THEORETICAL INSIGHTS INTO INFORMED VALUE ESTIMATION FOR ASYMMETRIC RECURRENT NATURAL ACTOR-CRITIC

Anonymous authors
Paper under double-blind review

ABSTRACT

In partially observable environments, reinforcement learning agents must act based on incomplete or noisy observations. While asymmetric actor-critic methods empirically improve learning, their theoretical properties remain largely unexplored. We generalize the Recurrent Natural Actor-Critic (Rec-NAC) algorithm to the asymmetric setting, where a state-dependent privileged signal is available only during training and is used by the critic. Both actor and critic consist of a recurrent encoder of the observable history followed by a linear head; the critic's head is learnable and integrates the privileged signal, while the actor does not. By adapting the finite-time and finite-width analysis of policy evaluation via recurrent temporal-difference learning to this setting, we identify distinct information and representation benefits of an informed critic and uncover a fundamental trade-off: privileged information reduces conditional target variance and removes the irreducible error due to information aliasing, while the informed architecture can improve best-in-class approximation but also induces additional finite-time complexity. We further analyze the resulting recurrent natural policy-gradient update and show that privileged signals induce an information mismatch in the actor's compatible function regression. Our results characterize when privileged information improves value estimation and its implications for policy learning.

1 INTRODUCTION

In many real-world reinforcement learning problems, the true state of the environment is not directly observable. Instead, agents must make decisions based on partial or noisy observations of the underlying system. The partially observable Markov decision process (POMDP) formalism (Kaelbling et al., 1998) models this setting, where optimal policies depend on histories of observations and actions rather than just the last state. Since agents may need to rely on the entire interaction history to act optimally (Singh et al., 1994), recurrent neural networks (RNNs) provide a natural parameterization of such history-dependent policies by encoding histories into a fixed-size latent representation (Zhu et al., 2017; Zhang et al., 2016; Hausknecht & Stone, 2015; Wierstra et al., 2010; Bakker, 2001). However, learning in these settings is often sample-inefficient due to the large history space and possibly suboptimal due to sparse reward-based supervision.

Asymmetric actor-critic methods mitigate this challenge by granting the critic access to privileged signals during training, while constraining the actor to use only histories of past observations and actions. This asymmetry should enable more accurate value estimation and improve policy updates, and has demonstrated strong empirical performance across domains (Dürr et al., 2026; Vasco et al., 2024; Kaufmann et al., 2023; Degrave et al., 2022; Pinto et al., 2017). Early asymmetric formulations with critics conditioning solely on the environment's full state (Pinto et al., 2017) generally induce biased policy gradients (Baisero & Amato, 2022). Subsequent work resolves this issue by introducing history-state value functions that explicitly condition on both the history and state, yielding unbiased gradients (Baisero & Amato, 2022). However, many training environments lie between the extremes of full-state observability and no privileged signals (Lambrechts et al., 2024): some internal variables, such as auxiliary measurements, forecasts, features, or simulator proxies, that correlate with the unobservable state without fully revealing it, may be available during training. Hence, Ebi et al. (2026) generalize the asymmetric framework of Baisero & Amato (2022) to critics

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

that leverage arbitrary state-dependent privileged signals, and raise the question of which additional information is most beneficial for value estimation, proposing practical informativeness criteria for privileged signal selection.

From a theoretical perspective, convergence guarantees for policy-gradient and actor-critic methods are well established in fully observable MDPs, including natural policy gradient methods under linear and neural function approximation (Agarwal et al., 2021; Wang et al., 2020; Liu et al., 2019) and actor-critic algorithms under both i.i.d. and Markovian sampling assumptions (Wu et al., 2020; Tian et al., 2023; Cayci et al., 2024a). Extensions to partially observable settings have recently emerged. Under symmetric observability in POMDPs, Cayci et al. (2024b) analyze a natural actor-critic method with an internal memory state, while Cayci & Eryilmaz (2025) investigate the recurrent natural actor-critic (Rec-NAC) method. Asymmetric variants have also been studied with a fixed agent state and linear critics (Lambrechts et al., 2025) or tabular belief-weighted formulations (Cai et al., 2024). However, a rigorous theoretical understanding of asymmetric actor-critic algorithms with history-dependent function approximators leveraging privileged information is still lacking.

In this work, we address this gap by generalizing the Rec-NAC framework of Cayci & Eryilmaz (2025) to the asymmetric setting where the critic has access to arbitrary state-dependent privileged signals during training. Our contributions are fourfold. First, we propose an informed critic architecture that combines a recurrent encoder of observation-action histories with privileged inputs through a learnable linear head, while the actor uses a separate recurrent encoder with a fixed linear readout that depends only on observable inputs. We further modify the Rec-NAC update rules for this asymmetric learning setting. Second, we extend the finite-time and finite-width convergence analysis of recurrent temporal-difference learning (Rec-TD) to the informed critic. Third, we characterize the effects of privileged information on value estimation by comparing the informed and symmetric critics, identifying information-aliasing and representation effects and deriving a sufficient condition under which the resulting benefits outweigh the additional finite-time and finite-width costs. Fourth, we analyze the effects of informed value estimation on the recurrent natural policy-gradient (Rec-NPG) updates, showing that the actor's limited information can induce an information-mismatch contribution to compatible function approximation.

Overall, to the best of our knowledge, our results provide the first theoretical foundation for asymmetric recurrent actor-critic methods in partially observable environments and clarify the role of privileged information in value estimation and its effect on policy optimization.

2 Background

In this section, we describe the decision processes as well as the history and history-action value functions considered in this work. An overview of the notation can be found in Appendix A.

2.1 Partially Observable Markov Decision Processes

A POMDP (Kaelbling et al., 1998) is a discrete-time partially observable control problem defined by an 8-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, P, \gamma)$. Here, $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{O}$ denote the state, action and observation spaces, respectively. The initial state distribution $P \in \Delta(\mathcal{S})$ gives the probability of $\boldsymbol{s}_0 \in \mathcal{S}$ being the process' initial state. The transition distribution $T\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ defines the dynamics of the system by providing the probability $T(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$ of transitioning to state $\boldsymbol{s}_{t+1} \in \mathcal{S}$ after taking action $\boldsymbol{a}_t \in \mathcal{A}$ in state $\boldsymbol{s}_t \in \mathcal{S}$. At time t, the agent receives an observation $\boldsymbol{o}_t \sim O(\cdot|\boldsymbol{s}_t)$, where $O\colon \mathcal{S} \to \Delta(\mathcal{O})$ gives the probability of observing $\boldsymbol{o}_t \in \mathcal{O}$ in state $\boldsymbol{s}_t$. The immediate reward for taking action $\boldsymbol{a}_t$ in state $\boldsymbol{s}_t$ is induced by the reward function $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. We assume uniformly bounded rewards.

Assumption 1 (Bounded rewards)**.** *For all* $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$, $|R(\boldsymbol{s}, \boldsymbol{a})| \leq r_{\max}$ *for some* $r_{max} > 0$.

In the partially observable setting, the agent must select actions based on the observable history $\boldsymbol{h}_t := (\boldsymbol{o}_0, \boldsymbol{a}_0, \ldots, \boldsymbol{o}_t)$, i.e., the sequence of past observations and actions. We define the set of 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. A policy, therefore, maps histories to action distributions, $\pi\colon \mathcal{H} \to \Delta(\mathcal{A})$. The objective is to maximize the expected discounted return $J(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(\boldsymbol{s}_t, \boldsymbol{a}_t)\right]$, where $\gamma \in [0, 1)$ is a discount factor.

2.2 INFORMED POMDP

The informed POMDP (Lambrechts et al., 2024) extends the standard POMDP formulation by introducing a so-called information space $\mathcal{I}$ and a corresponding information distribution $I : \mathcal{S} \to \Delta(\mathcal{I})$, which gives the probability of receiving information $\boldsymbol{i}_t \in \mathcal{I}$ given the unobservable state $\boldsymbol{s}_t \in \mathcal{S}$. Consequently, the informed POMDP is defined by the 10-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{O}, T, I, \widetilde{O}, R, P, \gamma)$. In contrast to the standard POMDP, the observation distribution is defined as $\widetilde{O} : \mathcal{I} \to \Delta(\mathcal{O})$, which specifies the probability of observing $\boldsymbol{o}_t$ given $\boldsymbol{i}_t$, i.e., $p(\boldsymbol{o}_t|\boldsymbol{s}_t, \boldsymbol{i}_t) = \widetilde{O}(\boldsymbol{o}_t|\boldsymbol{i}_t)$. For each informed POMDP, there is an underlying execution POMDP defined as $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, P, \gamma)$, where $O(\boldsymbol{o}_t|\boldsymbol{s}_t) = \mathbb{E}_{\mathbf{i}_t|\boldsymbol{s}_t}\left[\widetilde{O}(\boldsymbol{o}_t|\mathbf{i}_t)\right]$.

Assumption 2 (Independent trajectories)**.** *Given an initial state distribution P, we can obtain 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 $\{\boldsymbol{h}_{k,t}\}_{t\in\mathbb{Z}_{\geq 0}}$ for $k \in \mathbb{Z}_{\geq 0}$ by independently starting the system from $\boldsymbol{s}_{k,0} \sim P$.*

2.3 HISTORY-BASED VALUE FUNCTIONS AND POLICY GRADIENTS

To treat the standard and informed history-based value functions within a unified notation, we set $\boldsymbol{i}_t = \boldsymbol{o}_t$ when the privileged information is absent in the standard POMDP setting. We interpret this as reducing to the standard belief model, i.e., $p(\boldsymbol{s}_t|\boldsymbol{h}_t, \boldsymbol{i}_t) = p(\boldsymbol{s}_t|\boldsymbol{h}_t)$ for $\boldsymbol{i}_t = \boldsymbol{o}_t$.

Hence, for a history $\boldsymbol{h}_t$ and a privileged signal $\boldsymbol{i}_t \in \mathcal{I}$, the (informed) history Q-function is given by $Q^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) = \mathbb{E}^\pi_{\boldsymbol{s}_{t:\infty}, \boldsymbol{a}_{t+1:\infty}|\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t}\left[\sum_{j=0}^\infty \gamma^j R(\boldsymbol{s}_{t+j}, \boldsymbol{a}_{t+j})\right]$. The corresponding informed history value function $V^\pi : \mathcal{H} \times \mathcal{I} \to \mathbb{R}$ satisfies the Bellman relation

$$V^\pi(\boldsymbol{h}, \boldsymbol{i}) = \textstyle\sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{h}) Q^\pi(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a}), \quad Q^\pi(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a}) = R(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a}) + \gamma\, \mathbb{E}_{\mathbf{o}', \mathbf{i}'|\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a}}\left[V^\pi(\boldsymbol{h}', \mathbf{i}')\right],$$

where $R(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) = \mathbb{E}_{\mathbf{s}_t|\boldsymbol{h}_t, \boldsymbol{i}_t}\left[R(\mathrm{s}_t, \boldsymbol{a}_t)\right]$ is the (informed) history-based reward and $\boldsymbol{h}' = \boldsymbol{h}\boldsymbol{a}\boldsymbol{o}'$.

The value functions $Q(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t)$ and $V^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t)$ coincide with the informed history-based Q- and value functions of Ebi et al. (2026). When $\boldsymbol{i}_t = \boldsymbol{o}_t$, the above definitions reduce to the standard history-based functions, i.e., $V^\pi(\boldsymbol{h}_t, \boldsymbol{o}_t) = V^\pi(\boldsymbol{h}_t)$ and $Q^\pi(\boldsymbol{h}_t, \boldsymbol{o}_t, \boldsymbol{a}_t) = Q^\pi(\boldsymbol{h}_t, \boldsymbol{a}_t)$. For $\boldsymbol{i}_t = \boldsymbol{s}_t$, the informed history value reduces to the history-state value $V^\pi(\boldsymbol{h}_t, \boldsymbol{s}_t)$ of Baisero & Amato (2022). Moreover, the informed history-based functions are unbiased estimators of their symmetric counterparts, in the sense that $\mathbb{E}_{\mathbf{i}|\boldsymbol{h}}[Q^\pi(\boldsymbol{h}, \mathbf{i}, \boldsymbol{a})] = Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ and $\mathbb{E}_{\mathbf{i}|\boldsymbol{h}}[V^\pi(\boldsymbol{h}, \mathbf{i})] = V^\pi(\boldsymbol{h})$.

The standard policy gradient is $\nabla_\theta J(\pi_\theta) = \mathbb{E}\left[\sum_t \gamma^t Q^\pi(\boldsymbol{h}_t, \boldsymbol{a}_t) \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\right]$, whereas the informed asymmetric policy gradient using the unbiased informed history Q-function is

$$\nabla^{\mathrm{IAAC}}_\theta J(\pi_\theta) = \mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t Q^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\right]. \tag{1}$$

As shown by Ebi et al. (2026), the two gradients are equivalent, i.e., $\nabla^{\mathrm{IAAC}}_\theta J(\pi_\theta) = \nabla_\theta J(\pi_\theta)$.

