highlight three key results from our evaluation: (a) empirical learning curve comparison; (b) results from the hypothesis test assessing our HSCIC-based informativeness criterion; and (c) the boxplot distribution of the test statistic for our post-hoc informativeness criterion.

Learning curves. First, we compare *informed-asym-A2C* against the baselines across the six benchmark navigation tasks (Figure 1 (a)–(f)). In *Heaven-Hell-3*, *informed-asym-A2C* exhibits strong performance relative to *A2C* and *asym-A2C-s*, though it is slightly outperformed by *asym-A2C-hs*, which benefits from full-state access in the critic. In *Shopping-5*, the informed asymmetric actor-critic converges faster than *asym-A2C-hs* and achieves comparable final returns. In *Car-Flag*, *informed-asym-A2C* outperforms all baselines, demonstrating both higher sample efficiency and improved final performance. For *Cleaner*, *asym-A2C-hs* achieves marginally higher returns, but *informed-asym-A2C* converges at a similar rate with greater stability than *A2C*, which suffers a performance drop after 2.5 million steps. In the Memory-Four-Rooms tasks, *informed-asym-A2C* significantly outperforms both asymmetric baselines, particularly *asym-A2C-s*, which lacks critic access to history.

Overall, *informed-asym-A2C* matches or surpasses the performance of *asym-A2C-hs* and/or achieves faster convergence in most tasks while relying on less privileged information. This underscores the importance of not only leveraging asymmetric information but also structuring it to align effectively with task requirements.

Information	\bar{J} (mean \pm std)	p -val. (mean \pm std)
$i_t = \emptyset$	66.287 \pm 29.840	1.000 \pm 0.000
$i_t: \varsigma = 0.0$	132.435 \pm 90.216	0.108 \pm 0.144
$i_t: \varsigma = 0.1$	140.454 \pm 78.242	0.147 \pm 0.194
$i_t: \varsigma = 0.5$	122.671 \pm 56.914	0.136 \pm 0.229
$i_t: \varsigma = 0.9$	98.173 \pm 43.427	0.239 \pm 0.235
$i_t = s_t$	126.793 \pm 63.656	0.122 \pm 0.251

Table 1: HSCIC- and p -value statistics for different privileged signals, computed across 12 synthetic informed POMDP instances.

Kernel-based independence criterion. To investigate the role and informativeness of different privileged signals, we estimate the empirical mean pointwise HSCIC value $\bar{\mathcal{J}}$ and its null distribution across 12 synthetic informed POMDP instances. Gaussian RBF kernels k_G , k_I , and k_Z are used with bandwidths selected via the median heuristic, i.e., using the median pairwise distance between samples. We consider three configurations: (1) no privileged input, (2) noisy latent vectors i_t with varying noise levels ς , and (3) full-state access ($i_t = s_t$). Using random policies, we collect 20 episodes of length 25 and perform $B = 30$ permutations.

Table 1 reports the mean and standard deviation of $\bar{\mathcal{J}}$ and the corresponding p -values across informed POMDP instances. As noise increases, both the mean and variance of p -values generally increase (except for $\varsigma = 0.5$), suggesting reduced statistical significance and potential informativeness. Interestingly, full-state access appears on average less informative than noiseless partial input, consistent with trends observed in the learning curve analysis. However, the substantial variance across instances highlights the instance-specific nature of informativeness.

