Informed Asymmetric Actor-Critic: Theoretical Insights and Open Questions

Daniel Ebi
Karlsruhe Institute of Technology
Karlsruhe, Germany
daniel.ebi@kit.edu

Gaspard Lambrechts
University of Liège
Liège, Belgium
gaspard.lambrechts@uliege.be

Damien Ernst
University of Liège
Liège, Belgium
dernst@uliege.be

Klemens Böhm
Karlsruhe Institute of Technology
Karlsruhe, Germany
klemens.boehm@kit.edu

Abstract

Reinforcement learning in partially observable environments requires agents to make decisions under uncertainty, based on incomplete and noisy observations. Asymmetric actor-critic methods improve learning in these settings by exploiting privileged information available during training. Most existing approaches, however, assume full access to the true state. In this work, we present a novel asymmetric actor-critic formulation grounded in informed partially observable Markov decision processes, allowing the critic to leverage arbitrary privileged information without requiring full-state access. We show that the method preserves the policy gradient theorem and yields unbiased gradient estimates even when the critic conditions on privileged partial information. Furthermore, we provide a theoretical analysis of the informed asymmetric recurrent natural policy gradient algorithm derived from our informed asymmetric learning paradigm. Our findings challenge the assumption that full-state access is necessary for unbiased policy learning, motivating the need to develop well-defined criteria to quantify the informativeness of additional training signals and opening new directions for asymmetric reinforcement learning.

1 Introduction

Reinforcement learning (RL) has emerged as an effective framework for optimizing control policies in a variety of domains, including heating, ventilation, and air conditioning (HVAC) control [1], energy system management [2, 3], autonomous driving [4], and robotics [5].

However, in real-world deployments, RL agents frequently operate under partial observability, where decisions must be made from incomplete and noisy observations. This scenario is formalized by the partially observable Markov decision process (POMDP) formulation [6], in which optimal decision-making depends on the sequence of prior observations and actions. To address this, methods originally designed for fully observable settings are extended to learn history-dependent policies, typically by employing recurrent neural networks (RNNs) to encode observation–action histories [7, 8, 9, 10, 11]. While these methods are theoretically capable of learning optimal history-dependent policies, they often assume that the agent's observability is identical during training and deployment. As a result, policy learning is constrained to the limited information available at execution, which can be unnecessarily restrictive and possibly suboptimal. In practice, training environments frequently provide privileged signals that are unavailable at deployment, such as measurements from diagnostic

18th European Workshop on Reinforcement Learning (EWRL 2025).

sensors or internal variables from simulators, without requiring full access to the true state. The paradigm of asymmetric learning seeks to leverage such additional training-time information to guide policy optimization, while ensuring that the resulting history-dependent policies remain executable under deployment-time observability.

A common strategy to address asymmetric observability is first to learn privileged policies conditioned on the true state and subsequently imitate them [12]. While effective in practice, these approaches often lack theoretical guarantees, which can result in suboptimal policies for POMDPs [13]. To mitigate this limitation, Warrington et al. [13] introduce constraints that enable safe imitation of expert policies, ensuring optimality under partial observability. Privileged information has also been leveraged in model-based RL by constructing world models that either summarize past histories or incorporate additional state signals. Approaches of this type include the Informed Dreamer [14], the Wasserstein Believer [15], and the Scaffolder [16].

Another category of methods is asymmetric actor-critic approaches, in which the critic is conditioned on the full state while the actor relies on the history, aiming for more accurate policy updates. The early asymmetric actor-critic approach by Pinto et al. [17] achieves strong empirical performance but suffers from biased gradient estimates [18]. This issue is addressed by Baisero and Amato [18], who propose the history-state value function to explicitly model the relationship between histories and latent states, ensuring unbiased gradients. Applications of asymmetric actor-critic methods demonstrate good performance across domains [19, 20, 21]. Most actor-critic methods either assume full access to the true state during training or rely solely on information available at deployment. However, many environments often fall between these extremes: some internal variables may be observable during training, while others remain hidden or only partially accessible. In this work, we focus on methods that can effectively leverage privileged partial information, a largely unexplored setting.

Recent theoretical work has established convergence guarantees for both policy gradient and actor-critic methods in fully observable Markov decision process (MDP) settings. Natural policy gradient (NPG) methods are analyzed using linear and feedforward neural network approximators [22, 23], while actor-critic methods demonstrate provable convergence under both i.i.d. and Markovian sampling assumptions [24, 25, 26]. Extending these analyses to partially observable settings, recent work has begun to address recurrent policy optimization in POMDPs [27]. Notably, Cayci and Eryilmaz [27] provide convergence guarantees for symmetric natural actor-critic methods employing RNNs. Theoretical analysis of asymmetric actor-critic algorithms using linear function approximation has also emerged [28, 29], along with belief-weighted asymmetric variants [30]. Building on the results of [27] for the symmetric recurrent natural policy gradient method, our work extends the analysis to asymmetric settings where the critic has access to a privileged partial signal during training.

To this end, we present an asymmetric actor-critic method grounded in the framework of informed POMDPs. Our approach relaxes the standard assumption of full-state observability during training by allowing the critic to condition on privileged partial information, falling between fully privileged and unprivileged settings. The policy remains history-dependent and executable solely from past observations and actions. We show that the informed asymmetric critic is well-defined and unbiased, preserving the policy gradient theorem under partial privileged conditioning. By demonstrating that any state-conditional random variable can be leveraged in an unbiased manner, our work raises new questions regarding which privileged information is sufficient or necessary for effective learning. Additionally, we analyze the finite-time and finite-width convergence properties of the informed asymmetric variant of the recurrent natural actor-critic (Rec-NAC) algorithm and empirically evaluate its performance in a simulated partially observable environment. Our results highlight the need for a sound approach to evaluate the informativeness of privileged signals, enabling a systematic assessment of the practical benefits of asymmetric actor-critic methods while accounting for the trade-offs with model complexity.

2 Background

In this section, we describe the decision processes and actor-critic methods considered in this work, with a focus on the informed asymmetric actor-critic framework. We also introduce the infinite-width limit and neural tangent kernel, which provide a nonparametric perspective on learning dynamics and form the basis of our theoretical analysis. An overview of the notation can be found in Appendix A.

2.1 Partially observable Markov decision processes

A partially observable Markov decision process (POMDP) [6] is a discrete-time partially observable control problem defined by an 8-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, P, \gamma)$, where $\mathcal{S}$ denotes the state space, and $\mathcal{A}$ is the action space. $P \in \Delta(\mathcal{S})$ is the initial state distribution that gives the probability of $\boldsymbol{s}_0 \in \mathcal{S}$ being the process' initial state. The transition function $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ defines the dynamics of the system by providing the probability of transitioning to state $\boldsymbol{s}_{t+1} \in \mathcal{S}$ after taking action $\boldsymbol{a}_t \in \mathcal{A}$ in state $\boldsymbol{s}_t \in \mathcal{S}$. The reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifies the expected immediate reward received for taking a particular action in a given state, and $\gamma \in [0, 1)$ is a discount factor that specifies the relative importance of future rewards. Together, the tuple $(\mathcal{S}, \mathcal{A}, T, R, P, \gamma)$ describes a fully observable Markov decision process (MDP). What sets a POMDP apart from a fully observable MDP is the fact that the agent cannot directly observe the true state $\boldsymbol{s}_t \in \mathcal{S}$ of the environment. Instead, it receives an observation $\boldsymbol{o}_t \in \mathcal{O}$, where $\mathcal{O}$ denotes the observation space. The observation function $O : \mathcal{S} \to \Delta(\mathcal{O})$ gives the probability of obtaining observation $\boldsymbol{o}_t \in \mathcal{O}$ in state $\boldsymbol{s}_t \in \mathcal{S}$.

In the partially observable setting, the agent must select actions based on the observable history, defined as the sequence of past observations and actions. We represent each history as $\boldsymbol{h}$ by flattening this sequence and 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. To achieve optimal behavior under partial observability, the agent typically needs to consider the entire history, meaning its policy maps histories to action distributions, i.e., $\pi : \mathcal{H} \to \Delta(\mathcal{A})$. The agent's objective is to maximize the expected return, specifically the expected discounted sum of rewards, given by $J(\pi_\theta) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{s}_t, \mathbf{a}_t)\right]$. The history reward function is given by $R(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{\mathbf{s}|\boldsymbol{h}}\left[R(\mathbf{s}, \boldsymbol{a})\right]$. Moreover, the policy's history value function $V^{\pi} : \mathcal{H} \to \mathbb{R}$ is defined as the expected return following an observable history $\boldsymbol{h}$:

$$V^{\pi}(\boldsymbol{h}) = \mathbb{E}^{\pi}_{\mathbf{s}_{0:\infty}, \mathbf{a}_{0:\infty}|\boldsymbol{h}}\left[\sum_{j=0}^{\infty} \gamma^j \, R(\mathbf{s}_j, \mathbf{a}_j)\right], \tag{1}$$

supporting an indirect recursive Bellman form:

$$V^{\pi}(\boldsymbol{h}) = \sum_{\boldsymbol{a} \in \mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{h}) Q^{\pi}(\boldsymbol{h}, \boldsymbol{a}), \tag{2}$$

where the history Q-function is $Q^{\pi}(\boldsymbol{h}, \boldsymbol{a}) = R(\boldsymbol{h}, \boldsymbol{a}) + \gamma \mathbb{E}_{\mathbf{o}'|\boldsymbol{h}, \boldsymbol{a}}\left[V^{\pi}(\boldsymbol{h}')\right]$ with $\boldsymbol{h}' = \boldsymbol{h}\boldsymbol{a}\mathbf{o}'$.

2.2 Informed POMDP

The informed POMDP [14] extends the standard POMDP formulation by introducing a so-called information space $\mathcal{I}$ and a corresponding information function $I : \mathcal{S} \to \Delta(\mathcal{I})$, which specifies the probability of receiving information $\boldsymbol{i}_t \in \mathcal{I}$ given the true state $\boldsymbol{s}_t \in \mathcal{S}$. Accordingly, the informed POMDP is defined by the 10-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{O}, T, I, \widetilde{O}, R, P, \gamma)$. Unlike in a standard POMDP, the observation function is defined as $\widetilde{O} : \mathcal{I} \to \Delta(\mathcal{O})$, which gives the probability to get $\boldsymbol{o}_t \in \mathcal{O}$ given information $\boldsymbol{i}_t \in \mathcal{I}$. The key assumption in an informed POMDP is that the observation $\boldsymbol{o}_t$ is conditionally independent of the true state $\boldsymbol{s}_t$ given the information $\boldsymbol{i}_t$, i.e., $\boldsymbol{o}_t \perp\!\!\!\perp \boldsymbol{s}_t \mid \boldsymbol{i}_t$. For each informed POMDP, there is an underlying execution POMDP defined as $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, P, \gamma)$, where $O(\boldsymbol{o}_t|\boldsymbol{s}_t) = \sum_{\boldsymbol{i} \in \mathcal{I}} \widetilde{O}(\boldsymbol{o}_t|\boldsymbol{i}) I(\boldsymbol{i}|\boldsymbol{s}_t)$.

Assumption 2.1 (Independent trajectories)**.** *Given the initial state distribution P, we assume that independent trajectories $\{\boldsymbol{i}_{k,t}, \boldsymbol{o}_{k,t}, \boldsymbol{a}_{k,t}, r_{k,t} : t \in \mathbb{Z}_{\geq 0}\}$ and histories $\{\boldsymbol{h}_{k,t} : t \in \mathbb{Z}_{\geq 0}\}$ can be obtained, where each trajectory starts from an independent initial state $\boldsymbol{s}_{k,0} \sim P$, and $r_{k,t}$ denotes the reward observed at time step t along trajectory $k \in \mathbb{N}$.*

In the following, we present several reward and value functions that are central to the derivation and analysis of the informed asymmetric actor-critic. We first introduce the time-invariant informed history-based reward function $R(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a})$, which incorporates additional state-conditioned information $\boldsymbol{i} \sim I(\boldsymbol{i}|\boldsymbol{s})$ in the informed POMDP setting.

Definition 2.1 (Informed history-based reward function)**.** *The informed history-based reward function $R(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a})$ is the expected state-based reward $R(\boldsymbol{s}, \boldsymbol{a})$ given the belief $p(\boldsymbol{s}|\boldsymbol{h}, \boldsymbol{i})$ about the true state $\boldsymbol{s} \in \mathcal{S}$, i.e.,*

$$R(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a}) = \mathbb{E}_{\mathbf{s}|\boldsymbol{h}, \boldsymbol{i}}\left[R(\mathbf{s}, \boldsymbol{a})\right]. \tag{3}$$

Note that $R(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a})$ is an unbiased estimate of the expected immediate reward conditioned on history $\boldsymbol{h}$ and action $\boldsymbol{a} \in \mathcal{A}$, i.e., $\mathbb{E}_{\mathbf{i}|\boldsymbol{h}}\left[R(\boldsymbol{h}, \mathbf{i}, \boldsymbol{a})\right] = R(\boldsymbol{h}, \boldsymbol{a})$ (see Lemma D.1 in Appendix D).

We assume that the reward function is uniformly bounded:

Assumption 2.2 (Bounded rewards)**.** *For any $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$, $|R(\boldsymbol{s}, \boldsymbol{a})| \leq r_{\max}$, where $r_{max} \in \mathbb{R}_{>0}$.*

Assumption 2.2 implies that both the standard history-based rewards and, by Lemma D.1, the informed history-based rewards are also bounded by $r_{\max}$.

Next, we define the informed history Q-function, which conditions on $\boldsymbol{h}$, $\boldsymbol{i}$, and $\boldsymbol{a}$.

Definition 2.2 (Informed history Q-function)**.** *The informed history Q-function $Q^{\pi}(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a})$ denotes the expected discounted return when starting from history $\boldsymbol{h} \in \mathcal{H}$, privileged information $\boldsymbol{i} \in \mathcal{I}$, and action $\boldsymbol{a} \in \mathcal{A}$, and then following policy π:*

$$Q^{\pi}(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a}) = \mathbb{E}^{\pi}_{\mathbf{s}_{0:\infty}, \mathbf{a}_{0:\infty} | \boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a}} \left[\sum_{j=0}^{\infty} \gamma^j R(\mathbf{s}_j, \mathbf{a}_j) \right]. \tag{4}$$

Lemma D.2 in Appendix D establishes that the informed history Q-function is an unbiased estimate of the history Q-function. Lastly, we define the time-invariant informed asymmetric value function that evaluates a history $\boldsymbol{h}$ of past observations and actions together with state-conditioned information $\boldsymbol{i}$.

Definition 2.3 (Informed history value function)**.** *The informed history value function $V^{\pi}(\boldsymbol{h}, \boldsymbol{i})$ denotes the expected return starting from history $\boldsymbol{h} \in \mathcal{H}$ and additional information $\boldsymbol{i} \in \mathcal{I}$:*

$$V^{\pi}(\boldsymbol{h}, \boldsymbol{i}) = \mathbb{E}^{\pi}_{\mathbf{s}_{0:\infty}, \mathbf{a}_{0:\infty} | \boldsymbol{h}, \boldsymbol{i}} \left[\sum_{j=0}^{\infty} \gamma^j R(\mathbf{s}_j, \mathbf{a}_j) \right]. \tag{5}$$

It satisfies the recursive form:

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

where the informed Q-function satisfies the Bellman equation:

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

with $\boldsymbol{h}' = \boldsymbol{h}\boldsymbol{a}\mathbf{o}'$.