3 RECURRENT ACTOR-CRITIC ARCHITECTURES

Figure 1 in Appendix C illustrates the architectures of the actor and the informed critic neural network studied in this work. Both networks are implemented as Elman-type RNN with smooth activations and diagonal recurrence matrix (see Appendix B.1 for details), also known as independently recurrent neural networks (IndRNNs). They encode the observable history using weights shared across time and are followed by a linear head. The actor and critic use separate recurrent encoders and differ in their readout heads.

3.1 ACTOR

The actor is an IndRNN $G_t(\cdot;\theta)$ of width $m_G \in \mathbb{N}$, parameterized by $\theta = (\boldsymbol{w}_G, \boldsymbol{U}_G)$, where $\boldsymbol{w}_G \in \mathbb{R}^{m_G}$ defines the diagonal recurrent weight matrix $\boldsymbol{W}_G := \mathrm{diag}(\boldsymbol{w}_G) \in \mathbb{R}^{m_G \times m_G}$, 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 $\boldsymbol{x}_t = \begin{pmatrix}\boldsymbol{o}_t & \boldsymbol{a}_t\end{pmatrix}^\top$, and the resulting hidden state is denoted by $\boldsymbol{g}_t = G_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)$. A linear projection with fixed weights $\boldsymbol{c} \in \mathbb{R}^{m_G}$ produces a scalar output $D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta) = \frac{1}{\sqrt{m_G}}\langle \boldsymbol{c}, \boldsymbol{g}_t\rangle = \frac{1}{\sqrt{m_G}}\langle \boldsymbol{c}, G_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta)\rangle$. We define the policy distribution over actions as the softmax over the output logits, i.e., $\pi_\theta(\boldsymbol{a}|\boldsymbol{h}_t) := \frac{\exp(D_t(\boldsymbol{h}_t, \boldsymbol{a}; \theta))}{\sum_{\boldsymbol{a}'\in\mathcal{A}} \exp(D_t(\boldsymbol{h}_t, \boldsymbol{a}'; \theta))}$.

3.2 ASYMMETRIC CRITIC

The asymmetric critic, which we also refer to as the informed critic, estimates the informed history Q-function $Q(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t)$. It comprises an IndRNN $F_t(\cdot; \vartheta^F)$ of width m_F, parameterized by $\vartheta^F = (\boldsymbol{w}_F, \boldsymbol{U}_F)$, and a linear readout 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 privileged signal $\boldsymbol{i}_t \in \mathbb{R}^{d_i}$. The recurrent weights are defined analogously to the actor: $\boldsymbol{W}_F := \mathrm{diag}(\boldsymbol{w}_F) \in \mathbb{R}^{m_F \times m_F}$ is diagonal and $\boldsymbol{U}_F \in \mathbb{R}^{m_F \times d_x}$ 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 $\boldsymbol{z}_t = \boldsymbol{f}_t \oplus \boldsymbol{i}_t$. With parameters $\vartheta = (\boldsymbol{w}_F, \boldsymbol{U}_F, \boldsymbol{b})$, the critic output is $\Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta) := \frac{1}{\sqrt{d_z}} \langle \boldsymbol{b}, \boldsymbol{z}_t \rangle = \frac{1}{\sqrt{d_z}} \langle \boldsymbol{b}, \boldsymbol{f}_t \oplus \boldsymbol{i}_t \rangle = \frac{1}{\sqrt{d_z}} \langle \boldsymbol{b}, F_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F) \oplus \boldsymbol{i}_t \rangle$.

Note that the symmetric critic architecture is obtained from the asymmetric critic by removing access to the privileged signal $\boldsymbol{i}_t$ (see Appendix C.1 for details).

Infinite-width limit of the informed critic. To characterize the function class and training dynamics, we consider the infinite-width limit of the informed critic, i.e., as $m_F \to \infty$. At initialization, we represent the parameters of the recurrent encoder as $\boldsymbol{w}_0 \in \mathbb{R}^{m_F}$ and $\boldsymbol{U}_0 \in \mathbb{R}^{m_F \times d_x}$, where the entries of $\boldsymbol{w}_0$ and the rows of $\boldsymbol{U}_0$ are independently sampled as $w_0^i \sim \mathrm{Unif}(-\alpha, \alpha)$ with $\alpha > 0$ and $\boldsymbol{U}_0^i \sim \mathcal{N}(0, \boldsymbol{I}_{d_x})$. The weights of the linear readout are initialized as $\boldsymbol{b}_0 \sim \mathcal{N}(0, \boldsymbol{I}_{d_z})$. Let $\vartheta_0 = (\vartheta_0^F, \boldsymbol{b}_0)$ with $\vartheta_0^F = (\boldsymbol{w}_0, \boldsymbol{U}_0)$ denote the initial parameters of the informed critic.

In the infinite-width regime, the network is well approximated by its first-order Taylor expansion around random initialization, inducing the neural tangent kernel (NTK) (Jacot et al., 2018). Intuitively, as the width grows, individual parameter updates become small, and the network behaves nearly linearly in its parameters around initialization. Hence, training corresponds to kernel regression in the reproducing kernel Hilbert space (RKHS) associated with the NTK. Equivalently, the NTK admits a representation in terms of neural tangent random features (NTRFs), where each feature corresponds to the gradient of the network output w.r.t. its initialization (see Appendix B.4 for details). The NTRF feature map for the informed critic at time t is

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

where $\boldsymbol{\psi}_t^F$ coincides with the recurrent NTRF of the symmetric critic presented in Cayci & Eryilmaz (2025), and the additional block corresponds to the linear readout of the privileged embedding. In prior work (Cayci & Eryilmaz, 2025), the readout weights were assumed fixed; here, we generalize this formulation by making the readout learnable.

For a sequence $((\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t))_{0 \le t \le T-1}$, stacking features $\boldsymbol{\psi}_t^\top(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta_0)$ across time yields the NTRF matrix $\boldsymbol{\Psi}_T(\cdot; \vartheta_0)$ with $\boldsymbol{\Psi}(\cdot; \vartheta_0) := \boldsymbol{\Psi}_\infty(\cdot; \vartheta_0)$.

Definition 3.1 (Reference function class of the informed critic)**.** *Let $\mathcal{M}_F$ denote the set of admissible transportation mappings* [1]

$$v : \mathbb{R}^{1+d_x+d_z} \to \mathbb{R}^{1+d_x+d_z}, \quad \vartheta_0 = (w_0^i, \boldsymbol{U}_0^i, \boldsymbol{b}_0) \mapsto \begin{pmatrix} v_w(w_0^i) \\ v_u(\boldsymbol{U}_0^i) \\ v_b(\boldsymbol{b}_0) \end{pmatrix}, \quad \textit{with } \mathbb{E}\left[\|v(\vartheta_0)\|_2^2\right] < \infty.$$

Then, for any sequence length $T \ge 1$, the reference function class of the informed critic is

$$\mathcal{F}_T := \left\{ (\boldsymbol{h}, \boldsymbol{a}, \boldsymbol{i}) \mapsto \mathbb{E}_{\vartheta_0}\left[\boldsymbol{\Psi}_k(\boldsymbol{h}, \boldsymbol{a}, \boldsymbol{i}; \vartheta_0)\, v(\vartheta_0)\right] = \begin{pmatrix} \Phi_0^\star(\boldsymbol{h}_0, \boldsymbol{a}_0, \boldsymbol{i}_0; v) \\ \vdots \\ \Phi_T^\star(\boldsymbol{h}_T, \boldsymbol{a}_T, \boldsymbol{i}_T; v) \end{pmatrix} : v \in \mathcal{M}_F \right\},$$

where $(\boldsymbol{h}, \boldsymbol{a}, \boldsymbol{i}) \in \mathcal{H} \times \mathcal{A} \times \mathcal{I}$ and $\Phi_t^\star(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; v) = \mathbb{E}\left[\langle v(\vartheta_0), \boldsymbol{\psi}_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta_0) \rangle\right]$. We denote the infinite-horizon class by $\mathcal{F} := \mathcal{F}_\infty$.

[1] A transportation mapping describes how the infinitely many model parameters are displaced from their initialization during training to approximate a target function. We call such a mapping *admissible* if it satisfies mild regularity conditions (e.g., $\mathbb{E}[\|v(\vartheta_0)\|_2^2] < \infty$) ensuring that the induced function remains finite and the infinite-width limit is well defined.

The informed critic inherits the NTK structure of the symmetric critic while augmenting it with a privileged-input block, which enables richer value approximations at training time. Consequently, its reference function class contains that of the symmetric critic, i.e., $\tilde{\mathcal{F}} \subseteq \mathcal{F}$.

4 Asymmetric Recurrent Natural Actor-Critic Algorithm

In this section, we generalize the recurrent natural actor-critic (Rec-NAC) algorithm of Cayci & Eryilmaz (2025), originally formulated for standard POMDPs where both actor and critic are conditioned only on the observable history $\boldsymbol{h}_t \in \mathcal{H}$, to the asymmetric learning setting. The actor remains history-only, while the critic is informed: during training it receives the privileged signal $\boldsymbol{i}_t \sim I(\cdot|\boldsymbol{s}_t)$. Asymmetric Rec-NAC still comprises two loops: (i) recurrent temporal-difference (Rec-TD) learning for policy evaluation, and (ii) recurrent natural policy gradient (Rec-NPG) updates for the actor. The full Rec-NAC procedure is summarized in Algorithm 1 in Appendix C.

When the privileged signal is unavailable, the asymmetric Rec-NAC collapses to the original method of Cayci & Eryilmaz (2025) for symmetric settings. The sole modification is that we make the critic's readout weights learnable, whereas in prior work they were frozen at initialization.

4.1 Asymmetric Recurrent Temporal-Difference (Rec-TD) Learning

Building on the symmetric critic's objective (Cayci & Eryilmaz, 2025), we aim to find the parameters $\vartheta \in \Omega^{\vartheta}_{\boldsymbol{\rho},m_F,d_z}$ that minimize $\mathcal{R}^{\pi}_T(\vartheta) := \mathbb{E}^{\pi}\left[\sum_{t=0}^{T-1} \gamma^t \left(\Phi_t(\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{i}_t;\vartheta) - Q^{\pi}_t(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)\right)^2\right]$, given the symmetric random initialization $(\boldsymbol{w}_0, \boldsymbol{U}_0, \boldsymbol{b}_0)$, projection radii $\boldsymbol{\rho} = \begin{pmatrix}\rho_w & \rho_u & \rho_b\end{pmatrix}^{\top} \in \mathbb{R}^3_{>0}$, and $\Omega^{\vartheta}_{\boldsymbol{\rho},m_F,d_z} = \left\{\vartheta : \max_i |w^i - w^i_0| \le \frac{\rho_w}{\sqrt{m_F}}, \max_i \|\boldsymbol{U}^i - \boldsymbol{U}^i_0\| \le \frac{\rho_{\boldsymbol{u}}}{\sqrt{m_F}}, \max_j |b^j - b^j_0| \le \frac{\rho_{\boldsymbol{b}}}{\sqrt{d_z}}\right\}$.

To perform optimization, Rec-TD employs projected semi-gradient TD updates based on the TD error $\delta_t(\boldsymbol{h}_{t+1},\boldsymbol{a}_{t+1},\boldsymbol{i}_{t+1};\vartheta) := r_t + \gamma\Phi_{t+1}(\boldsymbol{h}_{t+1},\boldsymbol{a}_{t+1},\boldsymbol{i}_{t+1};\vartheta) - \Phi_t(\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{i}_t;\vartheta)$, as the target function Q^{π}_t is unknown and to be learned. This leads to the following update direction:

$$\check{\nabla}\mathcal{R}_T(\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{i}_t;\vartheta) = \textstyle\sum_{t=0}^{T-1} \gamma^t \delta_t(\boldsymbol{h}_{t+1},\boldsymbol{a}_{t+1},\boldsymbol{i}_{t+1};\vartheta)\cdot\nabla_{\vartheta}\Phi_t(\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{i}_t;\vartheta). \tag{2}$$

Under Assumption 2, for step size $\eta_{\text{td}} > 0$ and $k \in \mathbb{Z}_{\ge 0}$, Rec-TD with max-norm regularization uses a projected semi-gradient TD update, i.e.,

$$\vartheta_{k+1} = \texttt{Proj}_{\Omega_{\boldsymbol{\rho},(m_F,d_z)}}[\vartheta_k + \eta_{\text{td}}\cdot\check{\nabla}\mathcal{R}_T(\boldsymbol{h}_{k,T},\boldsymbol{a}_{k,T},\boldsymbol{i}_{k,T};\vartheta_k)]. \tag{3}$$

The projection $\texttt{Proj}$ keeps the parameters within the admissible norm-constrained set. For a complete overview of the Rec-TD learning loop, refer to the pseudocode in Algorithm 2 in Appendix C.