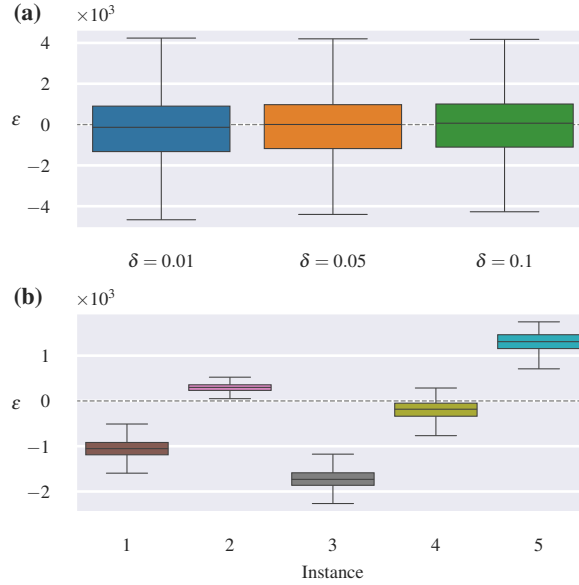


Figure 2: Boxplot distributions of ϵ over 1,000 test episodes for synthetic POMDP instances with privileged signal i_t ($\varsigma = 0.1$), computed for (a) different δ across 20 instances; (b) fixed $\delta = 0.05$ for five randomly sampled instances.

(ϵ, δ) -informativeness. We validate the post-hoc informativeness criterion on 20 synthetic informed POMDP instances with noisy privileged signals ($\varsigma = 0.1$). Specifically, we compare critic variants trained for 25,000 episodes with $T = 50$ using the Rec-NAC algorithm [14], with an Elman-type RNN of width 64 to encode the observation-action history. The RNN is followed by a fully connected layer with 256 units with ReLU activation, and a linear readout.

Figure 2 shows the boxplot distributions of ϵ computed over 1,000 test episodes for $\delta \in \{0.01, 0.05, 0.1\}$ across all instances (top), and for five randomly sampled instances with fixed $\delta = 0.05$ (bottom). The results indicate substantial variability across instances. In some cases, ϵ is consistently positive, reflecting a clear predictive benefit; in others, it centers near zero or is negative, suggesting limited informativeness or even harmful effects on return prediction. Across

instances, median values of ϵ remain close to zero for each evaluated δ , exceeding zero only for $\delta = 0.1$. These findings underscore the environment-dependent nature of (ϵ, δ) -informativeness.

7 CONCLUSION

We propose an informed asymmetric actor-critic method that generalizes the asymmetric approaches by allowing arbitrary privileged inputs in the critic without requiring full-state access, while preserving unbiased policy gradients. To guide the selection of such signals, we introduce two complementary informativeness criteria: a pre-training metric based on conditional dependence, and a post-hoc metric based on return prediction error. Our results show that privileged partial information can improve policy learning and value estimation, though its informativeness is task-dependent and not captured by return-based measures alone.

Future work may refine these criteria by exploring aspects such as signal complexity or direct policy impact, enabling better trade-offs between informativeness and model capacity for more robust and efficient learning under partial observability.

Acknowledgments

Daniel Ebi gratefully acknowledges the financial support of the German Research Foundation (DFG) as part of the Research Training Group GRK 2153: Energy Status Data – Informatics Methods for its Collection, Analysis and Exploitation. Gaspard Lambrechts gratefully acknowledges the financial support of the Wallonia-Brussels Federation and the Fonds de la Recherche Scientifique (FNRS) for his FRIA grant. Additionally, this work was supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition.

References

- [1] Khalil Al Sayed, Abhinandana Boodi, Roozbeh Sadeghian Broujeny, and Karim Beddiar. Reinforcement learning for HVAC control in intelligent buildings: A technical and conceptual review. *Journal of Building Engineering*, 95, 2024. ISSN 2352-7102.
- [2] Daniel Ebi, Edouard Fouché, Marco Heyden, and Klemens Böhm. MicroPPO: Safe power flow management in decentralized micro-grids with proximal policy optimization. In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2024.
- [3] Vincent François-Lavet, David Taralla, Damien Ernst, and Raphaël Fonteneau. Deep reinforcement learning solutions for energy microgrids management. In *European Workshop on Reinforcement Learning (EWRL 2016)*, 2016.
- [4] Ahmad Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017:70–76, 01 2017.
- [5] Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28694–28698, Apr. 2025.
- [6] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