Unlike the standard history value $V^{\pi}(\boldsymbol{h})$, the informed history value $V^{\pi}(\boldsymbol{h}, \boldsymbol{i})$ incorporates additional state-dependent information, providing richer context on the environment's dynamics and rewards. Compared to the state value $V^{\pi}(\boldsymbol{s})$, $(\boldsymbol{h}, \boldsymbol{i}) \in \mathcal{H} \times \mathcal{I}$ offers a more informative basis for predicting the agent's behavior.

Importantly, the informed history value function $V^{\pi}(\boldsymbol{h}, \boldsymbol{i})$ provides, in expectation, the same signal as the standard history value function, as it is an unbiased estimator of $V^{\pi}(\boldsymbol{h})$; that is $\mathbb{E}_{\mathbf{i}|\boldsymbol{h}}\left[V^{\pi}(\boldsymbol{h}, \mathbf{i})\right] = V^{\pi}(\boldsymbol{h})$ (see Lemma D.3). In the special case where the privileged information $\boldsymbol{i}$ equals the full environment state $\boldsymbol{s} \in \mathcal{S}$, i.e., $\boldsymbol{i} = \boldsymbol{s}$, the informed history value function reduces to the history-state value function $V^{\pi}(\boldsymbol{h}, \boldsymbol{s})$ introduced by Baisero and Amato [18] (see Corollary D.1).

2.3 Actor-critic methods under partial observability

Actor-critic methods are a class of policy gradient algorithms that consist of a policy model (the actor), parameterized by θ, and a value function estimator (the critic), parametrized by ϑ. The actor selects actions according to its policy $\pi_{\theta}(\boldsymbol{a}_t|\boldsymbol{h}_t)$, while the critic evaluates these actions to guide policy improvement. Under partial observability, both models typically condition on the history of past observations and actions, $\boldsymbol{h}_t \in \mathcal{H}$, and are trained using sample-based gradients.

Symmetric actor-critic. In the symmetric actor-critic setting, the policy gradient with respect to the policy parameters θ is given by:

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E}\left[\sum_{t} \gamma^t Q^{\pi}(\boldsymbol{h}_t, \boldsymbol{a}_t) \nabla_{\theta} \log \pi_{\theta}(\boldsymbol{a}_t|\boldsymbol{h}_t) \right]. \tag{8}$$

The critic serves two purposes: it bootstraps value estimates $\hat{V}(\boldsymbol{h}_t; \vartheta)$ and acts as a baseline to reduce variance in the policy gradient estimation. In practice, the history Q-function in Equation 8 is often replaced by the temporal-difference (TD) error $\delta_t = r_t + \gamma \hat{V}(\boldsymbol{h}_{t+1}; \vartheta) - \hat{V}(\boldsymbol{h}_t; \vartheta)$, where r_t denotes the observed reward at time step t, and $\hat{V}(\cdot; \vartheta)$ represents the value estimates from the critic.

Asymmetric actor-critic. Asymmetric actor-critic methods allow the critic to access privileged information, typically the true environment state $\boldsymbol{s}_t$, during training. This information is unavailable to the actor, and during execution. However, prior approaches often rely on state value functions $V^\pi(\boldsymbol{s})$ [17], which are proven to be generally ill-defined for agents operating on histories of past observations and actions [18]. To address this issue, we define the informed asymmetric policy gradient for an informed POMDP based on the unbiased informed history Q-function (cf. Equation 4):

$$\nabla_\theta^{\text{IAAC}} J(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t Q^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t | \boldsymbol{h}_t)\right], \tag{9}$$

where the policy depends on the observable history, and the critic additionally conditions on privileged information.

We find that the informed asymmetric policy gradient is equivalent to the standard policy gradient $\nabla_\theta J(\pi_\theta)$ encountered in symmetric settings.

Theorem 2.1 (Informed asymmetric policy gradient)**.** *Given an informed POMDP, the informed asymmetric policy gradient is equivalent to the standard policy gradient:*

$$\nabla_\theta^{IAAC} J(\pi_\theta) = \nabla_\theta J(\pi_\theta). \tag{10}$$

This implies that an informed asymmetric critic can exploit privileged information during training while preserving the unbiased nature of the policy updates. We provide the proof of Theorem 2.1 in Appendix C.1. Analogous to the informed history value function, the informed asymmetric policy gradient reduces to the asymmetric policy gradient formulation of Baisero and Amato [18] for $\boldsymbol{i}_t = \boldsymbol{s}_t$, with $\boldsymbol{s}_t \in \mathcal{S}$ (see Corollary D.2 in Appendix D).

Building on Theorem 2.1, we define the informed history critic $\hat{V} : \mathcal{H} \times \mathcal{I} \to \mathbb{R}$, which provides an estimate of the informed history value $V^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t)$, given $\boldsymbol{h}_t$ and $\boldsymbol{i}_t$. When combined with a history-dependent policy $\pi_\theta(\boldsymbol{a}_t | \boldsymbol{h}_t)$, this forms an asymmetric actor-critic method, which we refer to as informed asymmetric actor-critic (IAAC). The informed asymmetric policy gradient is approximated as $\hat{\nabla}_\theta^{\text{IAAC}} J(\pi_\theta) = \mathbb{E}\left[\sum_t \gamma^t \, \delta_t \, \nabla_\theta \log \pi(\boldsymbol{a}_t | \boldsymbol{h}_t)\right]$, where the TD errors $\delta_t = r_t + \gamma \, \hat{V}(\boldsymbol{h}_{t+1}, \boldsymbol{i}_{t+1}) - \hat{V}(\boldsymbol{h}_t, \boldsymbol{i}_t)$ are computed using the critic's informed value estimates.

2.4 Infinite-width limit and neural tangent kernel

Let $f(\boldsymbol{x}; \theta)$ denote a parametric function computed by a neural network with input $\boldsymbol{x}$ and parameters θ. In the infinite-width limit, i.e., as the width of each hidden layer $m \to \infty$, the network is well-approximated by its first-order Taylor expansion around the random initialization θ_0:

$$f(\boldsymbol{x}; \theta) \approx f(\boldsymbol{x}; \theta_0) + \langle \nabla_\theta f(\boldsymbol{x}; \theta_0), \theta - \theta_0 \rangle. \tag{11}$$

This linearization defines the time-independent neural tangent kernel (NTK) [31], given by $\kappa(\boldsymbol{x}, \boldsymbol{x}') := \langle \nabla_\theta f(\boldsymbol{x}; \theta_0), \nabla_\theta f(\boldsymbol{x}'; \theta_0) \rangle$, which governs the training dynamics under gradient descent. In the infinite-width regime, the NTK remains fixed during training and converges to a deterministic kernel, independent of the parameter trajectory. Consequently, learning reduces to kernel regression in the reproducing kernel Hilbert space (RKHS) $\mathcal{G}_\kappa$ associated with κ.

The NTK also admits a random feature interpretation, often referred to as the neural tangent random feature (NTRF) model. Defining the feature map $\psi(\boldsymbol{x}; \theta_0) := \nabla_\theta f(\boldsymbol{x}; \theta_0)$, the kernel becomes $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \psi(\boldsymbol{x}; \theta_0), \psi(\boldsymbol{x}'; \theta_0) \rangle$, showing that learning occurs in a feature space induced by the network's initialization θ_0. Functions in this space take the form: $f(\boldsymbol{x}) = \mathbb{E}_{\theta_0}\left[\langle v(\theta_0), \psi(\boldsymbol{x}; \theta_0) \rangle\right]$, for some square-integrable function $v(\cdot)$ over initializations. This representation defines a nonparametric function class:

$$\mathcal{F} := \left\{ \boldsymbol{x} \mapsto \mathbb{E}_{\theta_0}\left[\langle v(\theta_0), \psi(\boldsymbol{x}; \theta_0) \rangle\right] \mid v \in \mathcal{M} \right\}, \tag{12}$$

with RKHS norm $\|f\|_{\mathcal{F}} = \sqrt{\mathbb{E}_{\theta_0}[\|v(\theta_0)\|_2^2]}$. We refer to $v \in \mathcal{M}$ as a transportation mapping. The set of admissible mappings $\mathcal{M}$ controls the complexity of the function class and ensures that the

norm is finite. This perspective abstracts away finite-width parameterizations, reframing training as optimization in an infinite-dimensional feature space derived from the initialization geometry.

3 Informed Asymmetric Rec-NAC Algorithm

In partially observable environments, agents must often rely on the entire sequence of past observations and actions to act optimally [32]. However, explicitly storing and processing unbounded histories is computationally infeasible. Recurrent neural networks (RNNs) offer a practical solution by encoding the history into a fixed-size latent representation, enabling the agent to retain relevant temporal information while maintaining scalability. Due to their ability to model sequential dependencies, RNNs are particularly well-suited for learning non-Markovian policies and value functions within actor-critic frameworks.

In what follows, we introduce the informed asymmetric recurrent natural actor-critic (Rec-NAC) algorithm, which generalizes the recurrent NAC method [27] to settings with privileged information available. This variant employs a recurrent policy and an asymmetric critic that has access to additional input $\boldsymbol{i}_t \sim I(\boldsymbol{i}_t|\boldsymbol{s}_t)$ during training.

3.1 Network architectures

Figure 1 illustrates the architectures of the actor and the informed asymmetric critic neural network. Both networks are implemented as Elman-type recurrent neural networks (RNNs) (see Appendix B.1) that process sequences of observations and actions, with weights shared across time steps to ensure that the hidden state provides a compact encoding of the history $\boldsymbol{h}_t$. Each network is equipped with a task-specific readout head, described in detail below.

Actor. The actor is modeled as an Elman-type RNN $G_t(\cdot;\theta)$ of width $m_G \in \mathbb{N}$, parameterized by $\theta = (\boldsymbol{W}_G \quad \boldsymbol{U}_G)^\top$, where $\boldsymbol{W}_G \in \mathbb{R}^{m_G \times m_G}$ is diagonal and $\boldsymbol{U}_G \in \mathbb{R}^{m_G \times d_x}$ is a general input weight matrix, with d_x denoting the input dimension of the network. The input at each time step is $\boldsymbol{x}_t = (\boldsymbol{o}_t \quad \boldsymbol{a}_t)^\top$, and the resulting hidden state is denoted by $\boldsymbol{y}_t$. A linear projection with fixed weights $\boldsymbol{c} \in \mathbb{R}^{m_G}$ produces a scalar output $G_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta, \boldsymbol{c}) = \frac{1}{\sqrt{m_G}}\langle \boldsymbol{c}, \boldsymbol{y}_t \rangle$. The policy distribution over actions is defined using a softmax over the output logits:

$$\pi_t^\theta(\boldsymbol{a}|\boldsymbol{h}_t) := \frac{\exp\left(G_t(\boldsymbol{h}_t, \boldsymbol{a}; \theta, \boldsymbol{c})\right)}{\sum_{\boldsymbol{a}' \in \mathcal{A}} \exp\left(G_t(\boldsymbol{h}_t, \boldsymbol{a}'; \theta, \boldsymbol{c})\right)}, \quad \boldsymbol{a} \in \mathcal{A},\ \boldsymbol{h}_t \in \mathcal{H}. \tag{13}$$

Critic. The critic estimates the informed history Q-function $Q(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t)$, where $\boldsymbol{i}_t$ denotes privileged information available only during training. It comprises an Elman-type RNN $F_t(\cdot;\vartheta^F)$ of

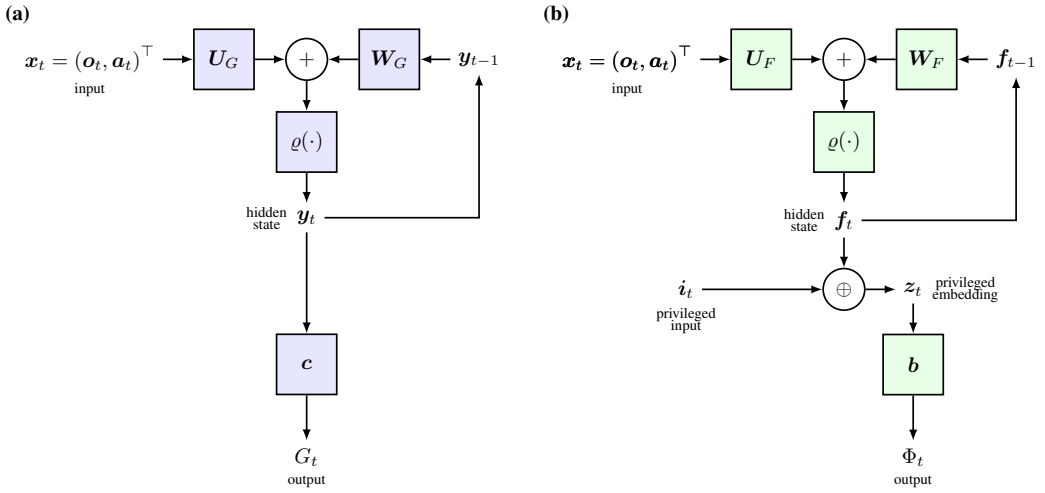

Figure 1: Network architecture of (a) the actor and (b) the informed asymmetric critic.

width m_F, parameterized by $\vartheta^F = (\boldsymbol{W}_F \quad \boldsymbol{U}_F)^\top$, and a single linear layer with learnable parameters $\boldsymbol{b} \in \mathbb{R}^{d_z}$, where $d_z = m_F + d_i$, with d_i denoting the dimensionality of the additional signal $\boldsymbol{i}_t$. The recurrent weights are defined analogously to the actor: $\boldsymbol{W}_G \in \mathbb{R}^{m_G \times m_G}$ is diagonal and $\boldsymbol{U}_G \in \mathbb{R}^{m_G \times d_x}$ is a general input weight matrix. The RNN receives $\boldsymbol{x}_t$ as input, analogous to the actor network, and computes the hidden state $\boldsymbol{f}_t = F_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F)$. This hidden state is then concatenated with the privileged information $\boldsymbol{i}_t$ to form $\boldsymbol{z}_t = \boldsymbol{f}_t \oplus \boldsymbol{i}_t$. The final output is given by a linear projection, i.e., $\phi_t(\boldsymbol{z}_t; \boldsymbol{b}) = \frac{1}{\sqrt{m_F + d_i}} \langle \boldsymbol{b}, \boldsymbol{z}_t \rangle$. The full critic network, parameterized by $\vartheta := (\boldsymbol{W}_F \quad \boldsymbol{U}_F \quad \boldsymbol{b})^\top$, is defined as

$$\Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta) := \phi_t\left(F_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F) \oplus \boldsymbol{i}_t; \boldsymbol{b}\right) = \phi_t(\boldsymbol{f}_t \oplus \boldsymbol{i}_t; \boldsymbol{b}) = \phi_t(\boldsymbol{z}_t; \boldsymbol{b}). \tag{14}$$

3.2 Infinite-width limit of the informed asymmetric critic

In the following, we give the characterization of the infinite-width limit for the informed asymmetric critic. Let $\mathrm{w}_0 \sim \mathrm{Unif}(-\alpha, \alpha), \mathbf{u}_0 \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\mathbf{b}_0 \sim \mathcal{N}(0, \boldsymbol{I}_{d_z})$, be independent random variables, and $\vartheta_0 := (\mathrm{w}_0 \quad \mathbf{u}_0 \quad \mathbf{b}_0)^\top$. The corresponding neural tangent random feature (NTRF) for the informed asymmetric critic at time t is defined as:

$$\boldsymbol{\psi}_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta_0) := \begin{pmatrix} \boldsymbol{\psi}_t^F(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta_0^F) \\ \frac{1}{\sqrt{d_z}} \boldsymbol{z}_t \end{pmatrix} = \begin{pmatrix} \boldsymbol{\psi}_t^F(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta_0^F) \\ \frac{1}{\sqrt{d_z}} (\boldsymbol{f}_t \oplus \boldsymbol{i}_t) \end{pmatrix},$$

where $\boldsymbol{\psi}_t^F$ is given by

$$\boldsymbol{\psi}_t^F(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta_0^F) := \sum_{k=0}^{t} \mathrm{w}_0^k \begin{pmatrix} F_{t-k-1}(\boldsymbol{h}_{t-k-1}, \boldsymbol{a}_{t-k-1}; \vartheta_0^F) \\ \boldsymbol{x}_{t-k} \end{pmatrix} \prod_{j=0}^{k} \Upsilon_{t-j}(\boldsymbol{h}_{t-j}, \boldsymbol{a}_{t-j}; \vartheta^{F_0}),$$

with $F_{-1} := \mathbf{0}$ and $\Upsilon_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta_0^F) = \varrho'\left(\mathrm{w}_0 \boldsymbol{f}_{t-1}(\boldsymbol{h}_{t-1}, \boldsymbol{a}_{t-1}; \vartheta_0^F) + \langle \mathbf{u}_0, \boldsymbol{x}_t \rangle\right)$.