4.2 Asymmetric Recurrent Natural Policy Gradient (Rec-NPG)

To approximate the natural gradient of the informed asymmetric policy objective, we adopt a truncated path-based compatible function approximation (Kakade, 2001) as in Rec-NPG of Cayci & Eryilmaz (2025). Given a truncation horizon $T \in \mathbb{N}$ and a history-dependent policy π_{θ_n}, the corresponding output of the critic is defined as $\hat{Q}^{\pi_{\theta_n}}_t := \Phi_t(\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{i}_t;\hat{\vartheta}_n)$, where $\hat{\vartheta}_n$ denotes the critic parameters obtained by Rec-TD at iteration $n \in \mathbb{Z}_{\ge 0}$. The actor then solves the following optimization problem to compute the compatible direction $\boldsymbol{q}$:

$$\min_{\boldsymbol{q}}\ \mathbb{E}\left[\ell_T(\boldsymbol{q};\theta,\hat{Q})\right] = \mathbb{E}\left[\textstyle\sum_{t=0}^{T-1}\gamma^t\left(\nabla^{\top}\log\pi^{\theta}_t(\boldsymbol{a}_t|\boldsymbol{h}_t)\boldsymbol{q} - \hat{A}^{\pi_{\theta_n}}_t(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)\right)^2\right],$$

subject to $\boldsymbol{q} \in \mathcal{B}^{(m_G)}_{2,\infty}(0,\boldsymbol{\rho}_G)$ (see Appendix B.3 for details), where the informed advantage estimate is $\hat{A}^{\pi_{\theta_n}}_t(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t) = \hat{Q}^{\pi_{\theta_n}}_t(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t) - \sum_{\boldsymbol{a}\in\mathcal{A}}\pi^{\theta}_t(\boldsymbol{a}|\boldsymbol{h}_t)\,\hat{Q}^{\pi_{\theta_n}}_t(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a})$.

Using compatible-function approximation, the natural-gradient direction is obtained by solving a convex regression problem via projected stochastic gradient descent (SGD). Hence, this method does not require constructing or inverting the Fisher information matrix (Bagnell & Schneider, 2003). Let $\hat{\boldsymbol{q}}_{n,k}$ denote the estimate after $k \in \mathbb{Z}_{\ge 0}$ SGD steps during the n-th policy update cycle. Then,

$$\hat{\boldsymbol{q}}_{n,k+1} = \texttt{Proj}_{\mathcal{B}^{(m_G)}_{2,\infty}}\left[\hat{\boldsymbol{q}}_{n,k} - \eta_{\text{sgd}}\nabla_{\boldsymbol{q}}\ell_T(\boldsymbol{q}_{n,k};\theta,\hat{Q})\right], \tag{4}$$

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

$$\theta_{(n+1)} = \theta^{(n)} + \eta_{\text{npg}} \cdot \frac{1}{K_{\text{sgd}}} \sum_{k=0}^{K_{\text{sgd}}-1} \hat{\boldsymbol{q}}_{n,k}. \tag{5}$$

5 Theoretical Insights into Rec-NAC with Informed Critic

In this section, we analyze the asymmetric variant of Rec-NAC in which the critic has access to a state-dependent privileged signal $\boldsymbol{i}_t$ during training, while the actor is conditioned only on the observable history. Our analysis first identifies three effects of privileged information on value estimation. The first is an information effect: conditioning the critic on $\boldsymbol{i}_t$ can reduce the conditional variance of the temporal-difference target and eliminate the information-aliasing error that is irreducible for a critic observing only $(\boldsymbol{h}_t, \boldsymbol{a}_t)$. The second is a representation effect: the informed architecture induces a larger function class and may provide a more effective representation of value-relevant structure, even when the privileged signal is, in principle, encoded in the observable history. The third is a complexity effect: leveraging $\boldsymbol{i}_t$ can introduce additional finite-time and finite-width costs through larger sensitivity, smoothness, parameter-norm, and finite-width constants. These effects lead to a trade-off between the benefits of privileged information for value estimation and the additional complexity of exploiting it.

We first characterize the information and representation effects in value estimation. We then derive a finite-time and finite-width bound for informed Rec-TD and compare it with the corresponding symmetric bound, yielding a sufficient condition under which the information-related benefits outweigh the additional finite-time costs. Finally, we investigate how informed value estimation affects Rec-NPG.

5.1 Information and representation effects in value estimation

The informed and symmetric critics condition on different information sets. The informed critic targets $Q_t^\pi := Q^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t)$, whereas the symmetric critic targets the conditional projection onto the information available to the actor, i.e., $\bar{Q}_t^\pi := \mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t,\boldsymbol{a}_t}\left[Q^\pi(\boldsymbol{h}_t, \mathbf{i}_t, \boldsymbol{a}_t)\right] = Q_t^\pi(\boldsymbol{h}_t, \boldsymbol{a}_t)$. The unbiasedness property of Q_t^π directly extends to the informed advantage function, i.e., $\mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t}\left[A^\pi(\boldsymbol{h}_t, \mathbf{i}_t, \boldsymbol{a}_t)\right] = A^\pi(\boldsymbol{h}_t, \boldsymbol{a}_t)$ (cf. Lemma E.1 in Appendix E.1). Thus, privileged information does not alter any population-level Bellman quantities, but only refines their conditional realizations.

We first isolate the information-theoretic effect of privileged information independently of the function class and optimization algorithm. Let $y_t := r_t + \gamma Q_{t+1}^\pi(\boldsymbol{h}_{t+1}, \boldsymbol{i}_{t+1}, \boldsymbol{a}_{t+1})$ denote the informed temporal-difference target. By the law of total variance, for almost every $(\boldsymbol{h}_t, \boldsymbol{a}_t)$, $\mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t}\left[\mathrm{Var}(y_t|\boldsymbol{h}_t, \boldsymbol{a}_t, \mathbf{i}_t)\right] \le \mathrm{Var}(y_t|\boldsymbol{h}_t, \boldsymbol{a}_t)$ (cf. Proposition E.1 in Appendix E.2). The inequality is strict whenever $\mathrm{Var}\left(\mathbb{E}\left[y_t|\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t\right]|\boldsymbol{h}_t, \boldsymbol{a}_t\right) > 0$. Privileged information, therefore, reduces the irreducible noise in TD targets without introducing bias, while preserving the same Bellman fixed point in expectation. Importantly, $\boldsymbol{i}_t$ need not reveal the full unobserved $\boldsymbol{s}_t$; any additional information that reduces conditional uncertainty about future returns can contribute to this effect.

This variance-reduction effect is distinct from the representational advantage of the informed critic architecture. Under the corresponding parameterization and projection constraints, the informed hypothesis class contains the symmetric class, $\tilde{\mathcal{F}}_T \subseteq \mathcal{F}_T$, but inclusion alone does not imply a strict approximation improvement. Nevertheless, privileged information may facilitate finite-class approximation even when it is already encoded in the history $\boldsymbol{h}_t$. Notably, the critic must learn to extract value-relevant structure from the history, whereas an explicit state-dependent signal can make such structure directly accessible.

A useful interpretation is to view the privileged signal as a realization of latent structure summarized by the history. The history induces a belief over possible latent states, which provides a natural control-relevant representation; a state-dependent signal can therefore expose an explicit instantiation of structure that would otherwise have to be inferred from $\boldsymbol{h}_t$. Thus, $\boldsymbol{i}_t$ can benefit the critic through two distinct mechanisms: it can resolve value-relevant uncertainty that remains after conditioning on $(\boldsymbol{h}_t, \boldsymbol{a}_t)$, or it can provide a more accessible representation of information already contained in the history. The former is an information effect, captured by the conditional variance

and information-aliasing terms below; the latter is a representation effect that can persist even when $\boldsymbol{i}_t$ is a deterministic function of $\boldsymbol{h}_t$.

To make the information-induced limitation of a symmetric critic explicit, we characterize its error incurred when it is evaluated against the refined target Q_t^π.

Lemma 5.1 (Decomposition of the symmetric critic's error w.r.t. Q_t^π)**.** *For any parameters $\tilde{\vartheta}$,*

$$\mathbb{E}\left[\left(\Phi_t^{sym}(\boldsymbol{h}_t, \boldsymbol{a}_t; \tilde{\vartheta}) - Q_t^\pi\right)^2\right] = \mathbb{E}\left[(\Phi_t^{sym}(\boldsymbol{h}_t, \boldsymbol{a}_t; \tilde{\vartheta}) - \bar{Q}_t^\pi)^2\right] + \mathbb{E}\left[\mathrm{Var}\left(Q_t^\pi | \boldsymbol{h}_t, \boldsymbol{a}_t\right)\right]. \quad (6)$$

We refer to the second term as the information-aliasing error, $\varepsilon_{\text{alias},t} := \mathbb{E}\left[\mathrm{Var}\left(Q_t^\pi | \boldsymbol{h}_t, \boldsymbol{a}_t\right)\right]$. It measures the variation in the value target across privileged signals that are indistinguishable to a critic observing only $(\boldsymbol{h}_t, \boldsymbol{a}_t)$. Consequently, it cannot be eliminated by increasing the capacity of a symmetric critic subject to this information constraint. The proof of Lemma 5.1 is provided in Appendix D.2.

The same decomposition applies to best-in-class approximation, i.e., when taking the infimum over the respective hypothesis classes. In particular, $\varepsilon_{\text{app},t}^{\tilde{\mathcal{F}}_T}(Q_t^\pi) = \varepsilon_{\text{app},t}^{\tilde{\mathcal{F}}_T}(\bar{Q}_t^\pi) + \varepsilon_{\text{alias},t}$, where $\varepsilon_{\text{app},t}^{\tilde{\mathcal{F}}_T}(\bar{Q}_t^\pi)$ is the symmetric critic's best-in-class error w.r.t. the conditional target $\bar{Q}_t^\pi$. Since $\tilde{\mathcal{F}}_T \subseteq \mathcal{F}_T$, $\varepsilon_{\text{app},t}^{\mathcal{F}_T}(Q_t^\pi) \leq \varepsilon_{\text{app},t}^{\tilde{\mathcal{F}}_T}(Q_t^\pi)$, even for $\varepsilon_{\text{alias},t} = 0$ (cf. Corollary E.1 in Appendix E.3).

At trajectory level, by Corollary E.3 in Appendix E.3, we have

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}_T} \mathbb{E}^\pi\left[\sum_{t<T} \gamma^t \left(\tilde{f}(\boldsymbol{h}_t, \boldsymbol{a}_t) - Q_t^\pi\right)^2\right] = \varepsilon_{\text{app}}^F(\tilde{\mathcal{F}}_T) + \sum_{t<T} \gamma^t \varepsilon_{\text{alias},t}, \quad (7)$$

where $\varepsilon_{\text{app}}^F(\tilde{\mathcal{F}}_T) := \inf_{\tilde{f} \in \tilde{\mathcal{F}}_T} \mathbb{E}^\pi\left[\sum_{t<T} \gamma^t \left(\tilde{f}(\boldsymbol{h}_t, \boldsymbol{a}_t) - \bar{Q}_t^\pi\right)^2\right]$. If $\bar{Q}_t^\pi$ is realizable in the symmetric class $\tilde{\mathcal{F}}_T$, the best-in-class trajectory-level approximation error of the symmetric critic w.r.t. Q_π^t is lower bounded by $\sum_{t<T} \gamma^t \varepsilon_{\text{alias},t}$. It is zero precisely when Q_t^π is conditionally deterministic given $(\boldsymbol{h}_t, \boldsymbol{a}_t)$. However, the informed class may retain an additional representation advantage beyond this information-induced effect.

5.2 Finite-time trade-off between aliasing and complexity

The preceding results characterize information and approximation effects independently of the finite-time learning costs of the critic. In the following, we therefore derive a finite-time and finite-width bound for Rec-TD with an informed critic and compare it with the corresponding symmetric Rec-TD bound established by Cayci & Eryilmaz (2025). This comparison yields a sufficient condition under which the information and approximation benefits of the informed critic lead to a tighter theoretical bound.

First, we present a regularity condition on the informed history-based Q-functions, which generalizes Assumption 5.3 in Cayci & Eryilmaz (2025) to the asymmetric setting and assumes that $(Q_t^\pi)_t$ lies in the RKHS induced by $\boldsymbol{\Psi}_T(\cdot; \vartheta_0)$.

Assumption 3 (Regularity of $(Q_t^\pi)_t$)**.** $\{Q_t^\pi : t \in \mathbb{Z}_{\geq 0}\} \in \mathcal{F}$ *with a transportation mapping* $v = (v_w, v_u, v_b) \in \mathcal{M}_F$ *such that* $\sup_{w \in \mathbb{R}} |v_w(w)| \leq \boldsymbol{\nu}_w$, $\sup_{\boldsymbol{u} \in \mathbb{R}^{d_x}} ||v_u(\boldsymbol{u})||_2 \leq \boldsymbol{\nu}_u$, *and* $\sup_{\boldsymbol{b} \in \mathbb{R}^{d_z}} ||v_b(\boldsymbol{b})||_2 \leq \boldsymbol{\nu}_b$.