- [7] Matthew Hausknecht and Peter Stone. Deep recurrent Q-learning for partially observable MDPs. In *2015 AAAI fall symposium series*, 2015.
- [8] Marvin Zhang, Zoe McCarthy, Chelsea Finn, Sergey Levine, and Pieter Abbeel. Learning deep neural network policies with continuous memory states. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 520–527. IEEE, 2016.
- [9] Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. *arXiv preprint arXiv:1710.06542*, 2017.
- [10] Andrea Baisero and Christopher Amato. Unbiased asymmetric reinforcement learning under partial observability. In *Proceedings of the Conference on Autonomous Agents and Multiagent Systems*, 2022.
- [11] Pengfei Zhu, Xin Li, Pascal Poupart, and Guanghui Miao. On improving deep reinforcement learning for POMDPs. *arXiv preprint arXiv:1704.07978*, 2017.
- [12] Daan Wierstra, Alexander Förster, Jan Peters, and Jürgen Schmidhuber. Recurrent policy gradients. *Logic Journal of IGPL*, 18(5):620–634, 2010.
- [13] Bram Bakker. Reinforcement learning with long short-term memory. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.
- [14] Semih Cayci and Atilla Eryilmaz. Recurrent natural policy gradient for POMDPs. In *ICML 2024 Workshop: Foundations of Reinforcement Learning and Control – Connections and Perspectives*, 2024.
- [15] Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive rl. *arXiv preprint arXiv:2401.08898*, 2024.
- [16] Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022.
- [17] Sanjiban Choudhury, Mohak Bhardwaj, Sankalp Arora, Ashish Kapoor, Gireeja Ranade, Sebastian Scherer, and Debadeepta Dey. Data-driven planning via imitation learning. *The International Journal of Robotics Research*, 37(13-14):1632–1672, 2018.
- [18] Andrew Warrington, Jonathan W Lavington, Adam Scibior, Mark Schmidt, and Frank Wood. Robust asymmetric learning in POMDPs. In *International Conference on Machine Learning*, pages 11013–11023. PMLR, 2021.
- [19] Gaspard Lambrechts, Adrien Bolland, and Damien Ernst. Informed POMDP: Leveraging additional information in model-based RL. *Reinforcement Learning Journal*, 2024.
- [20] Raphaël Avalos, Florent Delgrange, Ann Nowe, Guillermo Perez, and Diederik M Roijers. The wasserstein believer: Learning belief updates for partially observable environments through reliable latent space models. In *The Twelfth International Conference on Learning Representations*, 2024.

- [21] Edward S. Hu, James Springer, Oleh Rybkin, and Dinesh Jayaraman. Privileged sensing scaffolds reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.
- [22] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.
- [23] Gaspard Lambrechts, Damien Ernst, and Aditya Mahajan. A theoretical justification for asymmetric actor-critic algorithms. In *Forty-second International Conference on Machine Learning*, 2025.
- [24] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.
- [25] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
- [26] Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21247–21259. Curran Associates, Inc., 2020.
- [27] Vincent François-Lavet, Guillaume Rabusseau, Joelle Pineau, Damien Ernst, and Raphael Fonteneau. On overfitting and asymptotic bias in batch reinforcement learning with partial observability. *J. Artif. Int. Res.*, 65(1):1–30, 2019.
- [28] Hector Geffner and Blai Bonet. Solving large POMDPs using real time dynamic programming. In *Working Notes Fall AAAI Symposium on POMDPs*, 1998.
- [29] Andrea Baisero. gym-pomdps: Gym environments from POMDP files. <https://github.com/abaisero/gym-pomdps>, 2019. Accessed: 2025-08-01.
- [30] Hai Nguyen. POMDP Robot Domains. <https://github.com/hai-h-nguyen/pomdp-domains>, 2021. Accessed: 2025-08-01.
- [31] Shuo Jiang and Christopher Amato. Multi-agent reinforcement learning with directed exploration and selective memory reuse. In *Proceedings of the ACM Symposium on Applied Computing*, pages 777–784, 03 2021.
- [32] Andrea Baisero and Sammie Katt. gym-gridverse: Gridworld domains for fully and partially observable settings. <https://github.com/abaisero/gym-gridverse>, 2021. Accessed: 2025-08-01.
- [33] Andrea Baisero and Sammie Katt. asym-porl: Asymmetric methods for partially observable reinforcement learning. <https://github.com/abaisero/asym-rlpo>, 2021. Accessed: 2025-08-01.