For a sequence $((\boldsymbol{h}_k, \boldsymbol{a}_k, \boldsymbol{i}_k))_{0 \le k \le T-1}$, let the stacked NTRF matrix be

$$\boldsymbol{\Psi}_T(\cdot; \vartheta_0) := \begin{pmatrix} \boldsymbol{\psi}_0^\top(\boldsymbol{h}_0, \boldsymbol{a}_0, \boldsymbol{i}_0; \vartheta_0) \\ \boldsymbol{\psi}_1^\top(\boldsymbol{h}_1, \boldsymbol{a}_1, \boldsymbol{i}_1; \vartheta_0) \\ \vdots \\ \boldsymbol{\psi}_{T-1}^\top(\boldsymbol{h}_{T-1}, \boldsymbol{a}_{T-1}, \boldsymbol{i}_{T-1}; \vartheta_0) \end{pmatrix}, \quad \text{and } \boldsymbol{\Psi}(\cdot; \vartheta_0) := \boldsymbol{\Psi}_\infty(\cdot; \vartheta_0).$$

Following Definition B.2 in Appendix B.5, let the set of transportation mappings $\mathcal{M}_F$ be defined as

$$v : \mathbb{R}^{1+d_x+d_z} \to \mathbb{R}^{1+d_x+d_z}, \quad \vartheta_0 = (\mathrm{w}_0, \mathbf{u}_0, \mathbf{b}_0) \mapsto \begin{pmatrix} v_w(\mathrm{w}_0) \\ v_u(\mathbf{u}_0) \\ v_b(\mathbf{b}_0) \end{pmatrix},$$

subject to $\mathbb{E}\left[|v_w(\mathrm{w}_0)|^2\right] < \infty$, $\mathbb{E}\left[\|v_u(\mathbf{u}_0)\|_2^2\right] < \infty$, and $\mathbb{E}\left[\|v_b(\mathbf{b}_0)\|_2^2\right] < \infty$.

Then, the function class of the recurrent informed asymmetric critic $\Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta)$ is defined by:

$$\mathcal{F} := \left\{ \mathcal{H} \times \mathcal{A} \times \mathcal{I} \ni (\boldsymbol{h}, \boldsymbol{a}, \boldsymbol{i}) \mapsto \mathbb{E}_{\vartheta_0}\left[\boldsymbol{\Psi}_k\left(\boldsymbol{h}, \boldsymbol{a}, \boldsymbol{i}; \vartheta_0\right) v(\vartheta_0)\right] : v \in \mathcal{M}_F \right\}.$$

Accordingly, any target mapping $\Phi_t^\star \in \mathcal{F}$ is given by:

$$\Phi_t^\star(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; v) = \mathbb{E}\left[\langle v(\vartheta_0), \boldsymbol{\psi}_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta_0) \rangle\right].$$

3.3 Recurrent temporal-difference learning

The critic is trained using Recurrent Temporal-Difference learning (Rec-TD) by minimizing

$$\mathcal{R}_T^\pi(\vartheta) := \mathbb{E}_{s_0 \sim P}^\pi \left[\sum_{t=0}^{T-1} \gamma^t \left(\Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta) - Q_t^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) \right)^2 \right], \tag{15}$$

such that $\vartheta \in \Omega_{\boldsymbol{\rho}, (m_F, d_z)}$ (see Appendix B.4).

To perform optimization, Rec-TD employs semi-gradient TD updates based on the empirical TD error:

$$\delta_t(\boldsymbol{h}_{t+1}, \boldsymbol{a}_{t+1}, \boldsymbol{i}_{t+1}; \vartheta) := r_t + \gamma \Phi_{t+1}(\boldsymbol{h}_{t+1}, \boldsymbol{a}_{t+1}, \boldsymbol{i}_{t+1}; \vartheta) - \Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta) \tag{16}$$

leading to the following update direction:

$$\check{\nabla}\mathcal{R}_T(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta) = \sum_{t=0}^{T} \gamma^t \delta_t(\boldsymbol{h}_{t+1}, \boldsymbol{a}_{t+1}, \boldsymbol{i}_{t+1}; \vartheta) \nabla_\vartheta \Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta). \tag{17}$$

Under Assumption 2.1, Rec-TD proceeds with projected gradient descent:

$$\check{\vartheta}_{k+1} = \vartheta_{k+1} + \eta \cdot \check{\nabla}\mathcal{R}_T(\boldsymbol{h}_{k,T}, \boldsymbol{a}_{k,T}, \boldsymbol{i}_{k,T}; \vartheta_k), \tag{18}$$

for $k \in \mathbb{Z}_{\geq 0}$. For Rec-TD with max-norm regularization, the updated parameters are:

$$\vartheta_{k+1} = \texttt{Proj}_{\Omega_{\boldsymbol{\rho},(m_F,d_z)}}[\check{\vartheta}_{k+1}], \tag{19}$$

where $\eta > 0$ is the learning rate, and the projection ensures that the parameter update remains within the admissible norm-constrained set. For a complete overview of the Rec-TD learning loop, please refer to the pseudocode in Algorithm 1 in Appendix F.

3.4 Recurrent natural policy gradient

To approximate the natural gradient of the informed asymmetric policy objective, we adopt a truncated path-based compatible function approximation following [27]. Given a truncation horizon $T \in \mathbb{N}$ and a history-dependent policy $\pi_t^{\theta_n}$, the corresponding output of the critic is defined as $\hat{Q}_t^{\pi_{\theta_n}} := \Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \hat{\vartheta}_n)$, where $\hat{\vartheta}_n$ denotes the critic parameters obtained by Rec-TD at iteration $n \in \mathbb{Z}_{\geq 0}$. The actor then solves the following optimization problem to compute the compatible direction $\boldsymbol{g}$:

$$\min_{\boldsymbol{g}} \ \mathbb{E}\left[\ell_T(\boldsymbol{g}; \theta, \hat{Q})\right] = \left[\sum_{t=0}^{T-1} \gamma^t \left(\nabla \log \pi_t^\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\boldsymbol{g} - \hat{A}_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t)\right)^2\right], \tag{20}$$

subject to $\boldsymbol{g} \in \mathcal{B}_{2,\infty}^{(m_G)}(0, \boldsymbol{\rho}_G)$ (see Appendix B.3), where the advantage estimate is given by $\hat{A}_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t) = \hat{Q}_t(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) - \sum_{\boldsymbol{a} \in \mathcal{A}} \pi_t^\theta(\boldsymbol{a}|\boldsymbol{h}_t)\hat{Q}_t(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a})$.

This problem is solved via projected stochastic gradient descent (SGD). Let $\hat{\boldsymbol{g}}_{n,k}$ denote the estimate after $k \in \mathbb{Z}_{\geq 0}$ SGD steps during the n-th policy update cycle:

$$\tilde{\boldsymbol{g}}_{n,k+1} = \hat{\boldsymbol{g}}_{n,k} - \eta_{\text{sgd}} \nabla_{\boldsymbol{g}} \ell_T(\boldsymbol{g}_{n,k}; \theta, \hat{Q}), \tag{21}$$

$$\hat{\boldsymbol{g}}_{n,k+1} = \texttt{Proj}_{\mathcal{B}_{2,\infty}^{(m_G)}}\left[\tilde{\boldsymbol{g}}_{n,k+1}\right], \tag{22}$$

with initialization $\hat{\boldsymbol{g}}_{n,0} = \boldsymbol{0}$. The policy is then updated using the average compatible direction:

$$\theta_{(n+1)} = \theta^{(n)} + \eta_{\text{npg}} \cdot \frac{1}{K_{\text{sgd}}} \sum_{k < K_{\text{sgd}}} \hat{\boldsymbol{g}}_{n,k}. \tag{23}$$

The complete Rec-NPG procedure is outlined in Algorithm 2 in Appendix F. As an empirical validation, Appendix G presents experimental results evaluating the informed asymmetric Rec-NAC algorithm in an informed POMDP setting.

4 Theoretical Analysis of Informed Asymmetric Rec-NAC

This section provides a theoretical analysis of the informed asymmetric Rec-NAC algorithm. We derive a finite-time and finite-width bound for the informed asymmetric Rec-TD and discuss how incorporating additional state-dependent information into the critic affects the convergence behavior of Rec-NPG. Our results build on the analysis for the symmetric case presented in [27]. We conclude by highlighting the importance of developing quantitative criteria to assess the informativeness of privileged inputs with respect to the underlying true state.

4.1 Finite-time analysis of informed asymmetric Rec-TD

First, we show that informed asymmetric Rec-TD with max-norm regularization achieves global optimality in expectation, under appropriate smoothness and boundedness assumptions. Our analysis extends Theorem E.1 for symmetric Rec-TD [27], to the partially observed setting, where the critic conditions on an augmented input $(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t)$. This setup captures richer input context and accounts for information asymmetry during training.

The following result provides a non-asymptotic upper bound on the expected average Bellman residual over K gradient steps.

Theorem 4.1 (Finite-time bound of informed asymmetric Rec-TD)**.** *Let $\{Q_t^\pi : t \in \mathbb{Z}_{\geq 0}\} \in \mathcal{F}$ with a transportation mapping $v \in \mathcal{M}_F$ such that*

$$\sup_{w\in\mathbb{R}} \|v_w(w)\|_2 \leq \nu_w, \ \sup_{\boldsymbol{u}\in\mathbb{R}^d} \|v_u(\boldsymbol{u})\|_2 \leq \nu_u, \text{ and } \sup_{\boldsymbol{b}\in\mathbb{R}^{m_F+d_i}} \|v_b(\boldsymbol{b})\|_2 \leq \nu_b.$$

Then, for any projection radius $\boldsymbol{\rho} \succeq \boldsymbol{\nu} = \begin{pmatrix}\nu_w & \nu_u & \nu_b\end{pmatrix}^\top$ and step size $\eta_{td} > 0$, Rec-TD with max-norm regularization achieves the following error bound:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1} \mathcal{R}_T^\pi(\vartheta_k)\right] \leq & \frac{\|\boldsymbol{\nu}\|_2^2}{\eta_{td}\,(1-\gamma)\sqrt{K}} + \frac{\eta_{td}\, C_T^{(1)}}{(1-\gamma)^3\sqrt{K}} \\ & + \frac{C_T^{(2)}}{(1-\gamma)^2\sqrt{m_F}} + \frac{\gamma^T}{(1-\gamma)K}\sum_{k=0}^{K-1}\omega_{T,k}^2, \end{aligned} \tag{24}$$

for any $K \in \mathbb{N}$, where $\omega_{t,k} := \sqrt{\mathbb{E}\left[\Phi(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta_k) - Q_t^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t))^2\right]}$ denotes the critic approximation error for $t, k \in \mathbb{Z}_{\geq 0}$, and

$$C_T^{(1)}, C_T^{(2)} = \operatorname{poly}\left(\sum_{k=0}^{T-1}\left|\left(\varrho_1(\alpha + \frac{\boldsymbol{\rho}_w}{\sqrt{m_F}})\right)\right|^k, \|\boldsymbol{\rho}\|_2, \|\boldsymbol{\nu}\|_2\right)$$

are instance-dependent constants.

Theorem 4.1 establishes that, under appropriate smoothness and regularization assumptions, the informed asymmetric Rec-TD algorithm achieves a global error bound composed of (1) an optimization term and (2) a smoothness-related term, both decaying as $\mathcal{O}(1/\sqrt{K})$; (3) a statistical term that decreases with the RNN width m_F; and (4) an accumulation term that captures the impact of bootstrapped targets and scales with the discounted average critic approximation error. The result underscores a trade-off between optimization efficiency, function class complexity, and error propagation induced by temporal difference updates. The proof of Theorem 4.1 can be found in Appendix C.2.

4.2 Theoretical results for Rec-NPG

Next, we briefly investigate the theoretical performance of the Rec-NPG algorithm by analyzing its error properties in the presence of an informed asymmetric critic. Cayci and Eryilmaz [27] establish a non-asymptotic error bound for the best-iterate of Rec-NPG in the symmetric setting (see Theorem E.2 in Appendix E). Their results show that the algorithm's effectiveness is governed by the compatible function approximation error, defined as

$$\varepsilon_{\text{cfa}}^T := \mathbb{E}_{s_0\sim P}^{\pi_{\theta_n}} \sum_{t<T} \gamma^t \left|\nabla^\top \log \pi_t^\theta\left(\boldsymbol{a}_t|\boldsymbol{h}_t\right)\boldsymbol{g} - A_t^{\pi_\theta}\left(\boldsymbol{h}_t, \boldsymbol{a}_t\right)\right|^2.$$

According to Proposition E.2, for any $n \in \mathbb{Z}_+$, this error can be decomposed into (1) $\varepsilon_{\text{app},n}$, capturing the approximation error of the RNN; (2) $\varepsilon_{\text{td},n}$, denoting the statistical error in the critic's temporal-difference (TD) estimate (cf. Equation 15); and (3) $\varepsilon_{\text{sgd},n}$, reflecting the optimization error in the policy update based on the compatible function approximation.

Our primary focus is to understand the effect of using an informed asymmetric critic, instead of a symmetric one, on the actor's error bound, i.e., examining how $\varepsilon_{\text{td},n}$ is influenced. By Theorem 4.1,

the critic's statistical error satisfies

$$\varepsilon_{\mathrm{td},n} \leq \operatorname{poly}\left(\sum_{k=0}^{T-1} |\varrho_1 \alpha_m|^k\right) \mathcal{O}\left(\frac{1}{\sqrt{K_{\mathrm{td}}}} + \frac{1}{\sqrt{m_F}} + \gamma^T\right),$$

when the TD learning rate is set as $\eta_{\mathrm{td}} = \mathcal{O}\left(1/\sqrt{K_{\mathrm{td}}}\right)$. This error bound exactly matches the asymptotic rate achieved in the symmetric Rec-TD setting (see Theorem E.1). Generally, this bound suggests $\varepsilon_{\mathrm{td},n}$ can be made arbitrarily small by increasing the number of TD updates K_{td} and the width of the critic network.

4.3 Discussion

Despite asymptotic equivalence in the Rec-TD error bound, the presence of the additional input in the asymmetric setting influence the constant factors hidden in the bound, that is $C_T^{(1)}$ and $C_T^{(2)}$ in Equation 24. To understand the impact of auxiliary information during critic training, let us compare Theorem 4.1 with the corresponding result for the symmetric setting studied in [27] (cf. Theorem E.1).