Let $\omega_{k,t}^2 := \mathbb{E}\left[(\Phi_t(\boldsymbol{h}_{k,t}, \boldsymbol{a}_{k,t}, \boldsymbol{i}_{k,t}; \vartheta_k) - Q_{k,t}^\pi(\boldsymbol{h}_{k,t}, \boldsymbol{i}_{k,t}, \boldsymbol{a}_{k,t}))^2\right]$ denote the mean-squared error of the critic at time $t \in \mathbb{Z}_{\geq 0}$ and iteration $k \in \mathbb{Z}_{\geq 0}$.

Theorem 5.1 (Finite-time bound of asymmetric Rec-TD)**.** *Under Assumptions 1-3, for any projection radii $\boldsymbol{\rho} \succeq \boldsymbol{\nu} = (\nu_w \quad \nu_u \quad \nu_b)^\top$ and step size $\eta_{td} > 0$, Rec-TD with max-norm regularization achieves the following error bound:*

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

for any $K \in \mathbb{N}$, where $C_T^{(1)}, C_T^{(2)} = poly\left(\sum_{j=0}^{T-1} \left|\left(\varrho_1(\alpha + \frac{\boldsymbol{\rho}_w}{\sqrt{m_F}})\right)\right|^j, \|\boldsymbol{\rho}\|_2, \|\boldsymbol{\nu}\|_2\right)$ are instance-dependent constants that do not depend on K.

The proof of Theorem 5.1 can be found in Appendix D.1. The first two terms correspond to optimization and smoothness effects, respectively, while the third term captures the finite-width error of the neural network critic. Under the step-size choice $\eta_{\text{td}} = \mathcal{O}(K^{-1/2})$, the first two terms decay at rate $\mathcal{O}(K^{-1/2})$, while the finite-width contribution is independent of the number of optimization iterations K and decays as $\mathcal{O}(m_F^{-1/2})$. The fourth term arises from horizon truncation and depends on the iterate-dependent terminal mean-squared error. If Assumption 3 does not hold, the bound additionally contains a function-class approximation contribution of order $\mathcal{O}(\frac{1}{1-\gamma}\varepsilon^F_{\text{app}}(\mathcal{F}_T))$, where $\varepsilon^F_{\text{app}}(\mathcal{F}_T) := \inf_{f\in\mathcal{F}_T} \mathbb{E}^\pi \left[\sum_{t<T} \gamma^t \big(f(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t) - Q^\pi_t\big)^2\right]$.

Let $\bar{\omega}^2_{k,T} = \mathbb{E}\big[(\Phi^{\text{sym}}_T(\boldsymbol{h}_{k,T}, \boldsymbol{a}_{k,T}; \vartheta_k) - Q^\pi_{k,T})^2\big]$ and $\tilde{\omega}^2_{k,T} = \mathbb{E}\big[(\Phi^{\text{sym}}_T(\boldsymbol{h}_{k,T}, \boldsymbol{a}_{k,T}; \vartheta_k) - \bar{Q}^\pi_{k,T})^2\big]$. The approximation error decomposition in Lemma 5.1 gives $\bar{\omega}^2_{k,T} = \tilde{\omega}^2_{k,T} + \varepsilon_{\text{alias},k,T}$, where $\varepsilon_{\text{alias},k,T}$ denotes the information-aliasing error at time T and iteration k. Thus, the terminal error truncation contribution contains the additional information-induced term $\frac{\gamma^T}{(1-\gamma)K}\sum_{k<K}\varepsilon_{\text{alias},k,T}$, which is distinct from the trajectory-level approximation contribution in Equation 7, i.e., $\sum_{t<T}\gamma^t\varepsilon_{\text{alias},k,t}$.

To make the approximation-complexity trade-off explicit, let $\Delta L(K, m_F)$ denote the difference between the remaining finite-time terms in the informed and symmetric bounds (cf. Theorem 5.1 and Theorem F.1 in Appendix F.1), excluding the explicit information-aliasing contributions. Its full expression, including the corresponding norm, smoothness, Lipschitz, projection-radius, and finite-width constants, is provided in Appendix E.4. A positive $\Delta L(K, m_F)$ indicates a larger finite-time complexity contribution for the informed architecture under the corresponding bounds.

Proposition 5.1 (Sufficient condition for a tighter informed bound)**.** *Suppose that Q^π_t and $\bar{Q}^\pi_t$ are realizable in $\mathcal{F}_T$ and $\bar{\mathcal{F}}_T$, respectively, for all $0 \le t < T$. If the non-realizable extension of the finite-time bound includes an approximation contribution of the form $\frac{C_{app}}{1-\gamma}\varepsilon^F_{app}(\mathcal{F}_T)$ for some constant $C_{app} > 0$, then the informed critic has a smaller theoretical bound whenever*

$$C_{app}\sum_{t<T}\gamma^t\varepsilon_{alias,k,t} + \frac{\gamma^T}{K}\sum_{k<K}\varepsilon_{alias,k,T} > (1-\gamma)\Delta L(K, m_F) + \frac{\gamma^T}{K}\sum_{k<K}\left(\omega^2_{k,T} - \tilde{\omega}^2_{k,T}\right). \quad (8)$$

The condition exactly formalizes the trade-off: the information-induced reductions in approximation errors must compensate for the additional finite-time complexity and any difference in achievable terminal mean-squared errors. It is a sufficient theoretical comparison criterion, not a directly computable selection rule. In particular, the condition does not imply that conditional variance alone determines the benefit of privileged information; representation, approximation, and finite-time effects remain part of the overall comparison. The proof of Proposition 5.1 is deferred to Appendix E.4.

5.3 Effects of Informed Value Estimation on Rec-NPG

Finally, we analyze the effect of an informed critic on the Rec-NPG update. The actor remains restricted to the observable history $\boldsymbol{h}_t$, whereas the critic estimates Q^π_t. Consequently, the informed advantage may contain variation that cannot be represented by the actor's compatible features.

Cayci & Eryilmaz (2025) show that the best-iterate bound for Rec-NPG is governed by the compatible function approximation (CFA) error associated with the policy parameters θ and natural-gradient coefficients $\boldsymbol{q}$ (cf. Theorem F.2 in Appendix F.2). We generalize the symmetric CFA error and define the informed CFA error as $\varepsilon^T_{\text{cfa}}(\theta, \boldsymbol{q}) := \sum_{t<T}\gamma^t\left|\nabla^\top \log\pi^\theta_t(\boldsymbol{a}_t|\boldsymbol{h}_t)\,\boldsymbol{q} - A^{\pi_\theta}_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t)\right|^2$, where $A^{\pi_\theta}_t$ denotes the informed advantage function.

Proposition 5.2 (Error decomposition for $\varepsilon^T_{\text{cfa}}$)**.** *For any $n \in \mathbb{Z}_{\ge 0}$, we have*

$$\mathbb{E}\left[\mathbb{E}^{\pi_{\theta_n}}_{s_0\sim P}\left[\ell_T\left(\boldsymbol{q}_n; \theta_n, Q\right)\right]|\theta_k, k\le n\right] \le \frac{8\|\rho_G\|_2^2}{m_G}\sum_{t=0}^{T-1}\gamma^t\beta^2_{G,t} + 8\varepsilon^D_{app,n} + 6\varepsilon_{td,n} + 2\varepsilon_{sgd,n}$$

By Proposition 5.2, which generalizes Proposition 6.6 in Cayci & Eryilmaz (2025) to the informed setting, the CFA error can be bounded in terms of the actor's approximation error $\varepsilon^D_{\text{app},n}$, critic's statistical error $\varepsilon_{\text{td},n}$, the stochastic optimization error $\varepsilon_{\text{sgd},n}$, and a finite-width feature-drift term. The proof of Proposition 5.2 is provided in Appendix D.3.

The actor-side approximation error is given as

$$\varepsilon^D_{\text{app},n} := \inf_{\boldsymbol{q}\in\Omega_{\rho_G,m_G}} \mathbb{E}^{\pi_{\theta_n}}\left[\sum_{t<T}\gamma^t\left|\nabla^\top D_t(\boldsymbol{h}_t,\boldsymbol{a}_t;\theta_0)\boldsymbol{q} - Q_t^{\pi_{\theta_n}}(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)\right|^2\right].$$

Since the actor does not observe $\boldsymbol{i}_t$, the L^2-optimal target measurable w.r.t. $(\boldsymbol{h}_t,\boldsymbol{a}_t)$ is the conditional expectation $\bar{Q}_t^{\pi_{\theta_n}}$. Hence, Proposition E.2 in Appendix E.5 yields

$$\varepsilon^D_{\text{app},n} = \underbrace{\inf_{\boldsymbol{q}} \mathbb{E}^{\pi_{\theta_n}}\left[\sum_{t<T}\gamma^t|\nabla^\top D_t(\boldsymbol{h}_t,\boldsymbol{a}_t;\theta_0)\boldsymbol{q} - \bar{Q}_t^{\pi_{\theta_n}}|^2\right]}_{=\tilde{\varepsilon}^D_{\text{app},n}} + \underbrace{\sum_{t<T}\gamma^t\mathbb{E}^{\pi_{\theta_n}}\left[\mathrm{Var}\left(Q_t^{\pi_{\theta_n}}|\boldsymbol{h}_t,\boldsymbol{a}_t\right)\right]}_{=:\varepsilon_{\text{info},n}}.$$

This decomposition distinguishes two sources of actor approximation error. The first, $\tilde{\varepsilon}^D_{\text{app},n}$, reflects the expressive capacity of the actor's compatible features for the actor-observable target $\bar{Q}_t^\pi$, while the second, $\varepsilon_{\text{info},n}$, is induced by the critic's access to information unavailable to the actor. The latter is not merely an optimization error and cannot generally be eliminated by increasing the number of compatible-regression iterations. We refer to the non-negative term $\varepsilon_{\text{info},n}$ as information-mismatch error. It quantifies the variation in the informed value function that is unresolved by the actor's observable information and vanishes only when $Q_t^{\pi_{\theta_n}}$ is almost surely determined by $(\boldsymbol{h}_t,\boldsymbol{a}_t)$. By Proposition 5.2, this term enters the best-iterate bound for Rec-NPG through $\varepsilon^T_{\text{cfa}}(\theta,\boldsymbol{q})$.

This result is consistent with prior analysis of an asymmetric actor-critic method with a fixed agent state and linear function approximation (Lambrechts et al., 2025), which identifies increased residual error in the asymmetric compatible policy-gradient regression. Here, we interpret this residual as an actor-side information mismatch caused by the privileged dependence of the informed advantage.

The critic's statistical error inherits the finite-time behavior established in Theorem 5.1. In particular, for $\eta_{\text{td}} = \mathcal{O}(K_{\text{td}}^{-1/2})$, $\varepsilon_{\text{td},n} \le \text{poly}\left(\sum_{k=0}^{T-1}|\varrho_1\alpha_m|^k\right)\mathcal{O}\left(\frac{1}{\sqrt{K_{\text{td}}}} + \frac{1}{\sqrt{m_F}} + \gamma^T\right)$, which matches the asymptotic rate in the symmetric Rec-TD setting. For $\eta_{\text{sgd}} = \mathcal{O}(K_{\text{sgd}}^{-1/2})$, the stochastic optimization error in the policy update satisfies $\varepsilon_{\text{sgd},n} \le \text{poly}\left(\sum_{k=0}^{T-1}|\varrho_1^{(G)}\alpha_{m_G}|^k, \|\rho_G\|_2\right)\mathcal{O}\left(\sigma_D/\sqrt{K_{\text{sgd}}}\right)$, where σ_D denotes the second moment bound for the informed advantage target. Conditioning on the privileged signal can alter this constant because the additional information-dependent variation may increase the second moment of the stochastic gradients. The asymptotic rate, however, remains $\mathcal{O}(1/\sqrt{K_{\text{sgd}}})$, as in the symmetric setting. Hence, the statistical errors in the critic and the policy update, i.e., $\varepsilon_{\text{td},n}$ and $\varepsilon_{\text{sgd},n}$, can be made arbitrarily small by increasing K_{td}, K_{sgd}, and m_F.

Overall, privileged information can improve critic estimation through reduced conditional variance, a lower information-induced approximation floor, or a richer representation, at the expense of additional finite-time complexity. At the actor level, the same information can introduce a mismatch because the actor cannot condition on $\boldsymbol{i}_t$. Thus, the effect of privileged signals on Rec-NAC is governed by the balance between critic-side statistical and representation benefits, finite-time complexity, and the compatibility of the informed value target with the actor's observable features.

6 Conclusion

This work extends the Rec-NAC algorithm to asymmetric settings in which the critic leverages privileged state-dependent inputs during training. To this end, we propose a composed critic architecture that linearily combines RNN-based history representations with privileged signals. Our analysis shows that conditioning the critic on additional information can reduce the variance of the regression target and remove the information-aliasing error associated with the coarser observation set. However, it also introduces a richer hypothesis class, increasing estimation and optimization costs, and creates a mismatch with the actor's compatible features. More generally, our analysis highlights that improvements in value estimation do not automatically translate into uniformly tighter policy-learning guarantees when the critic has access to information unavailable to the actor.