A AUXILIARY RESULTS

This section collects our auxiliary results.

Corollary A.1 (Relation of $V^\pi(h, i)$ to the history-state value function of Baisero and Amato [10]). *The informed history value function $V^\pi(h, i)$ reduces to the history-state value function for $i = s$, where $s \in \mathcal{S}$ denotes the true environment state. In particular,*

$$V^\pi(h, s) = \sum_{a \in \mathcal{A}} \pi(a | h) Q^\pi(h, s, a),$$

where the history-state action-value function is defined as

$$Q^\pi(h, s, a) = R(s, a) + \gamma E_{s', o' | s, a} [V^\pi(h', s')],$$

with $s' \sim T(s' | s, a)$, $o' \sim \tilde{O}(o' | s')$, $i' = s'$, and h' denoting the updated history resulting from appending action a and observation o' to h .

By Lemma 4.3, this formulation provides an alternative unbiased estimator of the history value function:

$$V^\pi(h) = E_{s|h} [V^\pi(h, s)],$$

as previously established by Baisero and Amato [10].

Corollary A.2 (Relation of $\nabla_\theta^{\text{IAAC}} J(\pi_\theta)$ to the asymmetric policy gradient of Baisero and Amato [10]). *The informed asymmetric policy gradient $\nabla_\theta^{\text{IAAC}} J(\pi_\theta)$ reduces to the asymmetric policy gradient introduced by Baisero and Amato [10] for $i = s$, where $s \in \mathcal{S}$ denotes the true environment state. In particular,*

$$\nabla_\theta^{\text{AC}} J(\pi_\theta) = E \left[\sum_{t=0}^{\infty} \gamma^t Q^\pi(h_t, s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | h_t) \right].$$

Following Lemma 4.1-4.3, this formulation recovers an alternative asymmetric policy gradient estimator that is equivalent to the standard policy gradient:

$$\nabla_\theta^{\text{AC}} J(\pi_\theta) = \nabla_\theta J(\pi_\theta),$$

as established by Baisero and Amato [10].

Generality of state-dependent information functions. It is worth noting that the assumption of the information function $I : \mathcal{S} \rightarrow \Delta(\mathcal{I})$ depending solely on the current state s_t is not restrictive in the context of informed POMDPs. For instance, consider a more general setting where the information variable i_t is sampled from a distribution conditioned on the current state, action, and next state, i.e., $i_t \sim \mathbb{P}(i_t | s_t, a_t, s_{t+1})$. In such cases, the model can be reformulated as an informed POMDP by augmenting the state space to $\tilde{\mathcal{S}} = \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, and defining a corresponding transition function \tilde{T} and reward function \tilde{R} over this augmented space. Under this formulation, the information function retains the standard informed POMDP form $\tilde{I} : \tilde{\mathcal{S}} \rightarrow \Delta(\mathcal{I})$, thereby demonstrating that the state-only dependency assumption is without loss of generality.

B BENCHMARK ENVIRONMENTS

In this section, we describe the partially observable benchmark environments used in our experiments.

Heaven-Hell-3. The Heaven-Hell task [28, 29] is a partially observable navigation problem within a gridworld environment characterized by a corridor-like structure with a fork leading to three distinct terminal branches. Two of these branches correspond to terminal exits: one leading to a positive outcome (heaven) and the other to a negative outcome (hell). The third branch leads to a non-terminal location where the agent can interact with an oracle (referred to as a "priest") who provides information necessary to disambiguate the exits. The agent is initially unaware of which terminal corresponds to heaven.

The underlying state includes both the agent's position and the true location of the heaven exit. As observation, however, the agent either perceives its own location or, when visiting the priest, receives an observation that reveals the location of heaven. We construct privileged partial information by adding to the agent's location its distance to the heaven terminal using Earth Mover's distance.