One hidden key difference lies in the critic approximation error $\omega_{t,k}^2$, which is minimized over different function classes depending on whether the auxiliary input $\boldsymbol{i}_t$ is available. Let $\widetilde{\mathcal{F}}$ denote the function class of a symmetric critic RNN, parameterized with $\tilde{\vartheta} = \begin{pmatrix}\tilde{\boldsymbol{W}} & \tilde{\boldsymbol{U}}\end{pmatrix}^\top$, followed by a a linear readout with fixed weights. Functions in $\widetilde{F}$ take the form $\tilde{f}^\star(\boldsymbol{h}_t, \boldsymbol{a}_t; \tilde{v}) = \mathbb{E}[\langle \tilde{v}(\tilde{\vartheta}_0), \widetilde{\psi}_t^F(\boldsymbol{h}_t, \boldsymbol{a}_t; \tilde{\vartheta}_0)\rangle]$, where $\tilde{v}(\cdot)$ is the corresponding transportation mapping. Since these functions are independent of $\boldsymbol{i}_t$, we have $\widetilde{\mathcal{F}} \subset \mathcal{F}$. Any $\tilde{f}^\star \in \widetilde{\mathcal{F}}$ can be recovered in the richer class $\mathcal{F}$ by setting $\boldsymbol{i}_t = \boldsymbol{0}$. To contextualize the influence of the informativeness of $\boldsymbol{i}_t$, consider several canonical forms of $\boldsymbol{i}_t$. If $\boldsymbol{i}_t = \boldsymbol{0}$, we do not expect any gain in approximation accuracy compared to the symmetric setting. Conversely, if $\boldsymbol{i}_t$ provides helpful information about the underlying state, we expect the inclusion of $\boldsymbol{i}_t$ to be beneficial with respect to approximation error. For example, let $\boldsymbol{i}_t = e(\boldsymbol{s}_t) + \boldsymbol{\epsilon}_t$ be a noisy embedding of the true state with an injective map e and small noise $\boldsymbol{\epsilon}_t > \boldsymbol{0}$. In the limit $\boldsymbol{\epsilon}_t \to \boldsymbol{0}$, $\boldsymbol{i}_t$ converges to a deterministic encoding of $\boldsymbol{s}_t$, eliminating uncertainty about the true environment state.

However, incorporating informative signals $\boldsymbol{i}_t$ increases the complexity of the function class, requiring more expressive approximators. This typically leads to larger Lipschitz and smoothness constants. In particular, the instance-dependent terms hidden in Equation 24, such as L_T' and $\|\boldsymbol{\nu}\|_2$, are generally larger than their counterparts in the symmetric setting, i.e., $L_T' = L_T^F + \boldsymbol{\rho}_b$ and $\|\boldsymbol{\nu}\|_2 \geq \|\tilde{\boldsymbol{\nu}}\|_2$.

This introduces a fundamental trade-off: while informative additional inputs may improve approximation quality, they can also increase model complexity and variance, as well as slow down convergence in practice. Whether the net effect is provably beneficial likely depends on the informativeness of $\boldsymbol{i}_t$ relative to the increased model capacity required. Intuitively, auxiliary inputs improve training efficiency only if the average gain in approximation outweighs the added complexity cost captured by instance-dependent constants.

To quantify this trade-off, a criterion is needed that ideally balance informativeness with complexity and could guide the selection or construction of $\boldsymbol{i}_t$ in a task-adaptive or data-driven manner. Identifying such conditions remains a key challenge, and addressing it could lead to more effective training algorithms in partially observable environments.

5 Conclusion

This work introduces the informed asymmetric actor-critic paradigm, where the critic leverages privileged inputs during training without requiring full state access. We demonstrate that conditioning on such signals preserves unbiased policy gradients and convergence guarantees, even when with privileged partial information. Our finite-time analysis reveals a trade-off between potentially reduced approximation error and increased model complexity: while privileged inputs may enhance learning, their benefit primarily depends on their informativeness relative to their impact on training stability. A key open challenge is to develop quantitative criteria for assessing and selecting privileged inputs in a task-dependent manner. Future work may also extend the theoretical analysis to neural architectures that non-linearly integrate RNN outputs with privileged information.

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] Pengfei Zhu, Xin Li, Pascal Poupart, and Guanghui Miao. On improving deep reinforcement learning for PODMPs. *arXiv preprint arXiv:1704.07978*, 2017.

[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] Matthew Hausknecht and Peter Stone. Deep recurrent Q-learning for partially observable MDPs. In *2015 AAAI fall symposium series*, 2015.

[10] Daan Wierstra, Alexander Förster, Jan Peters, and Jürgen Schmidhuber. Recurrent policy gradients. *Logic Journal of IGPL*, 18(5):620–634, 2010.

[11] 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.

[12] 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.

[13] 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.

[14] Gaspard Lambrechts, Adrien Bolland, and Damien Ernst. Informed POMDP: Leveraging additional information in model-based RL. *Reinforcement Learning Journal*, 2024.

[15] 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.

[16] Edward S. Hu, James Springer, Oleh Rybkin, and Dinesh Jayaraman. Privileged sensing scaffolds reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] 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.

[18] Andrea Baisero and Christopher Amato. Unbiased asymmetric reinforcement learning under partial observability. In *Proceedings of the Conference on Autonomous Agents and Multiagent Systems*, 2022.

[19] Miguel Vasco, Takuma Seno, Kenta Kawamoto, Kaushik Subramanian, Peter R Wurman, and Peter Stone. A super-human vision-based reinforcement learning agent for autonomous racing in Gran Turismo. *arXiv preprint arXiv:2406.12563*, 2024.

[20] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

[21] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897): 414–419, 2022.

[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] Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[24] Semih Cayci, Niao He, and R. Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[25] Haoxing Tian, Alex Olshevsky, and Ioannis Paschalidis. Convergence of actor-critic with multi-layer neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[26] Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, 2020.

[27] 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.

[28] Gaspard Lambrechts, Damien Ernst, and Aditya Mahajan. A theoretical justification for asymmetric actor-critic algorithms. In *Forty-second International Conference on Machine Learning*, 2025.

[29] Semih Cayci, Niao He, and R. Srikant. Finite-time analysis of natural actor-critic for POMDPs. *SIAM Journal on Mathematics of Data Science*, 6(4):869–896, 2024.

[30] Yang Cai, Xiangyu Liu, Argyris Oikonomou, and Kaiqing Zhang. Provable partially observable reinforcement learning with privileged information. *CoRR*, abs/2412.00985, 2024.

[31] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8580–8589, Red Hook, NY, USA, 2018. Curran Associates Inc.

[32] Satinder P. Singh, Tommi Jaakkola, and Michael I. Jordan. Learning without state-estimation in partially observable markovian decision processes. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 284–292. Morgan Kaufmann, San Francisco (CA), 1994. ISBN 978-1-55860-335-6.

[33] 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.

A Notation

We use calligraphic letters (e.g., $\mathcal{X}$) to denote sets, and $\Delta(\mathcal{X})$ for the set of probability distributions over $\mathcal{X}$. Scalars are written in lowercase (e.g., x), vectors in bold lowercase (e.g., $\boldsymbol{x}$), and matrices in bold uppercase (e.g., $\boldsymbol{X}$). Sequences of vectors are denoted with a subscript range, e.g., $\boldsymbol{s}_{0:n} = (\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$

The i-th element of a vector $\boldsymbol{x}$ is denoted by x_i, the (i,j)-th entry of a matrix $\boldsymbol{X}$ by $X_{i,j}$, and the i-th row of $\boldsymbol{X}$ by $\boldsymbol{X}_{i,:}$. We write $\boldsymbol{I}_n$ for the $n \times n$ identity matrix, and $\mathrm{diag}(\boldsymbol{x})$ for a square diagonal matrix with diagonal entries given by $\boldsymbol{x}$.

We use the standard sets: $\mathbb{R}$ for the real numbers, $\mathbb{N}$ for the natural numbers, and $\mathbb{Z}$ for the integers. The notation $\{0, 1, \ldots, n\}$ denotes the set of integers from 0 to n. Functions are written as $x(\cdot)$ or $X(\cdot)$, and parametric functions as $x(\cdot;\theta)$ or $X(\cdot;\theta)$, where θ denotes the parameters.

Random variables are denoted in sans-serif: x for a scalar-valued random variable, $\mathbf{x}$ for a vector-valued one, and $\mathbf{X}$ for a matrix-valued one. For a random vector $\mathbf{y}$, we write $\mathbb{E}_{\mathbf{y}|\boldsymbol{x}}[f(\mathbf{y})]$ to denote the conditional expectation of $f(\mathbf{y})$ given $\mathbf{x} = \boldsymbol{x}$, i.e., $\mathbb{E}_{\mathbf{y}|\boldsymbol{x}}[f(\mathbf{y})] = \mathbb{E}[f(\mathbf{y}) \mid \mathbf{x} = \boldsymbol{x}]$.

B Algorithmic Tools for the Informed Asymmetric Rec-NAC

In this section, we present key concepts underlying the analysis of the informed asymmetric Rec-NAC method.

B.1 Elman-type Recurrent Neural Network (RNN)

An Elman-type RNN of width $m \in \mathbb{N}$ is parameterized by a recurrent weight matrix $\boldsymbol{W} \in \mathbb{R}^{m \times m}$ with all the off-diagonals set to zero and a general input weight matrix $\boldsymbol{U} \in \mathbb{R}^{m \times d_x}$, where $d_x \in \mathbb{N}$ denotes the input dimension. This structural choice of parameters simplifies the analysis of such a network while retaining essential modeling capabilities of recurrent architectures [27]. Given an input $\boldsymbol{x}_t \in \mathbb{R}^{d_x}$ for $t \in \mathbb{Z}_{\geq 0}$, the hidden state $\boldsymbol{y}_t \in \mathbb{R}^m$ evolves recursively via

$$\boldsymbol{y}_t = \overrightarrow{\varrho}\left(\boldsymbol{W}\boldsymbol{y}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t\right), \quad \boldsymbol{y}_0 = \overrightarrow{\varrho}\left(\boldsymbol{U}\boldsymbol{x}_0\right),$$

where $\overrightarrow{\varrho}(\cdot)$ denotes the element-wise application of a smooth activation function $\varrho \in \mathcal{C}^2(\mathbb{R}, \mathbb{R})$ with bounded derivatives $\|\varrho\|_\infty \leq \varrho_0, \quad \|\varrho'\|_\infty \leq \varrho_1, \quad \|\varrho''\|_\infty \leq \varrho_2$. Notably, the weights are shared across time steps, allowing the hidden state $\boldsymbol{y}_t$ to compactly encode the entire sequence of input with fixed-sized memory.

B.2 Symmetric Random Initialization for Recurrent Neural Networks

In this work, we consider a symmetric variant of random initialization for Elman-type RNNs with linear readout, which ensures certain desirable properties at initialization, such as zero-centered outputs.

Definition B.1 (Symmetric random initialization; Definition A.1 in [27])**.** *Let the RNN's width m be even. For $i \in \{1, \ldots, \frac{m}{2}\}$, draw independently output weights $c_i \sim \mathrm{Unif}(-1, 1)$, recurrent weights $\widetilde{w}_i \sim \mathrm{Unif}(-\alpha, \alpha)$, and input weights $\boldsymbol{U}_{0_{i,:}} \sim \mathcal{N}(0, \boldsymbol{I}_{d_x})$, and set for $i \in \{\frac{m}{2} + 1, \ldots, m\}$:*

$$c_i = -c_{i-\frac{m}{2}},$$
$$\widetilde{w}_i = \widetilde{w}_{i-\frac{m}{2}},$$
$$\boldsymbol{U}_{0_{i,:}} = \boldsymbol{U}_{0_{i-m/2,:}}.$$

The resulting initialization $(\boldsymbol{W}_0 \quad \boldsymbol{U}_0 \quad \boldsymbol{c})^\top$, with $\boldsymbol{W}_0 = \mathrm{diag}_m(\widetilde{\boldsymbol{w}})$, is referred to as symmetric random initialization.

B.3 Max-Norm Regularization for RNNs

Given an Elman-type RNN of width m with a linear output layer with symmetric random initialization $(\boldsymbol{W}_0, \boldsymbol{U}_0, \boldsymbol{c})$, where $\boldsymbol{c}$ denotes the fixed output-layer weights, we consider max-norm regularization

around the random initialization for sharp convergence guarantees. Let $\tilde{\boldsymbol{\rho}} = \begin{pmatrix} \tilde{\rho}_w & \tilde{\rho}_u \end{pmatrix}^\top \in \mathbb{R}^2_{>0}$ denote the projection radii. Then, the compactly-supported set of weights $\Omega_{\tilde{\boldsymbol{\rho}},m} \subset \mathbb{R}^{m(d_x+1)}$ is defined as

$$\Omega_{\tilde{\boldsymbol{\rho}},m} = \mathcal{B}^{(m)}_{2,\infty}\left(\begin{pmatrix} \boldsymbol{W}_0 & \boldsymbol{U}_0 \end{pmatrix}^\top, \tilde{\boldsymbol{\rho}}\right),$$

where $\mathcal{B}^{(d_x)}_p(\boldsymbol{x}, r)$ denotes the closed ℓ_p ball in $\mathbb{R}^d_x$ centered at $\boldsymbol{x}$ with radius τ:

$$\mathcal{B}^{(d)}_p(\boldsymbol{x}, \tau) = \{\boldsymbol{z} \in \mathbb{R}^d : \|\boldsymbol{z} - \boldsymbol{x}\|_p \leq \tau\}$$

Specifically, following [27], we define $\mathcal{B}^{(m)}_{2,\infty}\left(\begin{pmatrix} \boldsymbol{W} & \boldsymbol{U} \end{pmatrix}^\top, \tilde{\boldsymbol{\rho}}\right)$ as

$$\mathcal{B}^{(m)}_{2:\infty}(\begin{pmatrix} \boldsymbol{W} & \boldsymbol{U} \end{pmatrix}^\top, \tilde{\boldsymbol{\rho}}) := \otimes_{i=1}^m \left(\mathcal{B}^{(1)}_1\left(\boldsymbol{W}_{i,:}, \tfrac{\tilde{\rho}_w}{\sqrt{m}}\right), \mathcal{B}^{(d)}_2\left(\boldsymbol{U}_{i,:}, \tfrac{\tilde{\rho}_u}{\sqrt{m}}\right)\right),$$

where $\otimes$ is the Cartesian product.