Future work includes exploring nonlinear integrations of recurrent representations and privileged information, as well as developing principled criteria for selecting privileged signals that balance variance reduction, actor compatibility, and complexity costs. Another direction is to integrate strategies for mitigating the actor-critic mismatch, while preserving the variance-reduction benefits.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the clarity, grammar, and style of the manuscript. We have not used generative AI tools to generate research ideas, develop theoretical results or proofs, analyze results, or produce code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98), 2021.

J. Andrew Bagnell and Jeff Schneider. Covariant policy search. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 2003.

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

Bram Bakker. Reinforcement learning with long short-term memory. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Yang Cai, Xiangyu Liu, Argyris Oikonomou, and Kaiqing Zhang. Provable partially observable reinforcement learning with privileged information. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Semih Cayci and Atilla Eryilmaz. Recurrent natural policy gradient for POMDPs. *Transactions on Machine Learning Research*, 2025.

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

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), 2024b.

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), 2022.

Peter Dürr, Mireille Gheche, Guilherme Maeda, Nobuhiko Mukai, Naoya Takahashi, Stefan Heusser, Hamdi Sahloul, Yamen Saraiji, Pavel Adodin, Yin Bi, Sam Blakeman, Christian Conti, Dunai Hitos, Yunpu Hu, Farshad Khadivar, Raphaela Kreiser, Luz Martinez, Fabian Schilling, Ricardo Morales, and Michael Spranger. Outplaying elite table tennis players with an autonomous robot. *Nature*, 652, 2026.

Daniel Ebi, Damien Ernst, Klemens Böhm, and Gaspard Lambrechts. Informed asymmetric actor-critic: Leveraging privileged signals beyond full-state access. In *Forty-third International Conference on Machine Learning*, 2026.

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

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*, 2018.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2), 1998.

Sham Kakade. A natural policy gradient. In *Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 2001.

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

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

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

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.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POPGym: Benchmarking partially observable reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

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.

Satinder P. Singh, Tommi Jaakkola, and Michael I. Jordan. Learning without state-estimation in partially observable markovian decision processes. In *Machine Learning Proceedings 1994*. 1994.

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.

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.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020.

Daan Wierstra, Alexander Förster, Jan Peters, and Jürgen Schmidhuber. Recurrent policy gradients. *Logic Journal of IGPL*, 18(5), 2010.

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*, 2020.

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)*, 2016.

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

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 sequence of vectors is denoted by $\boldsymbol{x}_i$, the i-th element of a vector $\boldsymbol{x}$ is by x^i, the (i,j)-th entry of a matrix $\boldsymbol{X}$ by X^{ij}, 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 $\boldsymbol{x}$ and zeros elsewhere.

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., $\mathbb{E}_{\mathbf{y}|\boldsymbol{x}}[f(\mathbf{y})] = \mathbb{E}[f(\mathbf{y})|\mathbf{x} = \boldsymbol{x}]$.

B ALGORITHMIC TOOLS FOR ASYMMETRIC REC-NAC

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

B.1 INDEPENDENTLY RECURRENT NEURAL NETWORK (INDRNN)

An IndRNN of width $m \in \mathbb{N}$ is parameterized by a a recurrent weight vector $\boldsymbol{w} \in \mathbb{R}^m$, which defines the diagonal recurrent matrix $\boldsymbol{W} := \mathrm{diag}(\boldsymbol{w}) \in \mathbb{R}^{m \times m}$ 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. By construction, all off-diagonal entries of $\boldsymbol{W}$ are zero. Notably, IndRNNs with diagonal hidden-to-hidden weights $\boldsymbol{W}$ have been shown to effectively capture long-term dependencies (Morad et al., 2023) while offering theoretical advantages, as their diagonal structure enables explicit characterization of the reference function class and direct control of the spectral radius of $\boldsymbol{W}$, which are difficult to obtain for general recurrent matrices (Cayci & Eryilmaz, 2025).

Given an input $\boldsymbol{x}_t \in \mathbb{R}^{d_x}$ for $t \in \mathbb{Z}_{\geq 0}$, the hidden state $\boldsymbol{g}_t \in \mathbb{R}^m$ evolves recursively via

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

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 INDRNNS

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

Definition B.1 (Symmetric random initialization; cf. Definition 3.1 in (Cayci & Eryilmaz, 2025))**.** *Let the RNN's width m be even. For $i \in \{1, \ldots, \frac{m}{2}\}$, draw independently output weights $c_0^i \sim$ Unif$(-1, 1)$, recurrent weights $w_0^i \sim$ Unif$(-\alpha, \alpha)$, and input weights $\boldsymbol{U}_0^i \sim \mathcal{N}(0, \boldsymbol{I}_{d_x})$, and set*

$$c_0^{i+\frac{m}{2}} = -c^i,$$
$$w_0^{i+\frac{m}{2}} = w_0^i,$$
$$\boldsymbol{U}_0^{i+\frac{m}{2}} = \boldsymbol{U}_0^i.$$

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

B.3 MAX-NORM REGULARIZATION FOR INDRNNS

Given an IndRNN of width m followed by a linear output layer with symmetric random initialization $(\boldsymbol{w}_0, \boldsymbol{U}_0, \boldsymbol{c}_0)$, where $\boldsymbol{c}_0$ 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 $\Omega_{\tilde{\boldsymbol{\rho}},m} \subset \mathbb{R}^{m(1+d_x)}$ is defined as

$$\Omega_{\tilde{\boldsymbol{\rho}},m} = \mathcal{B}^{(m)}_{2,\infty}\left((\boldsymbol{w}_0, \boldsymbol{U}_0), \tilde{\boldsymbol{\rho}}\right),$$

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 Cayci & Eryilmaz (2025), we define $\mathcal{B}^{(m)}_{2,\infty}\left((\boldsymbol{w}, \boldsymbol{U}), \tilde{\boldsymbol{\rho}}\right)$ as

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

where $\otimes$ is the Cartesian product.

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

$$\max_{1\leq i\leq m} |w^i - w^i_0| \leq \frac{\tilde{\rho}_w}{\sqrt{m}},$$
$$\max_{1\leq i\leq m} \|\boldsymbol{U}^i - \boldsymbol{U}^i_0\| \leq \frac{\tilde{\rho}_u}{\sqrt{m}}.$$

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

$$\widetilde{\texttt{Proj}}_{\Omega_{\tilde{\boldsymbol{\rho}},m}}\left[(\boldsymbol{w}, \boldsymbol{U})\right] = \begin{bmatrix} \underset{\tilde{w}\in\mathcal{B}_2\left(w^i_0, \frac{\tilde{\rho}_w}{\sqrt{m}}\right)}{\arg\min} |w^i - \tilde{w}|, & \underset{\boldsymbol{u}\in\mathcal{B}_2\left(\boldsymbol{U}^i_0, \frac{\tilde{\rho}_u}{\sqrt{m}}\right)}{\arg\min} \|\boldsymbol{U}^i - \boldsymbol{u}\|_2 \end{bmatrix}_{i\in\{0,\ldots,m\}}.$$

Max-norm regularization for the informed critic. Given an informed critic with a composed architecture as described in Section 3, we define the compactly-supported set of weights

$$\Omega^{\vartheta}_{\boldsymbol{\rho},m_F,d_z} \in \mathbb{R}^{m_F(1+d_x)+d_z},$$

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

$$\Omega^{\vartheta}_{\boldsymbol{\rho},m_F,d_z} := \left(\mathcal{B}^{(m_F)}_{2,\infty}\left((\boldsymbol{w}_F, \boldsymbol{U}_F), \boldsymbol{\rho}_{\vartheta^F}\right), \mathcal{B}^{(d_z)}_2\left(\boldsymbol{b}, \tfrac{\rho_b}{\sqrt{d_z}}\right)\right).$$

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

$$\texttt{Proj}_{\Omega^{\vartheta}_{\boldsymbol{\rho},m_F,d_z}}\left[(\vartheta^F, \boldsymbol{b})\right] = \begin{bmatrix} \widetilde{\texttt{Proj}}_{\Omega_{\rho_{\vartheta^F}, m_F}}\left[\vartheta^F\right], & \underset{\mathrm{b}\in\mathcal{B}_2\left(\boldsymbol{b}, \frac{\rho_b}{\sqrt{d_z}}\right)}{\arg\min} \|\boldsymbol{b} - \mathrm{b}\|_2 \end{bmatrix},$$

where $\widetilde{\texttt{Proj}}_{\Omega_{\rho_{\vartheta^F}, m_F}}[\cdot]$ is the projection operator for the RNN weights.

B.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:

$$f(\boldsymbol{x}; \theta) \approx f(\boldsymbol{x}; \theta_0) + \langle \nabla_\theta f(\boldsymbol{x}; \theta_0), \theta - \theta_0 \rangle. \tag{9}$$

This linearization defines the time-independent neural tangent kernel (NTK) (Jacot et al., 2018), 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: $f(\boldsymbol{x}) = \mathbb{E}_{\theta_0} [\langle v(\theta_0), \psi(\boldsymbol{x}; \theta_0) \rangle]$, for some square-integrable function $v(\cdot)$ over initializations. This representation defines a nonparametric function class,

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

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.

Neural tangent random features of the informed critic. The NTRF feature map for the informed critic at time t is

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

where $\boldsymbol{\psi}_t^F$ coincides with the recurrent NTRF of the symmetric critic presented in Cayci & Eryilmaz (2025), and the additional block corresponds to the linear readout of the privileged embedding. In particular, $\boldsymbol{\psi}_t^F$ is given by

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

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 $((\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t))_{0 \le t \le T-1}$, stacking features $\boldsymbol{\psi}_t^\top(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \vartheta_0)$ across time yields the NTRF matrix

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

C Algorithmic Description of Rec-TD and Rec-NPG

Figure 1 illustrates the architectures of the actor and the informed critic neural network for the asymmetric variant of Rec-NAC. Section C.1 details how the symmetric critic is recovered as a special case of the proposed informed critic architecture.

For clarity and reproducibility, we provide pseudocode for the Rec-NAC algorithm and the informed Rec-TD learning loop used throughout the paper. Algorithm 1 provides a high-level description of Rec-NAC, while Algorithm 2 summarizes the corresponding Rec-TD update for the informed critic.

C.1 Symmetric Critic

The symmetric critic architecture is obtained from the asymmetric critic by removing access to the privileged signal $\boldsymbol{i}_t$. In this setting, the concatenation step before the linear readout disappears, i.e., $\boldsymbol{z}_t = \boldsymbol{f}_t$, and the dimensionality of $\boldsymbol{z}_t$ and the output weights $\tilde{\boldsymbol{b}}$ reduces to $d_z^{\mathrm{sym}} = m_F$. All other components remain identical to those of the asymmetric critic: the IndRNN dynamics, the input $\boldsymbol{x}_t$,

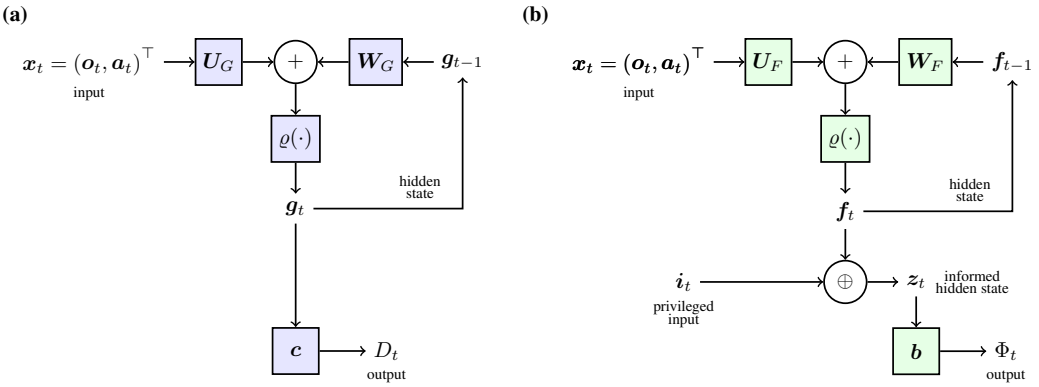


Figure 1: Network architecture of (a) the actor and (b) the informed critic.

Algorithm 1: Recurrent Natural Actor-Critic (Rec-NAC)

Initialize actor network $D_t(\cdot;\theta_0)$ via symmetric random initialization.
for $n = 0$ **to** $N-1$ **do**
  Critic: Obtain $\overline{Q}^{\pi_{\theta_n}}$ using Algorithm 2.
  Actor: Compute the gradient $\nabla_{\boldsymbol{q}} \ell_T \left(\boldsymbol{q}_n^k; \theta_n, \overline{Q}_k^{\pi_{\theta_n}} \right)$.
  Policy Update:
  Compute $\hat{\boldsymbol{q}}_{n,k+1}$ using Eq. 4.
  Update policy parameters $\theta_{(n+1)}$ using Eq. 5.
end
Output: Final policy π_{θ_N}.