At each time step, the agent selects an action from the discrete set **NORTH, SOUTH, EAST, WEST**. The environment is deterministic, and movement is constrained by the grid-world layout. To solve the task optimally, the agent must first visit the priest to acquire the necessary information about the correct exit, then return to the fork and proceed to the identified heaven location.

The agent receives sparse feedback in the form of a terminal reward:

- a reward of 1.0 for exiting to heaven, and
- a reward of -1.0 for exiting to hell.

Shopping-5. The Shopping-5 environment [29] models another grid-world navigation task in which an agent must buy a forgotten item from a store. The environment is modeled as a two-dimensional gridworld of size 5×5 , with the item placed randomly at one of the grid cells. The agent begins at an arbitrary location and must locate and buy the item. While the agent's position is fully observable, the item's position is hidden and must be explicitly queried.

Hence, the full state encodes both the agent's position and the item's location, represented compactly as integers. Observations are similarly encoded, but are partial: at each time step, the agent observes either its own position or, upon executing a query, the position of the item. Similar to the Heaven-Hell task, we introduce a privileged partial by computing the current Earth Mover's distance between the agent and the item.

At each time step, the agent selects an action from the discrete set **{UP, DOWN, LEFT, RIGHT, QUERY, BUY}**. The four movement actions update the agent's position deterministically within the bounds of the grid. Executing the **QUERY** action returns the location of the item, but is subject to a cost. The **BUY** action attempts to purchase the item at the agent's current position; if executed in the correct cell, it completes the task successfully.

The environment provides a dense reward signal to encourage efficient behavior:

- a reward of -1.0 for moving,
- a reward of -2.0 for querying the item's location,
- a reward of -5.0 for a **BUY** action in the wrong cell, and
- a reward of +10.0 for a **BUY** action in the correct cell.

Optimal behavior requires the agent to query the item’s location once, retain that information internally, and efficiently navigate to the target cell before executing a successful BUY action.

Car-Flag. The Car-Flag environment [30] models a continuous control task where an agent controls a car moving along a one-dimensional track via discrete force-control actions. At the two ends of the track are terminal flags: one representing a positive outcome (the good flag) and the other a negative outcome (the bad flag). Reaching either flag terminates the episode. Additionally, an intermediate information flag is placed along the track; when reached, it reveals the position of the good flag. While the task is conceptually similar to Heaven-Hell, key differences are the force-control and the position of the information flag.

Both the state and observation spaces are represented as three-dimensional real-valued vectors. The state includes the agent’s position, velocity, and the position of the good flag. The observation mirrors the state structure, but the third component (i.e., the good flag’s position) is masked, i.e., set to zero, when the agent is outside the observation range of the information flag; and the agent’s velocity is always hidden. In the informed setting, we provide the agent its velocity as a privileged partial signal.

At each time step, the agent selects an action from a discrete set of seven force-control inputs: LEFT_HIGH, LEFT_MEDIUM, LEFT_LOW, RIGHT_LOW, RIGHT_MEDIUM, RIGHT_HIGH, and NONE. These actions apply varying levels of acceleration to the left or right, or maintain zero acceleration.

The environment provides a sparse, terminal reward signal:

- a reward of 1.0 for reaching the good flag, and
- a reward of -1.0 for reaching the bad flag.

Optimal behavior requires the agent to first locate the information flag to identify the correct goal, then apply appropriate force controls to reach the good flag while avoiding the bad one.

Cleaner. Originally designed as a two-agent cooperative task, the Cleaner environment [31] is adapted in this work to a single-agent control problem via fully centralized training and execution. In this formulation, the joint actions and observations are constructed via the Cartesian product of the corresponding spaces of the two individual agents. The environment is a maze-like 13×13 grid-world in which two robots must collectively traverse and clean the entire area. The task is considered complete once every non-wall cell has been visited by at least one of the agents.