Hence, for any symmetric random initialization $(\boldsymbol{W}_0, \boldsymbol{U}_0, \boldsymbol{c})$, we have

$$\max_{1\leq i\leq m} |W_{ii} - W_{0_{ii}}| \leq \frac{\tilde{\rho}_w}{\sqrt{m}},$$
$$\max_{1\leq i\leq m} \|\boldsymbol{U}_{i,:} - \boldsymbol{U}_{0_{i,:}}\| \leq \frac{\tilde{\rho}_u}{\sqrt{m}}.$$

We denote the max-norm projection (or regularization) by the projection operator $\widetilde{\texttt{Proj}}_{\Omega_{\tilde{\boldsymbol{\rho}},m}}[\cdot]$, with

$$\widetilde{\texttt{Proj}}_{\Omega_{\tilde{\boldsymbol{\rho}},m}}\left[(\boldsymbol{W}, \boldsymbol{U})^\top\right] = \left[\begin{array}{cc} \underset{w\in\mathcal{B}_2\left(W_{0_{ii}}, \frac{\tilde{\rho}_w}{\sqrt{m}}\right)}{\operatorname{argmin}} |W_{ii} - w_i|, & \underset{u_i\in\mathcal{B}_2\left(\boldsymbol{U}_{0_{i,:}}, \frac{\tilde{\rho}_u}{\sqrt{m}}\right)}{\operatorname{argmin}} \|U_i - u_i\|_2 \end{array}\right]_{i\in\{0,\ldots,m\}}.$$

B.4 Max-Norm Regularization for the Informed Asymmetric Critic

Given an informed asymmetric critic composed of an Elman-type RNN of width m_F, parameterized by $\vartheta^F = \begin{pmatrix} \boldsymbol{W}_F & \boldsymbol{U}_F \end{pmatrix}^\top$, followed by a single linear layer, parameterized by $\boldsymbol{b}$, we define the compactly-supported set of weights $\Omega_{\boldsymbol{\rho},(m_F,d_z)} \subset \mathbb{R}^{m_F(d_x+1)+d_z}$, where d_z denotes the dimension of the linear-layer input. This set is defined relative to the projection radii vector $\boldsymbol{\rho} = \begin{pmatrix} \rho_w & \rho_u & \rho_b \end{pmatrix}^\top \in \mathbb{R}^3_{>0}$, and given by

$$\Omega_{\boldsymbol{\rho},(m_F,d_z)} := \left(\mathcal{B}^{(m)}_{2,\infty}\left(\begin{pmatrix} \boldsymbol{W}_F & \boldsymbol{U}_F \end{pmatrix}^\top, \boldsymbol{\rho}_{\vartheta^F}\right), \mathcal{B}^{(d_z)}_2\left(\boldsymbol{b}, \tfrac{\rho_b}{\sqrt{d_z}}\right)\right).$$

Let $\texttt{Proj}_{\Omega_{\boldsymbol{\rho},(m_F,d_z)}}[\cdot]$ denote the max-norm projection (or regularization), defined as

$$\texttt{Proj}_{\Omega_{\boldsymbol{\rho},(m_F,d_z)}}\left[\begin{pmatrix} \vartheta^F & \boldsymbol{b} \end{pmatrix}^\top\right] = \left[\begin{array}{c} \widetilde{\texttt{Proj}}_{\Omega_{\rho^F_\vartheta,m_F}}\left[\vartheta^F\right], \quad \underset{\mathrm{b}\in\mathcal{B}_2\left(\boldsymbol{b}, \frac{\rho_b}{d_z}\right)}{\operatorname{argmin}} \|\boldsymbol{b} - \mathrm{b}\|_2 \end{array}\right],$$

where $\widetilde{\texttt{Proj}}_{\Omega_{\rho^F_\vartheta,m_F}}[\cdot]$ is the projection operator for the RNN weights, detailed in Section B.3.

B.5 Transportation Mapping for the Informed Asymmetric Critic

Consider an informed asymmetric critic composed of an Elman-type RNN of width m_F, parameterized by $\vartheta^F = \begin{pmatrix} \boldsymbol{W}_F & \boldsymbol{U}_F \end{pmatrix}^\top$ and a single linear layer, parameterized by $\boldsymbol{b}$. Let $\mathrm{w} \sim \mathrm{Unif}(-\alpha, \alpha)$, $\mathbf{u}_0 \sim \mathcal{N}(0, \boldsymbol{I}_{d_x})$, and $\mathbf{b}_0 \sim \mathcal{N}(0, \boldsymbol{I}_{d_z})$ be independent random variables. We define the set of transportation mappings $\mathcal{M}_F$ as:

Definition B.2 (Transportation mapping for the informed asymmetric critic; cf. Definition 4.1 in [27])**.** *Let $\mathcal{M}_F$ be the set of mappings*

$$v : \mathbb{R}^{1+d_x+d_z} \to \mathbb{R}^{1+d_x+d_z}, \quad \vartheta_0 = (\mathrm{w}_0, \mathbf{u}_0, \mathbf{b}_0) \mapsto \begin{pmatrix} v_w(\mathrm{w}_0) \\ v_u(\mathbf{u}_0) \\ v_b(\mathbf{b}_0) \end{pmatrix},$$

subject to

$$\begin{aligned}
\mathbb{E}\big[|v_w(\mathrm{w}_0)|^2\big] &= \frac{1}{2}\left(|v_w(\alpha)|^2 + |v_w(-\alpha)|^2\right) < \infty, \\
\mathbb{E}\big[\|v_u(\mathbf{u}_0)\|_2^2\big] &= \frac{1}{(2\pi)^{d_x/2}} \int_{\mathbb{R}^d} \|v_u(\boldsymbol{u})\|_2^2 \, e^{-\frac{\|\mathbf{u}\|_2^2}{2}} du < \infty, \\
\mathbb{E}\big[\|v_b(\mathbf{b}_0)\|_2^2\big] &= \frac{1}{(2\pi)^{d_z/2}} \int_{\mathbb{R}^{d_z}} \|v_b(\boldsymbol{b})\|_2^2 \, e^{-\frac{\|\mathbf{b}\|_2^2}{2}} db < \infty.
\end{aligned}$$

We refer to each $v \in \mathcal{M}_F$ as a transportation mapping [27].

C Proofs

This section collects our proofs. Section C.1 contains the proof of Theorem 2.1, and Section C.2 that of Theorem 4.1.

C.1 Proof of Theorem 2.1

Proof. Given Equation 8 and following the Lemmas D.2-D.3, we have

$$\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \mathbb{E}\left[\sum_t \gamma^t\, Q^\pi(\mathbf{h}_t, \mathbf{a}_t)\, \nabla_\theta \log \pi(\mathbf{a}_t|\mathbf{h}_t)\right] \\
&\overset{\text{(a)}}{=} \sum_t \gamma^t\, \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t}\left[Q^\pi(\mathbf{h}_t, \mathbf{a}_t)\, \nabla_\theta \log \pi(\mathbf{a}_t|\mathbf{h}_t)\right] \\
&\overset{\text{(b)}}{=} \sum_t \gamma^t\, \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t}[\mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t}\left[Q^\pi(\boldsymbol{h}_t, \mathbf{i}_t, \mathbf{a}_t)\right] \nabla_\theta \log \pi(\mathbf{a}_t|\mathbf{h}_t)] \\
&\overset{\text{(c)}}{=} \sum_t \gamma^t\, \mathbb{E}_{\mathbf{h}_t,\mathbf{i}_t,\mathbf{a}_t}[Q^\pi(\mathbf{h}_t, \mathbf{i}_t, \mathbf{a}_t) \nabla_\theta \log \pi(\mathbf{a}_t|\mathbf{h}_t)] \\
&\overset{\text{(d)}}{=} \mathbb{E}\left[\sum_t \gamma^t\, Q^\pi(\mathbf{h}_t, \mathbf{i}_t, \mathbf{a}_t)\, \nabla_\theta \log \pi(\mathbf{a}_t|\mathbf{h}_t)\right] \\
&= \nabla_\theta^{\text{IAAC}} J(\pi_\theta).
\end{aligned}$$

In (a) and (d), we use the linearity of the expectation to decompose or combine the summation over t and the expectation over $(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t)$, respectively. In (b), using Lemma $D.2$, we substitute $Q^\pi(\mathbf{h}, \mathbf{a})$ with $\mathbb{E}_{\mathbf{i}|\boldsymbol{h}}\left[Q^\pi(\boldsymbol{h}, \mathbf{i}, \mathbf{a})\right]$, as the informed history-action value function is an unbiased estimate of $Q^\pi(\boldsymbol{h}, \boldsymbol{a})$. By applying the law of total expectation in (c), i.e., $\mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t,\mathbf{i}_t}\left[\cdot\right] = \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t}\left[\mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t}[\cdot]\right]$, we can rewrite the expression. This concludes the proof. □

C.2 Proof of Theorem 4.1

Before we prove Theorem 4.1, we derive Lipschitzness and smoothness for the parameters of the informed asymmetric critic. As this critic is composed of an Elman-type RNN and a single-layer perceptron with linear activation, we extend the results given by Lemma E.1 to the composed neural architecture of the informed asymmetric critic.

Let $\Gamma_{t_i}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F) := W_{ii} f_{t_i}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F)$ for any hidden state unit f_{t_i} with $i \in \{0, \dots, m_F - 1\}$ and $\vartheta^F \in \mathbb{R}^{m_F(d_x+1)}$; and let $p_t(\cdot) = \sum_{k=0}^{t-1} |\cdot|^k$ and $q_t(\cdot) = \sum_{k=0}^{t-1}(k+1)|\cdot|^k$.

Lemma C.1 (Local continuity of hidden states in the informed asymmetric critic)**.** *Given $\boldsymbol{\rho} \in \mathbb{R}^3_{>0}$ and $\alpha \geq 0$, let $\alpha_{m_F} = \alpha + \frac{\rho_{w_F}}{\sqrt{m_F}}$. Then, for any $(\boldsymbol{h}, \boldsymbol{a}, \boldsymbol{i}) \in \mathcal{H} \times \mathcal{A} \times \mathcal{I}$ with $\sup_{t\in\mathbb{Z}_{\geq 0}} \|\boldsymbol{x}_t\|_2 \leq 1$, $\sup_{t\in\mathbb{Z}_{\geq 0}} \|\boldsymbol{i}_t\|_2 \leq 1$, $t \in \mathbb{Z}_{\geq 0}$, $j \in \{0, \dots, m_F - 1\}$, and $l \in \{0, \dots, m_F + d_i - 1\}$,*

- $\vartheta^F_{j,:} \mapsto F_{t_j}\left(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F\right)$ *is L_t^F-Lipschitz continuous with $L_t^F = \left(\varrho_0^2 + 1\right)\varrho_0^2 \cdot p_t^2\left(\alpha_{m_F}\varrho_1\right)$,*
- $\vartheta^F_{j,:} \mapsto F_{t_j}\left(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F\right)$ *is β_t^F-smooth with $\beta_t^F = \mathcal{O}\left(d_x \cdot p_t\left(\alpha_{m_F}\varrho_1\right) \cdot q_t\left(\alpha_{m_F}\varrho_1\right)\right)$,*
- $\vartheta^F_{j,:} \mapsto \Gamma_{t_j}\left(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F\right)$ *is Λ_t^F-Lipschitz continuous with $\Lambda_t^F = \sqrt{2}\left(\varrho_0 + 1 + \alpha_{m_F} L_t^F\right)$,*
- $\vartheta^F_{j,:} \mapsto \Gamma_{t_j}\left(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F\right)$ *is χ_t^F-smooth with $\chi_t^F = \sqrt{2}\left(L_t^F + \alpha_{m_F}\beta_t^F\right)$,*
- $b_l \mapsto \Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \boldsymbol{b})$ *is L_t^Φ-Lipschitz continuous with $L_t^\Phi = \|\boldsymbol{b}\|_2$,*
- $b_l \mapsto \Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \boldsymbol{b})$ *is β_t^Φ-smooth with $\beta_t^\Phi = 0$*

in $\Omega_{\boldsymbol{\rho},(m_F, m_F+d_i)}$.

Consequently, for $\mathbb{H}_\infty := \mathcal{H}_\infty \times \mathcal{A} \times \mathcal{I}$, $T \in \mathbb{N}$, $t \in \mathbb{Z}_{\geq 0}$, *and any* $\vartheta \in \Omega_{\boldsymbol{\rho},(m_F, m_F+d_i)}$,

$$\sup_{(\boldsymbol{h},\boldsymbol{a},\boldsymbol{i})\in\mathbb{H}_\infty} \max_{0\leq t\leq T} \left|\Phi_t\left(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta\right)\right| \leq \rho_b \cdot \sqrt{L_T^F \cdot \|\boldsymbol{\rho}_{\vartheta^F}\|_2 + 1},$$

$$\sup_{(\boldsymbol{h},\boldsymbol{a},\boldsymbol{i})\in\mathbb{H}_\infty} \left|\Phi_t^{\mathrm{Lin}}\left(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta\right) - \Phi_t\left(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta\right)\right| \leq \frac{2}{\sqrt{m_F}} \left(\varrho_2 (\Lambda_t^F)^2 + \varrho_1 \chi_t^F\right) \|\vartheta^F - \vartheta_0^F\|_2^2,$$

$$\sup_{(\boldsymbol{h},\boldsymbol{a},\boldsymbol{i})\in\mathbb{H}_\infty} \left\langle \nabla\Phi_t\left(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta\right) - \nabla\Phi_t\left(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta_0\right), \vartheta - \bar{\vartheta}\right\rangle \leq \frac{2(\beta_t^F)^2 \|\boldsymbol{\rho}_{\vartheta^F}\|_2^2}{\sqrt{m_F}},$$

with probability 1 over the symmetric random parameter initialization $(\mathbf{W}_0^F, \mathbf{U}_0^F, \mathbf{b}_0)^\top$.

Furthermore, we extend the external result outlined in Lemma E.2 to the composed neural architecture of the informed asymmetric critic.

Lemma C.2 (Approximation error between NTRF and NTK in the informed asymmetric critic)**.** *Let* $\phi^\star \in \mathcal{F}$ *with the transportation mapping* $v \in \mathcal{M}_F$, *and let*

$$\bar{\vartheta}_{j,:} = \vartheta_{0_{j,:}} + \frac{1}{\sqrt{m_F}} \zeta_j v\left(\vartheta_{0_{j,:}}\right), \quad j \in \{1, \ldots, d + m_F + d_i - 1\}, \tag{25}$$

where ζ_j *are i.i.d. symmetric random variables independent of* ϑ_0. *Let*

$$\Phi_t^{\mathrm{Lin}}(\cdot; \vartheta) := \nabla_\vartheta \Phi_t(F_t^{\mathrm{Lin}}(\cdot; \vartheta_0^F), \cdot; \boldsymbol{b}_0) \cdot (\vartheta - \vartheta_0). \tag{26}$$

If $\mathbb{P}_T^{\pi,P}$ *induces a compactly-supported marginal distribution for* $\boldsymbol{x}_t$ *and* $\boldsymbol{i}_t$, $t \in \mathbb{Z}_{\geq 0}$ *such that* $\|\boldsymbol{x}_t\|_2 \leq 1$ *and* $\|\boldsymbol{i}_t\|_2 \leq 1$ *almost surely, and* $\{(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t) : t \in \mathbb{Z}_{\geq 0}\}$ *is independent from the random initialization* ϑ_0, *then we have*

$$\mathbb{E}_{\vartheta_0}\left[\mathbb{E}_P^\pi\left[\left(\phi_t^\star\left(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t\right) - \Phi_t^{\mathrm{Lin}}\left(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \bar{\vartheta}\right)\right)^2\right]\right] \leq \frac{2\|\boldsymbol{\nu}\|_2^2 (1 + \varrho_0^2) p_t^2(\alpha\varrho_1)}{m_F}. \tag{27}$$

We extend the non-stationary Bellman equation (cf. Proposition E.1) to the informed asymmetric setting:

Proposition C.1 (Non-stationary Bellman equation in the informed asymmetric setting)**.** *For* $\pi \in \Pi$, *we have*

$$\begin{aligned} Q_t^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) &= \mathbb{E}^\pi_{\mathbf{s}_t, \boldsymbol{h}_{t+1}, \mathbf{i}_{t+1}, \mathbf{a}_{t+1} | \boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t}\left[R(\mathbf{s}_t, \boldsymbol{a}_t) + \gamma Q_{t+1}^\pi(\mathbf{h}_{t+1}, \mathbf{i}_{t+1}, \mathbf{a}_{t+1})\right] \\ &= \mathbb{E}^\pi_{\mathbf{s}_t, \mathbf{h}_{t+1}, \mathbf{i}_{t+1} | \boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t}\left[R(\mathbf{s}_t, \boldsymbol{a}_t) + \gamma V_{t+1}^\pi(\mathbf{h}_{t+1}, \mathbf{i}_{t+1})\right], \end{aligned}$$

for any $(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) \in \mathcal{H} \times \mathcal{I} \times \mathcal{A}$ *and* $t \in \mathbb{Z}_{\geq 0}$.