Algorithm 2: Recurrent TD (Rec-TD) Learning Algorithm with Informed Critic

Input: policy π, bootstrap timestep L, step size η_{td}, number of updates K, and projection radius $\boldsymbol{\rho}$.
Initialize critic network $\Phi(\cdot;\vartheta^F, \boldsymbol{b})$ via symmetric random initialization.
for $k = 0$ **to** $K-1$ **do**
  Sample an initial state $\boldsymbol{s}_{k,0} \sim P$ independently.
  Get $\boldsymbol{i}_{k,0} \sim I(\cdot|\boldsymbol{s}_{k,0})$.
  Get $\boldsymbol{o}_{k,0} \sim \widetilde{O}(\cdot|\boldsymbol{i}_{k,0})$.
  Sample an action $\boldsymbol{a}_{k,0} \sim \pi_0(\cdot|\boldsymbol{h}_{k,0})$.
  Set $\check{\nabla}\mathcal{R}_T^k := 0$.
  for $l = 0$ **to** L **do**
    Transition to environment state $\boldsymbol{s}_{k,l+1} \sim T(\cdot|\boldsymbol{s}_{k,l}, \boldsymbol{a}_{k,l})$.
    Get $\boldsymbol{i}_{k,l+1} \sim I(\cdot|\boldsymbol{s}_{k,l+1})$.
    Get $\boldsymbol{o}_{k,l+1} \sim \widetilde{O}(\cdot|\boldsymbol{i}_{k,l+1})$.
    Sample an action $\boldsymbol{a}_{k,l+1} \sim \pi_\theta(\cdot|\boldsymbol{h}_{k,l+1})$.
    Compute $\delta_l := r_l + \gamma\Phi_{l+1}(\boldsymbol{h}_{k,l+1}, \boldsymbol{a}_{k,l+1}, \boldsymbol{i}_{k,l+1}; \vartheta_k) - \Phi_{l+1}(\boldsymbol{h}_{k,l}, \boldsymbol{a}_{k,l}, \boldsymbol{i}_{k,l}; \vartheta_k)$
    Compute $\check{\nabla}\mathcal{R}_L(\boldsymbol{h}_{k,L}, \boldsymbol{a}_{k,L}, \boldsymbol{i}_{k,L}; \vartheta_k)$ according to Eq. 2.
  end
  Compute ϑ_{k+1} using Eq. 3.
end
Output: Avg. estimate $\overline{Q}^\pi(\cdot) = \widehat{Q}_{\hat{\vartheta}}^{\pi_\theta}(\cdot)$ with $\hat{\vartheta} = \frac{1}{K}\sum_{k=0}^{K-1} \vartheta_k$.

the input weight matrix $\tilde{\boldsymbol{U}}_F$, and the diagonal recurrence matrix $\tilde{\boldsymbol{W}}_F$. Consequently, the symmetric critic output is $\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \tilde{\vartheta}) := \frac{1}{\sqrt{m_F}}\langle \tilde{\boldsymbol{b}}, \boldsymbol{f}_t \rangle$.

This architecture coincides with the symmetric critic introduced by Cayci & Eryilmaz (2025), except that they fixed the readout weights after initialization to isolate the effect of the recurrent dynamics in the theoretical analysis. We generalize the design by allowing $\tilde{\boldsymbol{b}}$ to be learned jointly with the recurrent parameters, which makes it directly comparable to the informed critic and enables a unified finite-width and finite-time analysis for both settings.

D PROOFS

This section collects the proofs of our results presented in Section 5.

Section D.1 presents the proof of Theorem 5.1 and Section D.2 the proof of Lemma 5.1. Section D.3 contains the proof of Proposition 5.2.

D.1 PROOF OF THEOREM 5.1 (FINITE-TIME BOUND OF ASYMMETRIC REC-TD)

The following proof closely follows the structure of the proof of Theorem 5.4 in (Cayci & Eryilmaz, 2025) (cf. Theorem F.1), with adaptations tailored to the informed critic's design presented in this work. For improved readability, throughout the proof we write, for example, $\boldsymbol{h}_t^k$ instead of $\boldsymbol{h}_{k,t}$.

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:

$$\Psi(\vartheta) = \|\vartheta - \bar{\vartheta}\|_2^2. \tag{10}$$

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

$$\begin{aligned}\Psi(\vartheta_{k+1}) \leq \Psi(\vartheta_k) &+ 2\eta \sum_{t=0}^{T-1} \gamma^t \delta_t \left(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k\right) \left\langle \nabla \Phi_t \left(F_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k; \vartheta_k^{\mathrm{F}}), \boldsymbol{i}_t^k; \boldsymbol{b}_k\right), \vartheta_k - \bar{\vartheta} \right\rangle \\ &+ \eta^2 \left\| \check{\nabla} \mathcal{R}_T \left(\boldsymbol{h}_T^k, \boldsymbol{a}_T^k, \boldsymbol{i}_T^k; \vartheta_k\right) \right\|_2^2.\end{aligned} \tag{11}$$

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

$$\begin{aligned}&\mathbb{E}[\Psi(\vartheta_{k+1}) - \Psi(\vartheta_k)] \\ &\qquad \leq 2\eta \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \underbrace{\check{\mathbb{E}}_t^k \left[\delta_t \left(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k\right)\right] \cdot \left\langle \nabla \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k),\ \vartheta_k - \bar{\vartheta} \right\rangle}_{(\boldsymbol{\Delta})_t}\right] \\ &\qquad\quad + \eta^2 \underbrace{\mathbb{E}\left[\left\| \check{\nabla} \mathcal{R}_T \left(\boldsymbol{h}_T^k, \boldsymbol{a}_T^k, \boldsymbol{i}_T^k; \vartheta_k\right)\right\|_2^2\right]}_{(\heartsuit)}\end{aligned} \tag{12}$$

To bound $\mathbb{E}(\boldsymbol{\Delta})_t$, we first apply the non-stationary Bellman equation (cf. Proposition E.3) and adapt it to the asymmetric setting:

$$\begin{aligned}\check{\mathbb{E}}_t^k \left[\delta_t \left(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k\right)\right] &= \check{\mathbb{E}}_t^k \left[r_t^k + \gamma \cdot \Phi_{t+1}(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, i_{t+1}^k; \vartheta_k)\right] - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k) \\ &= \gamma \cdot \check{\mathbb{E}}_t^k \left[\Phi_{t+1}(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k) - Q_{t+1}^{\pi_\theta}\left(\boldsymbol{h}_{t+1}^k, \boldsymbol{i}_{t+1}^k, \boldsymbol{a}_{t+1}^k\right)\right] \\ &\quad + Q_t^{\pi_\theta}\left(\boldsymbol{h}_t^k, \boldsymbol{i}_t^k, \boldsymbol{a}_t^k\right) - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k).\end{aligned} \tag{13}$$

We can reformulate the inner product using reparameterized features:

$$\left\langle \nabla \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k), \vartheta_k - \bar{\vartheta} \right\rangle = \left\langle \nabla \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_0), \vartheta_k - \bar{\vartheta} \right\rangle + \mathrm{err}_{k,t}^{(1)}, \tag{14}$$

where the residual term is

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

Using Lemma E.2, this error satisfies:

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

We can further decompose the inner product as follows:

$$\begin{aligned}\left\langle \nabla\Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_0),\ \vartheta_k - \bar{\vartheta}\right\rangle &= \Phi_t^{\mathrm{Lin}}(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k) - \Phi_t^{\mathrm{Lin}}(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \bar{\vartheta}) &(15)\\ &= \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k) - Q_t^{\pi_\theta}(\boldsymbol{h}_t^k, \boldsymbol{i}_t^k, \boldsymbol{a}_t^k) + \mathrm{err}_{k,t}^{(2)} + \mathrm{err}_{k,t}^{(3)}. &(16)\end{aligned}$$

The error terms are:

$$\mathrm{err}_{k,t}^{(2)} := \Phi_t^{\mathrm{Lin}}(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k) - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k), \tag{17}$$

$$\mathrm{err}_{k,t}^{(3)} := -\Phi_t^{\mathrm{Lin}}(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \bar{\vartheta}) + Q_t^{\pi_\theta}(\boldsymbol{h}_t^k, \boldsymbol{i}_t^k, \boldsymbol{a}_t^k). \tag{18}$$

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

$$\begin{aligned}(\boldsymbol{\Delta})_t = &-\left(Q_t^{\pi_\theta}(\boldsymbol{h}_t^k, \boldsymbol{i}_t^k, \boldsymbol{a}_t^k) - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k)\right)^2\\ &+ \gamma \cdot \check{\mathbb{E}}_t^k\left[\Phi_{t+1}(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k) - Q_{t+1}^{\pi_\theta}(\boldsymbol{h}_{t+1}^k, \boldsymbol{i}_{t+1}^k, \boldsymbol{a}_{t+1}^k)\right]\\ &\quad \cdot \left(Q_t^{\pi_\theta}(\boldsymbol{h}_t^k, \boldsymbol{i}_t^k, \boldsymbol{a}_t^k) - \Phi_t(\boldsymbol{h}_t^k, \boldsymbol{a}_t^k, \boldsymbol{i}_t^k; \vartheta_k)\right)\\ &+ \check{\mathbb{E}}_t^k\left[\delta_t(\boldsymbol{h}_{t+1}^k, \boldsymbol{a}_{t+1}^k, \boldsymbol{i}_{t+1}^k; \vartheta_k)\right] \cdot \left(\mathrm{err}_{k,t}^{(1)} + \mathrm{err}_{k,t}^{(2)} + \mathrm{err}_{k,t}^{(3)}\right)\end{aligned} \tag{19}$$

By Assumption 1 and Lemma E.2, we bound the temporal differences by:

$$\sup_{\boldsymbol{h},\boldsymbol{a},\boldsymbol{i}\in\mathbb{H}_\infty} |\delta_t(\boldsymbol{h}_{t+1}, \boldsymbol{a}_{t+1}, \boldsymbol{i}_{t+1}; \vartheta_k)| \leq r_{\max} + 2\rho_b \cdot \left(L_T^F \cdot \|\boldsymbol{\rho}\|_2 + 1\right) =: \delta_{\max}. \tag{20}$$

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

$$\mathbb{E}\left[(\boldsymbol{\Delta})_t\right] \leq -\omega_{k,t}^2 + \gamma\omega_{k,t+1}\,\omega_{k,t} + \delta_{\max}\sum_{j=1}^{3}\mathbb{E}\left|\mathrm{err}_{k,t}^{(j)}\right|. \tag{21}$$

Applying bounds from Lemmas E.2 and E.3, we obtain:

$$\mathbb{E}\left|\mathrm{err}_{k,t}^{(2)}\right| \leq \frac{2}{\sqrt{m_F}}\left(\varrho_2(\Lambda_t^F)^2 + \varrho_1\chi_t^F\right)\|\boldsymbol{\rho}_{\vartheta^F}\|_2^2 \tag{22}$$

and

$$\mathbb{E}\left|\mathrm{err}_{k,t}^{(3)}\right| \leq \sqrt{\mathbb{E}\left|\mathrm{err}_{k,t}^{(3)}\right|^2} \leq \frac{2\|\boldsymbol{\nu}\|_2\sqrt{1+\varrho_0^2}\cdot p_T(\alpha\varrho_1)}{\sqrt{m_F}}. \tag{23}$$

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

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

Using this inequality, we can derive the following bound for any time step $0 \leq t < T$:

$$\mathbb{E}\left[(\boldsymbol{\Delta})_t\right] \leq -\omega_{k,t}^2 + \frac{\gamma}{2}\left(\omega_{k,t+1}^2 + \omega_{k,t}^2\right) + \delta_{\max}\cdot\frac{c_T}{\sqrt{m_F}} \tag{24}$$

where c_T is defined as

$$c_T := 2(\beta_T^F)^2\|\boldsymbol{\rho}_{\vartheta^F}\|_2^2 + \left(\varrho_2(\Lambda_t^F)^2 + \varrho_1\chi_t^F\right)\|\boldsymbol{\rho}_{\vartheta^F}\|_2^2 + 2\|\boldsymbol{\nu}\|_2\sqrt{1+\varrho_0^2}\cdot p_T(\alpha\varrho_1). \tag{25}$$

Hence, we obtain the following upper bound:

$$\sum_{t=0}^{T-1}\gamma^t\mathbb{E}\left[(\boldsymbol{\Delta})_t\right] \leq -(1-\frac{\gamma}{2})\sum_{t=0}^{T-1}\gamma^t\omega_{k,t}^2 + \frac{\delta_{\max}\cdot c_T}{(1-\gamma)\sqrt{m_F}} + \underbrace{\frac{1}{2}\sum_{t=0}^{T-1}\gamma^{t+1}\omega_{k,t+1}^2}_{\leq\frac{1}{2}\left(\sum_{t=0}^{T-1}\gamma^t\omega_{k,t}^2+\gamma^T\omega_{k,T}^2\right)} \tag{26}$$

$$\leq -\frac{1-\gamma}{2}\sum_{t=0}^{T-1}\gamma^t\omega_{k,t}^2 + \frac{1}{2}\gamma^T\omega_{k,T}^2 + \frac{c_T\cdot\delta_{\max}}{(1-\gamma)\sqrt{m_F}}.$$

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

$$\left\|\sum_{t=0}^{T-1}\gamma^t\delta_t\left(\boldsymbol{h}_{t+1}^k,\boldsymbol{a}_{t+1}^k,\boldsymbol{i}_{t+1}^k;\vartheta_k\right)\nabla\Phi_t(\boldsymbol{h}_t^k,\boldsymbol{a}_t^k,\boldsymbol{i}_t^k;\vartheta_k)\right\|_2$$
$$\leq \sum_{t=0}^{T-1}\gamma^t\left|\delta_t\left(\boldsymbol{h}_{t+1}^k,\boldsymbol{a}_{t+1}^k,\boldsymbol{i}_{t+1}^k;\vartheta_k\right)\right|\cdot\left\|\nabla\Phi_t(\boldsymbol{h}_t^k,\boldsymbol{a}_t^k,\boldsymbol{i}_t^k;\vartheta_k)\right\|_2 \tag{27}$$

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

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

and

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

where $L_T' = \left((L_T^F)^2+\frac{1}{d_z}\left(L_T^F\|\boldsymbol{\rho}_{\vartheta^F}\|_2+1\right)^2\right)^{\frac{1}{2}}$ for every $t<T$ with probability 1, according to Lemma E.2.