The full environment state is represented as a binary tensor of shape $13 \times 13 \times 5$, where each channel encodes the presence of: (i) a wall, (ii) a dirty cell, (iii) a cleaned cell, (iv) the first agent, and (v) the second agent. Each agent’s local observation is a $3 \times 3 \times 3$ binary tensor that captures the immediate neighborhood centered around the agent, including information about walls, dirty cells, and clean cells. As privileged input, the critic receives a $13 \times 13 \times 5$ tensor encoding the agent’s own position within the grid world, while masking out the position of the other agent by setting its corresponding cells to zero.

Each agent independently selects from four movement actions: UP, DOWN, LEFT, and RIGHT. In the centralized setting, where both agents are controlled jointly, the action space is the Cartesian product of the individual action sets, yielding a total of 16 composite actions.

At each time step, the agent receives a reward proportional to the number of new cells cleaned during that step. The possible reward values are:

- a reward of 0.0 if no new cells are cleaned,
- a reward of 1.0 if one agent cleaned a new cell, and
- a reward of 2.0 if both agents cleaned a new cell.

Memory-Four-Rooms. The so-called Gridverse suite [32] defines a collection of partially observable environments in which agents interact within structured gridworlds. In this work, we consider the 7×7 -Memory-Four-Room and 9×9 -Memory-Four-Room environments. While actions are encoded as categorical indices, both states and observations are structured representations comprising multiple semantically meaningful components. Importantly, these components differ between state and observation, and some are only available in the state representation. The key components are:

- **Grid component:** A tensor of shape $3 \times 7 \times 7$ for 7×7 -Memory-Four-Room or $3 \times 9 \times 9$ for 9×9 -Memory-Four-Room, where each channel encodes a semantic property of the environment (e.g., cell type, cell color, or status). The observation includes a rotated, agent-centric $3 \times 2 \times 3$ -view of this grid rendered from the first-person perspective of the agent. Cells obstructed by walls are occluded in the observation.
- **Agent-ID-Grid component:** A binary matrix of size 7×7 or 9×9 , respectively, indicating the agent’s absolute position. This component is included only in the state.
- **Agent component:** A three-dimensional categorical array encoding the agent’s position and orientation. In the state, this is expressed in absolute coordinates, while in the observation, it is provided relative to the agent’s perspective and is thus constant, and not necessary for control.

The environment contains a good exit, a bad exit, and a beacon, each placed randomly at the start of each episode. The beacon shares its color with the good exit, and successful task completion requires the agent to first locate the beacon, memorize its color, and then navigate to the exit of matching color while avoiding the bad exit.

As privileged input, the critic is provided with an agent-centered $3 \times 3 \times 5$ tensor, offering an expanded view of the agent’s local surroundings.

At each time step, the agent selects from the following discrete action set: `MOVE_FORWARD`, `MOVE_BACKWARD`, `MOVE_LEFT`, `MOVE_RIGHT`, `TURN_LEFT`, `TURN_RIGHT`, `PICK_N_DROP`, and `ACTUATE`. The `MOVE_` actions are interpreted relative to the agent’s orientation, while `TURN_` modifies the orientation itself. Although the action set includes `PICK_N_DROP` and `ACTUATE` for generality, these are no-ops in the Memory-Four-Rooms tasks, as there are no doors or pickable objects.

The reward signal is composed of the following terms:

- a living reward of -0.05 per time step,
- a reward of +5.0 for reaching the good exit, and
- a reward of -5.0 for reaching the bad exit.

C IMPLEMENTATION DETAILS

In this section, we detail some parts of our implementation used in the evaluation. All experiments were conducted on a cluster node equipped with 64 cores running at 3.0 GHz and 72 GB of RAM allocated per task.

C.1 Model Architectures

In the following, we describe the model architectures employed for the actor and critic networks in each environment.

Benchmark tasks. For the benchmark tasks, we use the implementation [33] of environments and actor-critic methods provided by Baisero and Amato [10], extending them to the informed setting.