Proof of Theorem 4.1. The following proof closely follows the structure of the proof of Theorem 6.3 in [27] (cf. Theorem E.1), with adaptations tailored to informed asymmetric critic design presented in this work.

Proof. Given that $\{Q_t^\pi : t \in \mathbb{Z}_{\geq 0}\} \in \mathcal{F}$, let $\bar{\vartheta}$ denote the point of attraction. Then, the potential function is given as:

$$\Psi(\vartheta) = \|\vartheta - \bar{\vartheta}\|_2^2. \tag{28}$$

By leveraging the non-expansivity of the projection operator onto the convex set $\Omega_{\boldsymbol{\rho},(m_F, m_F+d_i)}$, we derive:

$$\begin{aligned} \Psi(\vartheta_{k+1}) \leq \Psi(\vartheta_k) &+ 2\eta \sum_{t=0}^{T-1} \gamma^t \delta_t\left(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k\right) \left\langle \nabla\Phi_t\left(F_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k; \vartheta_k^{\mathrm{F}}), i_t^k; \boldsymbol{b}_k\right), \vartheta_k - \bar{\vartheta}\right\rangle \\ &+ 2\eta^2 \left\|\mathcal{R}_T\left(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k\right)\right\|_2^2. \end{aligned} \tag{29}$$

To streamline notation, we define $\check{\mathbb{E}}_t^k[\cdot] := \mathbb{E}[\cdot|\vartheta_k, \ldots, \vartheta_0, \boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k]$. We obtain

$$\mathbb{E}[\Psi(\vartheta_{k+1}) - \Psi(\vartheta_k)] \leq 2\eta\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \underbrace{\check{\mathbb{E}}_t^k\left[\delta_t\left(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k\right)\right] \cdot \left\langle \nabla\Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k),\ \vartheta_k - \bar{\vartheta}\right\rangle}_{(\boldsymbol{\Delta})_t}\right] + \eta^2 \underbrace{\mathbb{E}\left[\left\|\check{\nabla}\mathcal{R}_T\left(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k\right)\right\|_2^2\right]}_{(\heartsuit)} \tag{30}$$

To bound $\mathbb{E}(\boldsymbol{\Delta})_t$, we first apply the non-stationary Bellman equation (cf. Proposition C.1) and adapt it to the informed asymmetric setting:

$$\begin{aligned}\check{\mathbb{E}}_t^k\left[\delta_t\left(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k\right)\right] &= \check{\mathbb{E}}_t^k\left[r_t^k + \gamma \cdot \Phi_{t+1}(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, i_{t+1}^k; \vartheta_k)\right] - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k) \\ &= \gamma \cdot \check{\mathbb{E}}_t^k\left[\Phi_{t+1}(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, i_{t+1}^k; \vartheta_k) - Q_{t+1}^{\pi_\theta}\left(h_{t+1}^k, i_{t+1}^k, a_{t+1}^k\right)\right] \\ &\quad + Q_t^{\pi_\theta}\left(h_t^k, i_t^k, \boldsymbol{a}_t^k\right) - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k).\end{aligned} \tag{31}$$

We can reformulate the inner product using reparameterized features:

$$\left\langle \nabla\Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k), \vartheta_k - \bar{\vartheta}\right\rangle = \left\langle \nabla\Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_0), \vartheta_k - \bar{\vartheta}\right\rangle + \text{err}_{t,k}^{(1)}, \tag{32}$$

where the residual term is

$$\text{err}_{t,k}^{(1)} := \left\langle \nabla\Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k) - \nabla\Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_0), \vartheta_k - \bar{\vartheta}\right\rangle.$$

Using Lemma C.1, this error satisfies:

$$\left|\text{err}_{t,k}^{(1)}\right| \leq \frac{2(\beta_t^F)^2\|\boldsymbol{\rho}_{\vartheta^F}\|_2^2}{\sqrt{m_F}} \leq \frac{2(\beta_T^F)^2\|\boldsymbol{\rho}_{\vartheta^F}\|_2^2}{\sqrt{m_F}}.$$

We can further decompose the inner product as follows:

$$\left\langle \nabla\Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_0),\ \vartheta_k - \bar{\vartheta}\right\rangle = \Phi_t^{\text{Lin}}(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k) - \Phi_t^{\text{Lin}}(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \bar{\vartheta}) \tag{33}$$

$$= \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k) - Q_t^{\pi_\theta}(\boldsymbol{h}_t^k, \boldsymbol{i}_t^k, \boldsymbol{a}_t^k) + \text{err}_{t,k}^{(2)} + \text{err}_{t,k}^{(3)}. \tag{34}$$

The error terms are:

$$\text{err}_{t,k}^{(2)} := \Phi_t^{\text{Lin}}(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k) - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k), \tag{35}$$

$$\text{err}_{t,k}^{(3)} := -\Phi_t^{\text{Lin}}(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \bar{\vartheta}) + Q_t^{\pi_\theta}(\boldsymbol{h}_t^k, \boldsymbol{i}_t^k, \boldsymbol{a}_t^k). \tag{36}$$

Substituting into $(\boldsymbol{\Delta})_t$ yields:

$$\begin{aligned}(\boldsymbol{\Delta})_t = &- \left(Q_t^{\pi_\theta}(\boldsymbol{h}_t^k, \boldsymbol{i}_t^k, \boldsymbol{a}_t^k) - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k)\right)^2 \\ &+ \gamma \cdot \check{\mathbb{E}}_t^k\left[\Phi_{t+1}(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k) - Q_{t+1}^{\pi_\theta}(\boldsymbol{h}_{t+1}^k, \boldsymbol{i}_{t+1}^k, \boldsymbol{a}_{t+1}^k)\right] \\ &\quad \cdot \left(Q_t^{\pi_\theta}(\boldsymbol{h}_t^k, \boldsymbol{i}_t^k, \boldsymbol{a}_t^k) - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k)\right) \\ &+ \check{\mathbb{E}}_t^k\left[\delta_t(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k)\right] \cdot \left(\text{err}_{t,k}^{(1)} + \text{err}_{t,k}^{(2)} + \text{err}_{t,k}^{(3)}\right)\end{aligned} \tag{37}$$

By Lemma C.1, we bound the temporal differences by:

$$\sup_{\boldsymbol{h},\boldsymbol{a},\boldsymbol{i}\in\mathbb{H}_\infty} |\delta_t\left(\boldsymbol{h}_{t+1}, \boldsymbol{a}_{t+1}, \boldsymbol{i}_{t+1}; \vartheta_k\right)| \leq r_{\max} + 2\rho_b \cdot \sqrt{L_T^F \cdot \|\boldsymbol{\rho}_{\vartheta^F}\|_2 + 1} =: \delta_{\max}. \tag{38}$$

Let $\omega_{t,k} := \left(\mathbb{E}_{\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k, \vartheta_k}\left[\left(Q_t^{\pi_\theta}(\boldsymbol{h}_t^k, \boldsymbol{i}_t^k, \boldsymbol{a}_t^k) - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k)\right)^2\right]\right)^{1/2}$, so:

$$\mathbb{E}\left[(\boldsymbol{\Delta})_t\right] \leq -\omega_{t,k}^2 + \gamma\omega_{t+1,k}\,\omega_{t,k} + \delta_{\max}\sum_{j=1}^{3}\mathbb{E}\left|\text{err}_{t,k}^{(j)}\right|. \tag{39}$$

Applying bounds from Lemmas C.1 and C.2, we obtain:

$$\mathbb{E}\left|\mathrm{err}_{t,k}^{(2)}\right| \leq \frac{2}{\sqrt{m_F}}\left(\varrho_2(\Lambda_t^F)^2 + \varrho_1\chi_t^F\right)\|\boldsymbol{\rho}_{\vartheta^F}\|_2^2 \tag{40}$$

and

$$\mathbb{E}\left|\mathrm{err}_{t,k}^{(3)}\right| \leq \sqrt{\mathbb{E}\left|\mathrm{err}_{t,k}^{(3)}\right|^2} \leq \frac{2\|\boldsymbol{\nu}\|_2\sqrt{1+\varrho_0^2}\cdot p_T(\alpha\varrho_1)}{\sqrt{m_F}}. \tag{41}$$

We can bound the product $\omega_{t+1,k}\,\omega_{t,k}$ using the arithmetic mean inequality:

$$\omega_{t+1,k}\,\omega_{t,k} \leq \frac{1}{2}\left(\omega_{t,k}^2 + \omega_{t+1,k}^2\right).$$

Using this inequality, we can derive the following bound for any time step $t \in \{0, 1, \ldots, T-1\}$:

$$\mathbb{E}\left[(\boldsymbol{\Delta})_t\right] \leq -\omega_{t,k}^2 + \frac{\gamma}{2}\left(\omega_{t+1,k}^2 + \omega_{t,k}^2\right) + \delta_{\max}\cdot\frac{\mathcal{C}_T}{\sqrt{m_F}} \tag{42}$$

where $\mathcal{C}_T$ is defined as

$$\mathcal{C}_T := 2(\beta_T^F)^2\|\boldsymbol{\rho}_{\vartheta^F}\|_2^2 + \left(\varrho_2(\Lambda_t^F)^2 + \varrho_1\chi_t^F\right)\|\boldsymbol{\rho}_{\vartheta^F}\|_2^2 + 2\|\boldsymbol{\nu}\|_2\sqrt{1+\varrho_0^2}\cdot p_T(\alpha\varrho_1). \tag{43}$$

Hence, we obtain the following upper bound:

$$\begin{aligned}\sum_{t=0}^{T-1}\gamma^t\mathbb{E}\left[(\boldsymbol{\Delta})_t\right] &\leq -(1-\frac{\gamma}{2})\sum_{t<T}\gamma^t\omega_{t,k}^2 + \frac{\delta_{\max}\cdot\mathcal{C}_T}{(1-\gamma)\sqrt{m_F}} + \underbrace{\frac{1}{2}\sum_{t<T}\gamma^{t+1}\omega_{t+1,k}^2}_{\leq\frac{1}{2}\left(\sum_{t<T}\gamma^t\omega_{t,k}^2+\gamma^T\omega_{T,k}^2\right)} \\ &\leq -\frac{1-\gamma}{2}\sum_{t<T}\gamma^t\omega_{t,k}^2 + \frac{1}{2}\gamma^T\omega_{T,k}^2 + \frac{\mathcal{C}_T\cdot\delta_{\max}}{(1-\gamma)\sqrt{m_F}}.\end{aligned} \tag{44}$$

We now derive an upper bound on the term $\mathbb{E}[(\heartsuit)]$. First, using the triangle inequality, we have:

$$\begin{aligned}&\left\|\sum_{t<T}\gamma^t\delta_t\left(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k\right)\nabla\Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k)\right\|_2 \\ &\leq \sum_{t<T}\gamma^t\left|\delta_t\left(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k\right)\right|\cdot\left\|\nabla\Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k)\right\|_2\end{aligned} \tag{45}$$

Since ϑ_k remains within the bounded set $\Omega_{\boldsymbol{\rho},(m_F,m_F+d_i)}$ for every $k \in \mathbb{Z}_{\geq 0}$ due to the max-norm regularization, we can ensure

$$\left|\delta_t\left(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k\right)\right| \leq \delta_{\max} = r_{\max} + 2\rho_b\cdot\sqrt{L_T^F\cdot\|\boldsymbol{\rho}_{\vartheta^F}\|_2 + 1},$$

$$\left\|\nabla\Phi_t\left(F_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k; \vartheta_k^{\mathrm{F}}), \boldsymbol{i}_t^k; \boldsymbol{b}_k\right)\right\|_2^2 \leq \left(L_T'\right)^2,$$

where $L_T' = L_T^F + \rho_b$ for every $t < T$ with probability 1, according to Lemma C.1.

Therefore, we obtain the upper bound:

$$\left\|\check{\nabla}\mathcal{R}_T\left(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k\right)\right\|_2 \leq \frac{\delta_{\max}L_T'}{1-\gamma}. \tag{46}$$

Taking the expectation over the stochastic components $\left(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k, \vartheta_k\right)$ in the update rule (cf. Equation 30), and substituting in the bounds obtained above (cf. Equation 44 and Equation 46), yields:

$$\mathbb{E}[\Psi(\vartheta_{k+1}) - \Psi(\vartheta_k)] \leq -\eta(1-\gamma)\sum_{t=0}^{T-1}\gamma^t\omega_{t,k}^2 + \eta\gamma^T\omega_{T,k}^2 \tag{47}$$

$$+ \eta\frac{\mathcal{C}_T\delta_{\max}}{(1-\gamma)\sqrt{m_F}} + \eta^2\frac{\delta_{\max}^2\left(L_T'\right)^2}{(1-\gamma)^2}, \tag{48}$$

for every $k \in \mathbb{Z}_{\geq 0}$. Since $\Psi(\vartheta_0) \leq \|\boldsymbol{\nu}\|_2^2$, applying a telescoping sum over $k = 0, 1, \ldots, K-1$, as in the proof of the symmetric Rec-TD bound in [27], results in:

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\mathcal{R}_T(\vartheta_k)\right] \leq \frac{\|\boldsymbol{\nu}\|_2^2}{\eta(1-\gamma)K} + \frac{\eta\delta_{\max}^2\,(L'_T)^2}{(1-\gamma)^3} + \frac{\mathcal{C}_T\delta_{\max}}{(1-\gamma)^2\sqrt{m_F}} + \frac{\gamma^T}{(1-\gamma)K}\sum_{k=0}^{K-1}\omega_{T,k}^2. \tag{49}$$

This concludes the proof. □

D Auxillary Results

This section collects our auxiliary results.