Therefore, we obtain the upper bound:

$$\left\|\check{\nabla}\mathcal{R}_T\left(\boldsymbol{h}_t^k,\boldsymbol{a}_t^k,\boldsymbol{i}_t^k;\vartheta_k\right)\right\|_2 \leq \frac{\delta_{\max}L_T'}{1-\gamma}. \tag{28}$$

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 12), and substituting in the bounds obtained above (cf. Eq. 26 and Eq. 28), yields:

$$\mathbb{E}[\Psi(\vartheta_{k+1})-\Psi(\vartheta_k)] \leq -\eta(1-\gamma)\sum_{t=0}^{T-1}\gamma^t\omega_{k,t}^2 + \eta\gamma^T\omega_{k,T}^2 + \eta\frac{c_T\delta_{\max}}{(1-\gamma)\sqrt{m_F}} + \eta^2\frac{\delta_{\max}^2\left(L_T'\right)^2}{(1-\gamma)^2},$$

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 (Cayci & Eryilmaz, 2025), results in:

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

This concludes the proof. □

Corollary D.1. *Choosing the step size $\eta_{td}=\mathcal{O}(\frac{1}{\sqrt{K}})$ gives the following bound:*

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

D.2 Proof of Lemma 5.1 (Decomposition of the symmetric critic's error w.r.t. Q_t^{π})

Proof. We start from $\tilde{\omega}_{k,t}^2=\mathbb{E}\left[(\Phi_t^{\mathrm{sym}}(\boldsymbol{h}_t,\boldsymbol{a}_t;\vartheta)-\bar{Q}_t^{\pi})^2\right]$. Using the identity

$$\Phi_t^{\mathrm{sym}}(\boldsymbol{h}_t,\boldsymbol{a}_t;\vartheta)-\bar{Q}_t^{\pi} = \Phi_t^{\mathrm{sym}}(\boldsymbol{h}_t,\boldsymbol{a}_t;\vartheta)-Q_t^{\pi}+Q_t^{\pi}-\bar{Q}_t^{\pi} = \left(\Phi_t^{\mathrm{sym}}(\boldsymbol{h}_t,\boldsymbol{a}_t;\vartheta)-Q_t^{\pi}\right)+\left(Q_t^{\pi}-\bar{Q}_t^{\pi}\right),$$

we obtain

$$\tilde{\omega}_{k,t}^2 = \mathbb{E}\left[\left((\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi) + (Q_t^\pi - \bar{Q}_t^\pi)\right)^2\right]. \tag{29}$$

Expanding the right-hand side in Eq. 29 yields

$$\begin{aligned}\tilde{\omega}_{k,t}^2 = \mathbb{E}&\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right)^2\right] \\ &+ 2\mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right)\left(Q_t^\pi - \bar{Q}_t^\pi\right)\right] \\ &+ \mathbb{E}\left[\left(Q_t^\pi - \bar{Q}_t^\pi\right)^2\right].\end{aligned} \tag{30}$$

First, we consider the cross term in Eq. 30, i.e., $\mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right)\left(Q_t^\pi - \bar{Q}_t^\pi\right)\right]$. Rewrite

$$\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right) = \left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - \bar{Q}_t^\pi\right) - \left(Q_t^\pi - \bar{Q}_t^\pi\right),$$

so we obtain

$$\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right)\left(Q_t^\pi - \bar{Q}_t^\pi\right) = \left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - \bar{Q}_t^\pi\right)\left(Q_t^\pi - \bar{Q}_t^\pi\right) - \left(Q_t^\pi - \bar{Q}_t^\pi\right)^2.$$

Taking expectations yields

$$\begin{aligned}\mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right)\left(Q_t^\pi - \bar{Q}_t^\pi\right)\right] = \mathbb{E}&\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - \bar{Q}_t^\pi\right)\left(Q_t^\pi - \bar{Q}_t^\pi\right)\right] \\ &- \mathbb{E}\left[\left(Q_t^\pi - \bar{Q}_t^\pi\right)^2\right].\end{aligned} \tag{31}$$

By applying the law of iterated expectations, we have

$$\mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - \bar{Q}_t^\pi\right)\left(Q_t^\pi - \bar{Q}_t^\pi\right)\right] = \mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - \bar{Q}_t^\pi\right)\mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t,\boldsymbol{a}_t}\left[\left(Q_t^\pi - \bar{Q}_t^\pi\right)\right]\right],$$

since Φ_t^{sym} and $\bar{Q}_t^\pi$ are functions of $(\boldsymbol{h}_t, \boldsymbol{a}_t)$.

By definition of $\bar{Q}_t^\pi$),

$$\mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t,\boldsymbol{a}_t}\left[\left(Q_t^\pi - \bar{Q}_t^\pi\right)\right] = 0.$$

Thus, we find

$$\mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - \bar{Q}_t^\pi\right)\left(Q_t^\pi - \bar{Q}_t^\pi\right)\right] = 0.$$

Therefore, Eq. 31 reduces to

$$\mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right)\left(Q_t^\pi - \bar{Q}_t^\pi\right)\right] = -\mathbb{E}\left[\left(Q_t^\pi - \bar{Q}_t^\pi\right)^2\right]. \tag{32}$$

Substituting this result back into Eq. 30 yields

$$\begin{aligned}\tilde{\omega}_{k,t}^2 &= \mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right)^2\right] - 2\mathbb{E}\left[\left(Q_t^\pi - \bar{Q}_t^\pi\right)^2\right] + \mathbb{E}\left[\left(Q_t^\pi - \bar{Q}_t^\pi\right)^2\right] \\ &= \mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right)^2\right] - \mathbb{E}\left[\left(Q_t^\pi - \bar{Q}_t^\pi\right)^2\right].\end{aligned} \tag{33}$$

By the definition of conditional variance, we have

$$\text{Var}(Q_t^\pi|\boldsymbol{h}_t, \boldsymbol{a}_t) = \mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t,\boldsymbol{a}_t}\left[\left(Q_t^\pi - \bar{Q}_t^\pi\right)^2\right].$$

Taking expectations, i.e., $\mathbb{E}\left[\left(Q_t^\pi - \bar{Q}_t^\pi\right)^2\right] = \mathbb{E}\big[\text{Var}(Q_t^\pi|\boldsymbol{h}_t, \boldsymbol{a}_t)\big]$, yields

$$\begin{aligned}\tilde{\omega}_{k,t}^2 &= \mathbb{E}\left[(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - \bar{Q}_t^\pi)^2\right] \\ &= \mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right)^2\right] - \mathbb{E}\left[\text{Var}(Q_t^\pi|\boldsymbol{h}_t, \boldsymbol{a}_t)\right].\end{aligned} \tag{34}$$

Rearranging terms in Eq. 32 by adding $\mathbb{E}\left[\text{Var}(Q_t^\pi|\boldsymbol{h}_t, \boldsymbol{a}_t)\right]$ to both sides yields

$$\mathbb{E}\left[\left(\Phi_t^{\text{sym}}(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta) - Q_t^\pi\right)^2\right] = \tilde{\omega}_{k,t}^2 + \mathbb{E}\left[\text{Var}(Q_t^\pi|\boldsymbol{h}_t, \boldsymbol{a}_t)\right] \tag{35}$$

This concludes the proof. □

D.3 Proof of Proposition 5.2 (Error decomposition for $\varepsilon_{\text{CFA}}^{T}$)

Let ε_{app} denote the actor's approximation error,

$$\varepsilon_{\text{app},n}^{D} = \inf_{\boldsymbol{q}\in\mathcal{B}_{2,\infty}^{(m_G)}(0,\boldsymbol{\rho}_G)} \mathbb{E}_{s_0\sim P}^{\pi_{\theta_n}} \sum_{t=0}^{T-1} \gamma^t \Big| \nabla^\top D_t\left(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0\right) \boldsymbol{q} - Q_t^{\pi_{\theta_n}}\left(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t\right) \Big|^2,$$

ε_{td} be the error of the critic,

$$\varepsilon_{\text{td},n} = \mathbb{E}\left[\mathcal{R}_T^{\pi_{\theta_n}}\left(\bar{\vartheta}_n\right) | \theta_k, k \le n\right],$$

and ε_{sgd} be the error in the policy update based on the compatible function approximation (CFA),

$$\varepsilon_{\text{sgd},n} = \mathbb{E}\left[\ell_T\left(\boldsymbol{q}_n; \theta_n, \hat{Q}\right) \Big| \bar{\theta}_n, \vartheta_k, k \le n\right] - \inf_{\boldsymbol{q}} \mathbb{E}\left[\ell_T\left(\boldsymbol{q}; \theta_n, \hat{Q}\right) | \bar{\vartheta}_n, \theta_k, k \le n\right].$$

Proof of Proposition 5.2.

Proof. For any $\boldsymbol{q}$, we have

$$\ell_T(\boldsymbol{q}; \theta_n, Q) \le 2\ell_T(\boldsymbol{q}; \theta_n, \hat{Q}) + 2\sum_{t=0}^{\infty} \gamma^t \left(A_t^{\pi_{\theta_n}}(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) - \hat{A}_t^{\pi_{\theta_n}}(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t)\right)^2. \tag{36}$$

Since

$$\varepsilon_{\text{sgd},n} = \mathbb{E}[\ell_T(\boldsymbol{q}_n; \theta_n, \hat{Q}) | \vartheta_n, \theta_k, k \le n] - \inf_{\boldsymbol{q}\in\mathcal{B}_{2,\infty}^{(m_G)}(0,\boldsymbol{\rho}_G)} \mathbb{E}[\ell_T(\boldsymbol{q}; \theta_n, \hat{Q}) | \vartheta_n, \theta_k, k \le n], \tag{37}$$

we obtain

$$\begin{aligned}\mathbb{E}[\ell_T(\boldsymbol{q}_n; \theta_n, Q | \vartheta_n, \theta_k, k \le n)] \le 2\mathbb{E}\left[\inf_{\boldsymbol{q}} \mathbb{E}[\ell_T(\boldsymbol{q}; \theta_n, \hat{Q}) | \vartheta_n, \theta_k, k \le n] \Big| \theta_k, k \le n\right] \\ + 2(\varepsilon_{\text{td},n} + \varepsilon_{\text{sgd},n}),\end{aligned} \tag{38}$$

where we use $Var(\boldsymbol{x}|\theta_k, k \le n) \le \mathbb{E}\left[|\boldsymbol{x}|^2 | \theta_k, k \le n\right]$ for any square-integrable $\boldsymbol{x}$. Moreover,

$$\begin{aligned}\inf_{\boldsymbol{q}} \mathbb{E}\left[\ell_T(\boldsymbol{q}; \theta_n, \hat{Q}) | \vartheta_n, \theta_k, k \le n\right] \le 2 \inf_{\boldsymbol{q}} \mathbb{E}\left[\ell_T(\boldsymbol{q}; \theta_n, Q) | \vartheta_n, \theta_k, k \le n\right] \\ + 2\sum_{t=0}^{\infty} \gamma^t \left(A_t^{\pi_{\theta_n}}(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) - \hat{A}_t^{\pi_{\theta_n}}(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t)\right)^2.\end{aligned} \tag{39}$$