In each task, a 128-dimensional single-layer gated recurrent unit (GRU) encodes the concatenated action and observation features into a history representation. While the actor and critic networks share this architectural component, their parameters are maintained separately. The subsequent actor and critic network components vary across environments as follows:

- For the Heaven-Hell-3 and Shopping-5 tasks, we employ a 64-dimensional embedding model to represent states, actions, and observations. Both the actor and critic networks consist of two-layer feedforward neural networks with 512 and 256 units, respectively, using ReLU activations in the hidden layers and a linear output layer.
- For the Car-Flag and Cleaner environments, actions are represented as one-hot encodings of their respective categorical indices. As the state and observation representations provided by these environments are already flattened and structurally simple, no additional embedding is applied. The actor and critic subsequent networks adopt the same architecture used for the Heaven-Hell-3 and Shopping-5 tasks.
- For the Memory-Four-Room tasks, the $3 \times 2 \times 3$ observation tensors are initially processed by an embedding layer that maps each categorical value to an 8-dimensional vector. The resulting embedded tensor is then flattened into a 144-dimensional feature vector, which serves as the observation input to both the actor and critic networks. Actions, provided as categorical indices, are represented using one-dimensional embedding layers. For the states, the grid component is first embedded and then concatenated with the agent-ID grid. A three-layer convolutional network subsequently processes this combined input. The output of the convolutional network is concatenated with the agent components. The actor and critic networks each consist of a hidden layer with 512 units using ReLU activation, followed by a linear output layer.

We encode the privileged information analogously to the observations. The embedded privileged information is then concatenated with the latent history representation before being passed to the task-specific feedforward neural network.

For each environment and method, we use the hyperparameter values recommended by Baisero and Amato [10] to ensure comparability with prior work. Table 2 summarizes the actor learning rate α_π , critic learning rate $\alpha_{\hat{v}}$, and the initial negative-entropy weight λ_0 selected for each environment. Additionally, the following model hyperparameters are applied across all environments: discount factor is set to $\gamma = 0.99$, episodes are automatically terminated if they exceed 100 time steps; two episodes are sampled per gradient update; a frozen target network is

used to stabilize critic training, with target parameters updated every 10,000 time steps; and the negative-entropy weight λ decays linearly over 2 million time steps to a final value equal to one-tenth of λ_0 .

Environment	α_π	$\alpha_{\hat{V}}$	λ_0
Heaven-Hell-3	0.001	0.001	0.1
Shopping-5	0.001	0.0003	3.0
Car-Flag	0.001	0.001	0.03
Cleaner	0.001	0.001	1.0
7×7 -Memory-Four-Room	0.0003	0.001	0.1
9×9 -Memory-Four-Room	0.001	0.0003	0.3

Table 2: Hyperparameters for the benchmark environments.

Synthetic informed POMDPs. For the synthetic informed POMDP environments, the actor is implemented as an Elman-type recurrent neural network (RNN) of width $m_a = 64$, followed by a linear readout layer, as in the symmetric Rec-NAC algorithm [14]. The informed critic consists of an Elman-type RNN of width $m_c = 64$, followed by a feedforward neural network with 256 hidden units and ReLU activation, and a linear output layer. Across all environments, we use a fixed discount factor of $\gamma = 0.99$.

C.2 Kernel-based informativeness criterion

In this work, we leverage the following result of [26] to construct a hypothesis test for conditional independence:

Theorem C.1 (Theorem 5.4 in [26]). *Suppose $k_X \otimes k_Y$ is a characteristic kernel on $X \times Y$, and that $\mathbb{P}(\cdot | Z)$ admits a regular version. Then $\mathcal{I}(X, Y | Z) = 0$ almost surely if and only if $X \perp\!\!\!\perp Y | Z$.*

We implement the HSCIC-based informativeness criterion using the `PyRKHSstats` library¹. Specifically, we employ Gaussian RBF kernels with bandwidth parameters selected via the median heuristic, i.e., using the median pairwise distance between samples. To improve test sensitivity while maintaining valid Type I error control, kernel parameters can be optimized by splitting the data set into two disjoint subsets, one for bandwidth selection and the other for computing the HSCIC statistic and null distribution.

¹<https://github.com/Black-Swan-ICL/PyRKHSstats>