Lemma D.1 (Unbiasedness of the informed history-based reward)**.** *In an informed POMDP, the informed history-based reward function $R(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a})$ satisfies*

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

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

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

$$R(\boldsymbol{h}, \boldsymbol{a}) = \mathbb{E}_{\mathbf{s}|\boldsymbol{h}}\big[R(\mathbf{s}, \boldsymbol{a})\big] = \sum_{\boldsymbol{s}\in\mathcal{S}} R(\boldsymbol{s}, \boldsymbol{a})\, p(\boldsymbol{s}|\boldsymbol{h}), \tag{50}$$

and applying the law of total probability, we obtain:

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

This concludes the proof. □

Lemma D.2 (Unbiasedness of the informed Q-function)**.** *In an informed POMDP, the informed history Q-function satisfies*

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

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

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

$$\begin{aligned}
Q^{\pi}(\boldsymbol{h}, \boldsymbol{a}) &= \mathbb{E}^{\pi}_{\mathbf{s}_{0:\infty}, \mathbf{a}_{0:\infty}|\boldsymbol{h}, \boldsymbol{a}} \Big[\sum_{j=0}^{\infty} \gamma^j\, R\big(\mathbf{s}_j, \mathbf{a}_j\big) \Big] \\
&= \mathbb{E}_{\mathbf{i}|\boldsymbol{h}} \Big[\mathbb{E}^{\pi}_{\mathbf{s}_{0:\infty}, \mathbf{a}_{0:\infty}|\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a}} \Big[\sum_{j=0}^{\infty} \gamma^j\, R\big(\mathbf{s}_j, \mathbf{a}_j\big) \Big] \Big] \\
&= \mathbb{E}_{\mathbf{i}|\boldsymbol{h}} \Big[Q^{\pi}(\boldsymbol{h}, \mathbf{i}, \boldsymbol{a}) \Big].
\end{aligned}$$

This concludes the proof. □

Lemma D.3 (Unbiasedness of the informed value function)**.** *In an informed POMDP, the informed value function satisfies for all $\boldsymbol{h} \in \mathcal{H}$:*

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

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

$$V^{\pi}(\boldsymbol{h}) = \mathbb{E}^{\pi}_{\mathbf{s}_{0:\infty},\, \mathbf{a}_{0:\infty}|\boldsymbol{h}} \left[\sum_{j=0}^{\infty} \gamma^j R(\mathbf{s}_j, \mathbf{a}_j) \right],$$

and using the law of total expectation, we have:

$$
\begin{aligned}
V^{\pi}(\boldsymbol{h}) &= \mathbb{E}^{\pi}_{\mathbf{s}_{0:\infty},\mathbf{a}_{0:\infty}|\boldsymbol{h}}\Big[\sum_{j}^{\infty}\gamma^{j}\, R\big(\mathbf{s}_j,\mathbf{a}_j\big)\Big] \\
&= \mathbb{E}_{\mathbf{i}|\boldsymbol{h}}\Big[\mathbb{E}^{\pi}_{\mathbf{s}_{0:\infty},\mathbf{a}_{0:\infty}|\boldsymbol{h},\boldsymbol{i}}\Big[\sum_{j}^{\infty}\gamma^{j}\, R\big(\mathbf{s}_j,\mathbf{a}_j\big)\Big]\Big] \\
&= \mathbb{E}_{\mathbf{i}|\boldsymbol{h}}\Big[V^{\pi}(\boldsymbol{h},\mathbf{i})\Big].
\end{aligned}
$$

This concludes the proof. □

Corollary D.1 (Relation of $V^{\pi}(h,i)$ to the history-state value function of Baisero and Amato [18])**.** *The informed history value function $V^{\pi}(\boldsymbol{h},\boldsymbol{i})$ reduces to the history-state value function for $\boldsymbol{i}=\boldsymbol{s}$, where $\boldsymbol{s}\in\mathcal{S}$ denotes the true environment state. In particular,*

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

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

$$
Q^{\pi}(\boldsymbol{h},\boldsymbol{s},\boldsymbol{a}) = R(\boldsymbol{s},\boldsymbol{a}) + \gamma\, \mathbb{E}_{\mathbf{s}',\mathbf{o}'|\boldsymbol{s},\boldsymbol{a}}\left[V^{\pi}(\boldsymbol{h}',\mathbf{s}')\right],
$$

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

By Lemma D.3, this formulation provides an alternative unbiased estimator of the history value function:

$$
V^{\pi}(\boldsymbol{h}) = \mathbb{E}_{\mathbf{s}|\boldsymbol{h}}\left[V^{\pi}(\boldsymbol{h},\mathbf{s})\right],
$$

as previously established by Baisero and Amato [18].

Corollary D.2 (Relation of $\nabla^{\text{IAAC}}_{\theta}J(\pi_\theta)$ to the asymmetric policy gradient of Baisero and Amato [18])**.** *The informed asymmetric policy gradient $\nabla^{IAAC}_{\theta}J(\pi_\theta)$ reduces to the asymmetric policy gradient introduced by Baisero and Amato [18] for $\boldsymbol{i}=\boldsymbol{s}$, where $\boldsymbol{s}\in\mathcal{S}$ denotes the true environment state. In particular,*

$$
\nabla^{AAC}_{\theta}J(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^{t}\, Q^{\pi}(\boldsymbol{h}_t,\boldsymbol{s}_t,\boldsymbol{a}_t)\,\nabla_{\theta}\log\pi_\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\right].
$$

Following Lemma D.1-D.3, this formulation recovers an alternative asymmetric policy gradient estimator that is equivalent to the standard policy gradient:

$$
\nabla^{AAC}_{\theta}J(\pi_\theta) = \nabla_{\theta}J(\pi_\theta),
$$

as established by Baisero and Amato [18].

E External Results

This section provides a summary of external results necessary to prove our lemmas and theorems. Section E.1 presents the Lipschitz-continuity and smoothness results for the hidden states in Elman-type RNNs. Section E.2 details the approximation error between NTRF and NTK in Elman-type RNNs, and Section E.1 introduces the non-stationary Bellman equation. Section E.4 contains the main results of the theoretical analysis of symmetric Rec-NAC. Section E.5 describes the decomposition of the compatible function approximation error in symmetric Rec-NAC.

E.1 Lipschitzness and smoothness of hidden states in Elman-type RNNs

Let $\widetilde{\Gamma}_{t_i}(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta) := W_{ii}\boldsymbol{y}_{t_i}(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)$ for any hidden state unit $\boldsymbol{y}_{t_i}$ with $i \in \{0, \ldots, m-1\}$ and $\theta \in \mathbb{R}^{m(d_x+1)}$. Then, we have the following Lipschitz-continuity and smoothness results for $\theta_i \mapsto \boldsymbol{y}_{t_i}(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)$ and $\theta_i \mapsto \widetilde{\Gamma} t_i(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)$.

Lemma E.1 (Local continuity of hidden states; Lemma B.1 in [27])**.** *Given $\tilde{\boldsymbol{\rho}} \in \mathbb{R}^2_{>0}$ and $\alpha \geq 0$, let $\alpha_m = \alpha + \frac{\tilde{\rho}_w}{\sqrt{m}}$. Then, for any $(\boldsymbol{h}, \boldsymbol{a}) \in \mathcal{H} \times \mathcal{A}$ with $\sup_{t \in \mathbb{Z}_{\geq 0}} \|\boldsymbol{x}_t\|_2 \leq 1$, $t \in \mathbb{Z}_{\geq 0}$ and $i \in \{0, \ldots, m-1\}$,*

- *$\theta_i \mapsto \boldsymbol{y}_{t_i}(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)$ is L_t-Lipschitz continuous with $L_t = \left(\varrho_0^2 + 1\right)\varrho_0^2 \cdot p_t^2(\alpha_m \varrho_1)$,*
- *$\theta_i \mapsto \boldsymbol{y}_{t_i}(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)$ is β_t-smooth with $\beta_t = \mathcal{O}\left(d_x \cdot p_t(\alpha_m \varrho_1) \cdot q_t(\alpha_m \varrho_1)\right)$,*
- *$\theta_i \mapsto \widetilde{\Gamma}_{t_i}(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)$ is Λ_t-Lipschitz continuous with $\Lambda_t = \sqrt{2}\left(\varrho_0 + 1 + \alpha_m L_t\right)$,*
- *$\theta_i \mapsto \widetilde{\Gamma}_{t_i}(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)$ is χ_t-smooth with $\chi_t = \sqrt{2}\left(L_t + \alpha_m \beta_t\right)$,*

in $\Omega_{\tilde{\boldsymbol{\rho}}, m}$. Consequently, for $\widetilde{\mathbb{H}}_\infty := \mathcal{H}_\infty \times \mathcal{A}$ and any $\theta \in \Omega_{\tilde{\boldsymbol{\rho}}, m}$,

$$\sup_{(\boldsymbol{h}, \boldsymbol{a}) \in \widetilde{\mathbb{H}}_\infty} \max_{0 \leq t \leq T} \left|G_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)\right| \leq L_T \cdot \|\tilde{\boldsymbol{\rho}}\|, T \in \mathbb{N}, \tag{51}$$

$$\sup_{(\boldsymbol{h}, \boldsymbol{a}) \in \widetilde{\mathbb{H}}_\infty} \left|G_t^{\mathrm{Lin}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta) - G_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)\right| \leq \frac{2}{\sqrt{m}}\left(\varrho_2 \Lambda_t^2 + \varrho_1 \chi_t\right)\|\theta - \theta_0\|_2^2, t \in \mathbb{Z}_{\geq 0}, \tag{52}$$

$$\sup_{(\boldsymbol{h}, \boldsymbol{a}) \in \widetilde{\mathbb{H}}_\infty} \left\langle \nabla G_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta) - \nabla G_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0), \theta - \bar{\theta} \right\rangle \leq \frac{2\beta_t^2 \|\tilde{\boldsymbol{\rho}}\|_2^2}{\sqrt{m}}, \tag{53}$$

with probability 1 over the symmetric random parameter initialization $(\mathbf{W}_0, \mathbf{U}_0, \mathbf{c})^\top$.

E.2 Approximation error between NTRF and NTK in Elman-type RNNs

The following lemma provides an upper bound on the approximation error between the neural tangent random feature (NTRF) and the neural tangent kernel (NTK) in Elman-type RNNs.

Lemma E.2 (Approximation error between RNN-NTRF and RNN-NTK; Lemma B.2 in [27])**.** *Let $g^\star \in \mathcal{G}$ with the transportation mapping $\tilde{v} \in \mathcal{M}_G$, and let*

$$\bar{\theta}_i = \theta_{0_i} + \frac{1}{\sqrt{m}} c_i \tilde{v}(\theta_{0_i}), \quad i \in \{1, \ldots, m-1\}, \tag{54}$$

for any symmetric random parameter initialization $\theta_0 = (\boldsymbol{W}_0, \boldsymbol{U}_0, \boldsymbol{c})^\top$ (cf. Defintion B.1). Let

$$G_t^{\mathrm{Lin}}(\cdot; \theta) = \nabla_\theta G_t(\cdot; \theta_0) \cdot (\theta - \theta_0). \tag{55}$$

If $\mathbb{P}_T^{\pi, P}$ induces a compactly-supported marginal distribution for $\boldsymbol{x}_t, t \in \mathbb{Z}_{\geq 0}$ such that $\|\boldsymbol{x}_t\|_2 \leq 1$ almost surely and $\{(\boldsymbol{h}_t, \boldsymbol{a}_t) : t \in \mathbb{Z}_{\geq 0}\}$ is independent from the random initialization θ_0, then we have

$$\mathbb{E}_{\theta_0}\left[\mathbb{E}_P^\pi\left[\left(g_t^\star(\boldsymbol{h}_t, \boldsymbol{a}_t) - G_t^{\mathrm{Lin}}\left(\boldsymbol{h}_t, \boldsymbol{a}_t; \bar{\theta}\right)\right)^2\right]\right] \leq \frac{2\|\tilde{\boldsymbol{\nu}}\|_2^2 (1 + \varrho_0^2) p_t^2(\alpha \varrho_1)}{m}. \tag{56}$$

E.3 Non-stationary Bellman equation

Proposition E.1 (Non-stationary Bellman equation; Proposition B.3 in [27]). *For $\pi \in \Pi$, we have*

$$\begin{aligned} Q_t^\pi(\boldsymbol{h}_t, \boldsymbol{a}_t) &= \mathbb{E}^\pi_{\mathbf{s}_t, \mathbf{h}_{t+1}, \mathbf{a}_{t+1} | \boldsymbol{h}_t, \boldsymbol{a}_t} \left[R(\mathbf{s}_t, \boldsymbol{a}_t) + \gamma Q^\pi_{t+1}(\mathbf{h}_{t+1}, \mathbf{a}_{t+1}) \right] \\ &= \mathbb{E}^\pi_{\mathbf{s}_t, \mathbf{h}_{t+1} | \boldsymbol{h}_t, \boldsymbol{a}_t} \left[R(\mathbf{s}_t, \boldsymbol{a}_t) + \gamma V^\pi_{t+1}(\boldsymbol{h}_{t+1}) \right], \end{aligned}$$

for any $(\boldsymbol{h}_t, \boldsymbol{a}_t) \in \mathcal{H} \times \mathcal{A}$ and $t \in \mathbb{Z}_{\geq 0}$.

E.4 Finite-time and finite-width bounds for symmetric Rec-NAC

Theorem E.1 (Finite-time bound for Rec-TD; Theorem 6.3 in [27]). *Assume for the symmetric history Q-function that $\{Q_t^\pi : t \in \mathbb{Z}_{\geq 0}\} \in \widetilde{\mathcal{F}}$ with a transportation mapping $\tilde{v} \in \widetilde{\mathcal{M}}$ such that*

$$\sup_{w \in \mathbb{R}} \|\tilde{v}_w(w)\|_2 \leq \tilde{\nu}_w, \text{ and } \sup_{\boldsymbol{u} \in \mathbb{R}^d_x} \|\tilde{v}_u(\boldsymbol{u})\|_2 \leq \tilde{\nu}_u.$$

Then, for any projection radius $\tilde{\boldsymbol{\rho}} \succeq \tilde{\boldsymbol{\nu}} = (\tilde{\nu}_w, \tilde{\nu}_u)^\top$ and step size $\eta_{td} > 0$, Rec-TD with max-norm regularization achieves the following error bound:

$$\mathbb{E}\left[\frac{1}{K} \sum_{k=0}^{K-1} \mathcal{R}_T^\pi(\theta_k) \right] \leq \frac{1}{\sqrt{K}} \left(\frac{\|\tilde{\boldsymbol{\nu}}\|_2^2}{(1-\gamma)} + \frac{\tilde{C}_T^{(1)}}{(1-\gamma)^3} \right) + \frac{\tilde{C}_T^{(2)}}{(1-\gamma)^2 \sqrt{m}} + \frac{\gamma^T}{(1-\gamma)K} \sum_{k=0}^{K-1} \tilde{\omega}_{T,k}^2, \tag{57}$$

for any $K \in \mathbb{N}$, where

$$\tilde{C}_T^{(1)}, \tilde{C}_T^{(2)} = \text{poly}\left(\sum_{k=0}^{T-1} \left| \left(\varrho_1(\alpha + \frac{\tilde{\rho}_w}{\sqrt{m}}) \right) \right|^k, \|\tilde{\boldsymbol{\rho}}\|_2, \|\tilde{\boldsymbol{\nu}}\|_2 \right)$$

are instance-dependent constants that do not depend on K, and

$$\tilde{\omega}_{t,k} := \sqrt{\mathbb{E}\left[\widetilde{F}_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_k) - Q_t^\pi(\boldsymbol{h}_t, \boldsymbol{a}_t))^2 \right]},$$

for $t, k \in \mathbb{Z}_{\geq 0}$.

For the average-iterate Rec-TD with $\bar{\theta}_K := \frac{1}{K} \sum_{k=0}^{K-1} \theta_k$, we have

$$\mathbb{E}\left[\mathcal{R}_T^\pi(\bar{\theta}_K) \right] \leq \frac{10}{(1-\gamma)\sqrt{K}} \left(\|\tilde{\boldsymbol{\nu}}\|_2^2 + \frac{\tilde{C}_T^{(1)}}{(1-\gamma)^2} \right) + \frac{10 \tilde{C}_T^{(2)}}{(1-\gamma)^2 \sqrt{m}} + \frac{10 \gamma^T}{(1-\gamma)K} \sum_{k=0}^{K-1} \tilde{\omega}_{T,k}^2. \tag{58}$$

Theorem E.2 (Finite-time bound for Rec-NPG; Theorem 7.3 in [27]). *Assume that $\mathbb{P}_T^{\pi^\star, P} \ll \mathbb{P}_T^{\pi_{\theta_n}, P}, n < N$, and let*

$$\Xi := \max_{0 \leq n < N} \left\| \frac{\mathbb{P}_T^{\pi^\star, P}}{\mathbb{P}_T^{\pi_{\theta_n}, P}} \right\|_\infty$$

We have the following result under Rec-NPG after $N \in \mathbb{Z}_{\geq 0}$ steps with step-size $\eta_{\text{npg}} = \frac{1}{\sqrt{N}}$ with projection radius $\boldsymbol{\rho}_G \in \mathbb{R}^2_{>0}$:

$$\begin{aligned} &\min_{0 \leq n < N} \mathbb{E}_0 \left[V^{\pi^\star}(P) - V^{\pi_{\theta_n}}(P) \right] \leq \frac{\log |\mathcal{A}|}{(1-\gamma)\sqrt{N}} \\ &+ \sqrt{p_T(\gamma)} \mathbb{E}_0 \left[\frac{1}{N} \sum_{n=0}^{N-1} \left(\Xi \varepsilon_{\text{cfa}}^T(\theta_n, \boldsymbol{g}_n) \right)^{\frac{1}{2}} \right] + \frac{2\gamma^T r_{\max}}{(1-\gamma)^2} \\ &+ \|\boldsymbol{\rho}_G\|_2^2 \sum_{t<T} \gamma^t \frac{2\beta_t + 12\left(\Lambda_t^2 \varrho_2 + \chi_t \varrho_1 \right) \sqrt{N}}{\sqrt{m_G}} \\ &+ \|\boldsymbol{\rho}_G\|_2^2 \sum_{t<T} \gamma^t \frac{12 L_t \sqrt{\Lambda_T^2 \varrho_2 + \chi_t \varrho_1}}{m_G^{1/4}} + \frac{\|\boldsymbol{\rho}_G\|_2^2}{2\sqrt{N}} \sum_{t<T} \gamma^t L_t^2 \end{aligned} \tag{59}$$

where $\varepsilon_{\text{cfa}}^T(\theta, \boldsymbol{g}) := \mathbb{E}_P^{\pi_{\theta_n}} \left[\sum_{t<T} \gamma^t \left| \nabla^\top \log \pi_t^\theta(\boldsymbol{a}_t | \boldsymbol{h}_t) \boldsymbol{g} - A_t^{\pi_\theta}(\boldsymbol{h}_t, \boldsymbol{a}_t) \right|^2 \right]$, and the sequence $(L_t, \beta_t, \Lambda_t, \chi_t)_t$ is defined in Lemma E.1.

E.5 Decomposition of the compatible function approximation error

Let $\varepsilon_{\mathrm{app}}$ denote the actor's approximation error,

$$\varepsilon_{\mathrm{app},n} = \inf \left\{ \mathbb{E}_{\mathbb{P}_T}^{\pi_{\theta_n},P} \sum_{t=0}^{T-1} \gamma^t \left[\nabla^\top G_t\left(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0\right) \boldsymbol{g} - Q_t^{\pi_{\theta_n}}\left(\boldsymbol{h}_t, \boldsymbol{a}_t\right) \right]^2 : \boldsymbol{g} \in \mathcal{B}_{2,\infty}^{(m_G)}(0, \boldsymbol{\rho}_G) \right\},$$

$\varepsilon_{\mathrm{td}}$ be the error of the critic,

$$\varepsilon_{\mathrm{td},n} = \mathbb{E}_{\theta_{n,k}, k \leq n} \left[\mathcal{R}_T^{\pi_{\theta_n}}\left(\bar{\theta}_n\right) \right],$$

and $\varepsilon_{\mathrm{sgd}}$ be the error in the policy update based on the compatible function approximation,

$$\varepsilon_{\mathrm{sgd},n} = \mathbb{E}_{\bar{\theta}_n, \theta_{n,k}, k \leq n} \left[\ell_T\left(\boldsymbol{g}_n; \theta_n, \hat{Q}^{(n)}\right) \right] - \inf_{\boldsymbol{g}} \mathbb{E}_{\bar{\theta}_n, \theta_{n,k}, k \leq n} \left[\ell_T\left(\boldsymbol{g}; \theta_n, \hat{Q}^{(n)}\right) \right].$$

We can decompose the compatible function approximation error $\varepsilon_{\mathrm{cfa}}^T$ into the approximation error for the RNN and the statistical errors as follows:

Proposition E.2 (Error decomposition for $\varepsilon_{\mathrm{cfa}}^T$; cf. Proposition 7.6 in [27])**.** *We have*

$$\begin{aligned} &\mathbb{E}\left[\mathbb{E}_P^{\pi_{\theta_n}} \left[\ell_T\left(\boldsymbol{g}_n; \theta_n, Q^{(n)}\right) \right] \mid \theta_k, k \leq n \right] \\ &\quad \leq \frac{8\|\tilde{\boldsymbol{\rho}}\|_2^2}{m_G} \sum_{t=0}^{T-1} \gamma^t \beta_t^2 + 8\varepsilon_{\mathrm{app},n} + 6\varepsilon_{\mathrm{td},n} + 2\varepsilon_{\mathrm{sgd},n} \end{aligned}$$

for any $n \in \mathbb{Z}_{\geq 0}$.

F Pseudocode of Rec-TD and Rec-NPG

This section contains the pseudocode of Rec-TD (cf. Algorithm 1) and Rec-NPG (cf. Algorithm 2).

Algorithm 1: Projected l-step TD learning algorithm

Input: policy π, bootstrap timestep l, step size α, number of updates K, projection radius $\boldsymbol{\rho}$.

for $k = 0$ **to** $K - 1$ **do**
  Initialize $\boldsymbol{s}_{k,0} \sim P$.
  Get $\boldsymbol{i}_{k,0} \sim I(\cdot|\boldsymbol{s}_{k,0})$.
  Get $\boldsymbol{o}_{k,0} \sim O(\cdot|\boldsymbol{i}_{k,0})$.
  for $l = 0$ **to** $L - 1$ **do**
    Sample action $\boldsymbol{a}_{k,l} \sim \pi(\cdot|\boldsymbol{h}_{k,l})$.
    Get reward $r_{k,l} \sim R(\cdot|\boldsymbol{s}_{k,l}, \boldsymbol{a}_{k,l})$.
    Get environment state $\boldsymbol{s}_{k,l+1} \sim T(\cdot|\boldsymbol{s}_{k,l}, \boldsymbol{a}_{k,l})$.
    Get information $\boldsymbol{i}_{k,l+1} \sim I(\cdot|\boldsymbol{s}_{k,l+1})$.
    Get observation $\boldsymbol{o}_{k,l+1} \sim O(\cdot|\boldsymbol{i}_{k,l+1})$.
  end
  Sample last action $\boldsymbol{a}_{k,L} \sim \pi(\cdot|\boldsymbol{h}_{k,L})$.
  Compute semi-gradient $\check{\nabla}\mathcal{R}_L(\boldsymbol{h}_{k,L}; \vartheta_k)$ according to Equation 17.
  Update $\check{\vartheta}_{k+1}$ according to Equation 18.
  Get ϑ_{k+1} using Equation 19.
end

Output: Average estimate $\overline{Q}^{\pi}(\cdot) = \widehat{Q}^{\pi}_{\hat{\vartheta}}(\cdot)$ with $\hat{\vartheta} = \frac{1}{K}\sum_{k=0}^{K-1} \vartheta_k$.

Algorithm 2: Recurrent Natural Policy Gradient (Rec-NPG) Algorithm

Input: Number of updates N, step sizes η_{npg}, projection radius $\tilde{\boldsymbol{\rho}}$.

Initialize actor RNN $G_t(\cdot; \theta_0, \boldsymbol{c})$.
for $n = 0$ **to** $N - 1$ **do**
  Obtain $\overline{Q}^{\pi}_n$ using Algorithm 1
  Initialize $\boldsymbol{s}_{n,0} \sim P$.
  for $k = 0$ **to** $K_{sgd} - 1$ **do**
    Get $\boldsymbol{i}_{n,k} \sim I(\cdot|\boldsymbol{s}_{n,k})$.
    Get $\boldsymbol{o}_{n,k} \sim O(\cdot|\boldsymbol{i}_{n,k})$.
    Sample action $\boldsymbol{a}_{n,k} \sim \pi(\cdot|\boldsymbol{h}_{n,k}; \theta_n)$.
    Get environment state $\boldsymbol{s}_{n,k+1} \sim T(\cdot|\boldsymbol{s}_{n,k}, \boldsymbol{a}_{n,k})$.
    Compute the gradient $\nabla_{\boldsymbol{g}} \ell_T \left(\boldsymbol{g}^k_n; \theta_n, \overline{Q}^{\pi_{\theta_n}}_k\right)$.
    Update $\tilde{\boldsymbol{g}}_{n,k+1}$ using Equation 21.
    Get $\hat{\boldsymbol{g}}_{n,k+1}$ using Equation 22.
  end
  Update policy parameters $\theta_{(n+1)}$ using Equation 23.
end

Output: Policy π_{θ_N}

G Numerical Experiments for Informed Asymmetric Rec-NAC

We evaluate the numerical performance of the informed asymmetric Rec-NAC method synthetic informed POMDP instances with a finite state space $\mathcal{S}$ (i.e., $|\mathcal{S}| = 10$), a discrete action space $\mathcal{A}$ (i.e., $|\mathcal{A}| = 4$), and continuous observation and information spaces.

Following the methodology of François-Lavet et al. [33], transition probabilities are initialized as follows: for each $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$, the corresponding entry in T is set to zero with probability 0.75 and otherwise sampled uniformly from $[0, 1]$. If all transitions from a given $(\boldsymbol{s}, \boldsymbol{a})$ pair are zero, we assign a non-zero probability to a randomly chosen next state to ensure reachability. The resulting probabilites are then normalized so that $\sum_{\boldsymbol{s}' \in \mathcal{S}} T(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) = 1$.

Rewards $r : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ are sampled independently at initialization:

$$r(\boldsymbol{s}, \boldsymbol{a}) \sim \text{Unif}(-1, 1).$$

Privileged information $\boldsymbol{i}_t \in \mathcal{I}$ is generated by sampling from a Gaussian distribution centered on a state-specific embedding:

$$\boldsymbol{i}_t \sim \mathcal{N}(\mu_{s_t}, \sigma^2 \boldsymbol{I}_{d_i}),$$

where $\sigma \in \mathbb{R}_{\geq 0}$ controls the noise level. Observations $\boldsymbol{o}_t \in \mathcal{O}$ are obtained by applying a noisy linear transformation to $\boldsymbol{i}_t$, with ς controlling the observation uncertainty. This construction ensures that $\boldsymbol{o}_t$ is conditionally independent of $\boldsymbol{s}_t$ given $\boldsymbol{i}_t$, i.e., $\boldsymbol{o}_t \perp \boldsymbol{s}_t | \boldsymbol{i}_t$.

The agent implements a recurrent policy $\pi(\boldsymbol{a}_t|\boldsymbol{h}_t)$, using an Elman-type RNN of width m followed by a linear softmax readout to produce:$\pi(\boldsymbol{a}_t|\boldsymbol{h}_t) = \text{Softmax}(\boldsymbol{c}_\pi^\top \boldsymbol{y}_t)$.

The critic uses a separate Elman RNN of width m. Its hidden state $\boldsymbol{f}_t$ is augmented with additional input $\boldsymbol{i}_t$, yielding the value estimate $\hat{Q}_\theta(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t; \vartheta) = \boldsymbol{b}^\top (\boldsymbol{f}_t \; \boldsymbol{i}_t)$. We train the critic using Rec-TD.

We compare the following variants of Rec-NAC, i.e, three canonical forms of $\boldsymbol{i}_t$, for various $m \in \{8, 64, 256\}$: (1) Rec-Nac with symmetric critic, i.e., $\boldsymbol{i}_t = \emptyset$; (2) Rec-NAC with informed asymmetric critic (IAAC) with privileged partial information, i.e., $\boldsymbol{i}_t \sim \mathcal{N}(\mu_{s_t}, \sigma^2 I_{d_i})$ with $\sigma = 0.1$; (3) Rec-NAC with informed asymmetric critic assuming full state access, i.e., $\boldsymbol{i}_t = \boldsymbol{s}_t$.

We train each configuration for $25{,}000$ episodes of length $T = 32$ using 10 random seeds and network widths $m \in \{8, 64, 256\}$. The actor and critic parameters are updated after each episode.

Figure 2 presents the TD error and episodic return during training, smoothed using a moving average over 100 episodes, for the different Rec-NAC variants evaluated at different network widths.

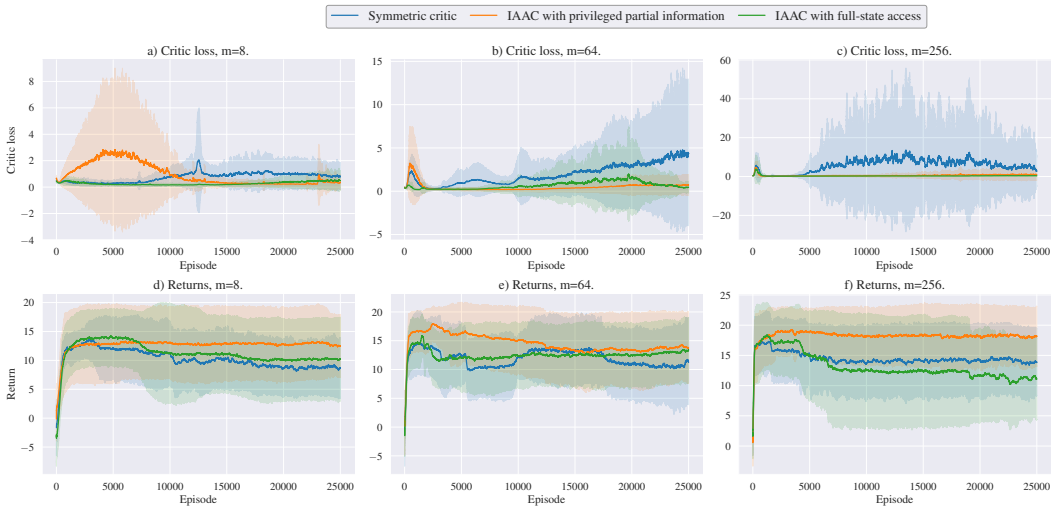

Figure 2: Critic loss (top row) and episodic return (bottom row) for different Rec-NAC variants across various network widths ($m \in \{8, 64, 256\}$). Each curve represents the mean over 10 independent runs and is smoothed using a moving average over 100 episodes; shaded regions denote the standard deviation.

Critic losses. Across all methods, except for the informed asymmetric actor-critic with privileged partial information at $m = 8$, critic losses decrease rapidly on average during the early training episodes, followed by fluctuations around a stable regime for the IAAC variants. For small network widths, all three models eventually converge to low-variance losses near zero, with the IAAC variant that has full-state access exhibiting the smoothest trajectory. For $m = 64$ and $m = 256$, the IAAC with privileged partial information achieves the most stable critic loss. In contrast, the alternative IAAC variant exhibits higher variance at $m = 64$, and the symmetric critic exhibits the most unstable loss curves. At the largest network width, both IAAC variants maintain comparable critic loss trajectories throughout training.

Episodic returns. The informed asymmetric actor–critic variants consistently achieve higher average returns than the symmetric variant after $25{,}000$ training episodes across all network widths. The performance gap is less pronounced at $m = 64$ than at $m = 8$ or $m = 256$. Notably, the symmetric Rec-NAC variant achieves the highest return for $m = 256$, indicating that it benefits from the larger network capacity, as smaller hidden dimensions may be insufficient to capture the sequence of past observations and actions accurately. Interestingly, the IAAC with only privileged partial information outperforms the full-state access variant across all network configurations. For $m = 256$, the performance of the full-state IAAC drops significantly after $5{,}000$ episodes, failing to reach high return levels and being surpassed by the symmetric variant. For $m = 8$, it initially converges to higher episodic returns than its partially-informed counterpart but subsequently experiences a performance drop. However, at $m = 64$, the IAAC with full-state access achieves only slightly lower returns. Across all configurations, return variance is generally high, consistent with the observed instability in critic loss, reflecting the stochasticity of environments sampled from the distribution of synthetic informed POMDPs.

Overall, the results indicate that the informed asymmetric actor–critic improves asymptotic learning performance compared to the symmetric critic, with the privileged-partial-information configuration yielding the highest episodic returns. Nonetheless, our findings suggest a trade-off between model capacity and learning performance.