Taking the corresponding conditional expectation and using the definition of $\varepsilon_{\text{td},n}$ gives

$$\begin{aligned}&\mathbb{E}\left[\inf_{\boldsymbol{q}} \mathbb{E}[\ell_T(\boldsymbol{q}; \theta_n, \hat{Q}) | \vartheta_n, \theta_k, k \le n \Big| \theta_k, k \le n\right] \\ &\le 2\mathbb{E}\left[\inf_{\boldsymbol{q}} \mathbb{E}[\ell_T(\boldsymbol{q}; \theta_n, Q) | \vartheta_n, \theta_k, k \le n \Big| \theta_k, k \le n\right] + 2\varepsilon_{\text{td},n}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[\ell_T(\boldsymbol{q}_n; \theta_n, Q) | \vartheta_n, \theta_k, k \le n] \le 4\mathbb{E}\left[\inf_{\boldsymbol{q}} \mathbb{E}[\ell_T(\boldsymbol{q}; \theta_n, Q) | \vartheta_n, \theta_k, k \le n \Big| \theta_k, k \le n\right] \\ + 6\varepsilon_{\text{td},n} + 2\varepsilon_{\text{sgd},n}.\end{aligned} \tag{40}$$

We next relate the advantage-based CFA loss to the Q-function approximation error, i.e.,

$$\begin{aligned}&\mathbb{E}\Big[\left(\nabla_\theta^\top \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\boldsymbol{q} - A_t^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t)\right)^2 \Big| \boldsymbol{h}_t, \boldsymbol{i}_t\Big] \\ &\quad= \mathbb{E}\left[\left(\nabla_\theta^\top \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\boldsymbol{q} - Q_t + V_t\right)^2 \Big| \boldsymbol{h}_t, \boldsymbol{i}_t\right] \\ &\quad= \mathbb{E}\left[\left(\nabla_\theta^\top \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\boldsymbol{q} - Q_t\right)^2 \Big| \boldsymbol{h}_t, \boldsymbol{i}_t\right] \\ &\qquad+ 2V_t\mathbb{E}\left[\nabla_\theta^\top \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\boldsymbol{q} - Q_t \Big| \boldsymbol{h}_t, \boldsymbol{i}_t\right] + V_t^2 \\ &\quad= \mathbb{E}\left[\left(\nabla_\theta^\top \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\boldsymbol{q} - Q_t\right)^2 \Big| \boldsymbol{h}_t, \boldsymbol{i}_t\right] - V_t^2 \\ &\quad\le \mathbb{E}\left[\left(\nabla_\theta^\top \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\boldsymbol{q} - Q_t\right)^2 \Big| \boldsymbol{h}_t, \boldsymbol{i}_t\right],\end{aligned}$$

since $\mathbb{E}\left[\nabla_\theta^\top \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\boldsymbol{q}|\boldsymbol{h}_t, \boldsymbol{i}_t\right] = 0$ and

$$V_t := \mathbb{E}_{\boldsymbol{a}_t\sim\pi(\cdot|\boldsymbol{h}_t)}[Q_t^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t)|\boldsymbol{h}_t, \boldsymbol{i}_t] = V_t^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t).$$

Therefore, for any $\boldsymbol{q} \in \mathcal{B}_{2,\infty}^{(m_G)}(0, \boldsymbol{\rho}_G)$,

$$\begin{aligned}\mathbb{E}[\ell_T(\boldsymbol{q}; \theta_n, Q)|\vartheta_n, \theta_k, k \le n] &\le \mathbb{E}\left[\sum_{t<T} \gamma^t \left(\nabla_\theta^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_n)\boldsymbol{q} - Q_t^{\pi_{\theta_n}}\right)^2 |\vartheta_n, \theta_k, k \le n\right], \\ &\le 2\mathbb{E}\Bigg[\sum_{t<T} \gamma^t \left(\nabla_\theta^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - Q_t^{\pi_{\theta_n}}\right)^2 \\ &\qquad + \left(\left[\nabla D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_n) - \nabla D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\right]^\top \boldsymbol{q}\right)^2 |\vartheta_n, \theta_k, k \le n\Bigg].\end{aligned}$$

Using $\|\nabla D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_n) - \nabla D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\|_2 \le \frac{\beta_{G,t}\|\rho_G\|_2}{\sqrt{m_G}}$, $t \in \mathbb{N}$, from the proof of Theorem 6.3 in (Cayci & Eryilmaz, 2025), we obtain

$$\begin{aligned}&\inf_{\boldsymbol{q}} \mathbb{E}[\ell_T(\boldsymbol{q}; \theta_n, Q|\vartheta_n, \theta_k, k \le n)] \\ &\quad \le 2\varepsilon_{\text{app},n} + 2\|\rho_G\|_2^2 \,\mathbb{E}\left[\sum_{t<T} \gamma^t \|\nabla D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_n) - \nabla D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\|_2^2 \Big| \vartheta_n, \theta_k, k \le n\right] \\ &\quad \le 2\varepsilon_{\text{app},n} + \frac{2\|\rho_G\|_2^4}{m_G} \sum_{t<T} \gamma^t \beta_{G,t}^2.\end{aligned}$$

Substituting this bound into Equation 40 yields

$$\mathbb{E}[\ell_T(\boldsymbol{q}_n; \theta_n, Q|\vartheta_n, \theta_k, k \le n)] \le \frac{8\|\rho_G\|_2^4}{m_G} \sum_{t<T} \gamma^t \beta_{G,t}^2 + 8\varepsilon_{\text{app},n} + 6\varepsilon_{\text{td},n} + 2\varepsilon_{\text{sgd},n}. \tag{41}$$

This concludes the proof. □

E Auxiliary Results

This section collects our auxiliary results.

Section E.2 presents the unbiasedness result for the informed advantage (Lemma E.1 and Section E.2 the conditional variance-reduction property w.r.t. TD targets (Proposition E.1). Section E.3 characterizes the error floors of the informed and symmetric critics (Corollaries E.1-E.3). Section E.4 compares the finite-time bounds of the informed and symmetric critics, and contains the proof of Proposition 5.1. Section E.5 presents the decomposition of the actor's approximation error arising from the use of an informed critic (Proposition E.2). Section E.6 contains the local continuity properties of the informed critic (Lemma E.2), Section E.7 the approximation error between NTRF and NTK in the informed critic (Lemma E.3), and Section E.8 the non-stationary Bellman equation in the asymmetric setting (Proposition E.3).

E.1 Unbiasedness of the Informed Advantage

Lemma E.1 (Unbiasedness of the informed advantage)**.** *In an informed POMDP, the informed advantage satisfies*

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

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

Proof. By definition,

$$A^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) = Q^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t) - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{h}_t)\, Q^\pi(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}).$$

Taking expectation over $\mathbf{i}_t$ conditioned on $\boldsymbol{h}_t$ and using linearity,

$$\mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t}[A^\pi(\boldsymbol{h}_t, \mathbf{i}_t, \boldsymbol{a}_t)] = \mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t}[Q^\pi(\boldsymbol{h}_t, \mathbf{i}_t, \boldsymbol{a}_t)] - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{h}_t)\, \mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t}[Q^\pi(\boldsymbol{h}_t, \mathbf{i}_t, \boldsymbol{a})].$$

Applying the unbiasedness property of the informed history-based Q-function (cf. Lemma 3.2 in Ebi et al. (2026)) to both terms on the right-hand side yields

$$\begin{aligned}\mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t}[A^\pi(\boldsymbol{h}_t, \mathbf{i}_t, \boldsymbol{a}_t)] &= Q^\pi(\boldsymbol{h}_t, \boldsymbol{a}_t) - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{h}_t)\, Q^\pi(\boldsymbol{h}_t, \boldsymbol{a}) \\ &= A^\pi(\boldsymbol{h}_t, \boldsymbol{a}_t).\end{aligned}$$

This concludes the proof. □

E.2 Conditional Variance Reduction of TD Targets

Proposition E.1 (Conditional variance reduction of TD targets)**.** *Let*

$$y_t := r_t + \gamma Q^\pi_{t+1}(\boldsymbol{h}_{t+1}, \boldsymbol{i}_{t+1}, \boldsymbol{a}_{t+1})$$

denote the informed TD target. Since the privileged signal provides additional conditioning information, the law of total variance implies that, for almost every $(\boldsymbol{h}_t, \boldsymbol{a}_t)$*,*

$$\mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t,\boldsymbol{a}_t}\left[\mathrm{Var}(y_t|\boldsymbol{h}_t, \boldsymbol{a}_t, \mathbf{i}_t)\right] \le \mathrm{Var}(y_t|\boldsymbol{h}_t, \boldsymbol{a}_t).$$

More precisely,

$$\mathrm{Var}(y_t|\boldsymbol{h}_t, \boldsymbol{a}_t) = \mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t,\boldsymbol{a}_t}\left[\mathrm{Var}(y_t|\boldsymbol{h}_t, \boldsymbol{a}_t, \mathbf{i}_t)\right] + \mathrm{Var}\left(\mathbb{E}\left[y_t|\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t\right]|\boldsymbol{h}_t, \boldsymbol{a}_t\right).$$

Hence, the inequality is strict for a given $(\boldsymbol{h}_t, \boldsymbol{a}_t)$ *whenever* $\mathrm{Var}\left(\mathbb{E}\left[y_t|\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t\right]|\boldsymbol{h}_t, \boldsymbol{a}_t\right) > 0$*.*

Proof. Let

$$y_t := r_t + \gamma Q^\pi_{t+1}(\boldsymbol{h}_{t+1}, \boldsymbol{i}_{t+1}, \boldsymbol{a}_{t+1})$$

denote the TD target. For fixed $(\boldsymbol{h}_t, \boldsymbol{a}_t)$, applying the law of total variance with respect to $\mathbf{i}_t$ gives

$$\mathrm{Var}(y_t|\boldsymbol{h}_t, \boldsymbol{a}_t) = \mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t,\boldsymbol{a}_t}[\mathrm{Var}(y_t|\boldsymbol{h}_t, \boldsymbol{a}_t, \mathbf{i}_t)] + \mathrm{Var}(\mathbb{E}[y_t|\boldsymbol{h}_t, \boldsymbol{a}_t, \mathbf{i}_t]|\boldsymbol{h}_t, \boldsymbol{a}_t)\,.$$

Since the second term is non-negative, we obtain, for almost every $(\boldsymbol{h}_t, \boldsymbol{a}_t)$,

$$\mathbb{E}_{\mathbf{i}_t|\boldsymbol{h}_t,\boldsymbol{a}_t}\left[\mathrm{Var}(y_t|\boldsymbol{h}_t, \boldsymbol{a}_t, \mathbf{i}_t)\right] \le \mathrm{Var}(y_t|\boldsymbol{h}_t, \boldsymbol{a}_t).$$

Moreover, the inequality is strict for a given $(\boldsymbol{h}_t, \boldsymbol{a}_t)$ whenever

$$\mathrm{Var}(\mathbb{E}[y_t|\boldsymbol{h}_t, \boldsymbol{a}_t, \mathbf{i}_t]|\boldsymbol{h}_t, \boldsymbol{a}_t) > 0.$$

This concludes the proof. □

E.3 Irreducible Error Floors of the Informed and Symmetric Critic

Corollary E.1 (Asymmetric versus symmetric mean-squared critic error w.r.t. Q^π_t)**.** *Under the hypothesis class inclusion* $\tilde{\mathcal{F}}_T \subseteq \mathcal{F}_T$*, for every* $0 \le t < T$*, we have*

$$\inf_{f\in\mathcal{F}_T} \mathbb{E}\left[\left(f(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t) - Q^\pi_t\right)^2\right] \le \inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \mathbb{E}\left[\left(\tilde{f}(\boldsymbol{h}_t, \boldsymbol{a}_t) - \bar{Q}^\pi_t\right)^2\right] + \mathbb{E}\left[\mathrm{Var}\left(Q^\pi_t|\boldsymbol{h}_t, \boldsymbol{a}_t\right)\right]. \quad (42)$$

Proof. From Lemma 5.1, for any $\tilde{f} \in \tilde{\mathcal{F}}_T$, we have

$$\mathbb{E}\left[\left(\tilde{f}(\boldsymbol{h}_t, \boldsymbol{a}_t) - Q^\pi_t\right)^2\right] = \mathbb{E}\left[\left(\tilde{f}(\boldsymbol{h}_t, \boldsymbol{a}_t) - \bar{Q}^\pi_t\right)^2\right] + \mathbb{E}\left[\mathrm{Var}\left(Q^\pi_t|\boldsymbol{h}_t, \boldsymbol{a}_t\right)\right]. \quad (43)$$

Since the variance term does not depend on $\tilde{f}$, it factors out of the infimum over $\tilde{f} \in \tilde{\mathcal{F}}_T$, i.e.,

$$\inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \mathbb{E}\left[\left(\tilde{f}(\boldsymbol{h}_t, \boldsymbol{a}_t) - Q^\pi_t\right)^2\right] = \inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \mathbb{E}\left[\left(\tilde{f}(\boldsymbol{h}_t, \boldsymbol{a}_t) - \bar{Q}^\pi_t\right)^2\right] + \mathbb{E}\left[\mathrm{Var}\left(Q^\pi_t|\boldsymbol{h}_t, \boldsymbol{a}_t\right)\right]. \quad (44)$$

Using the hypothesis class inclusion $\tilde{\mathcal{F}}_T \subseteq \mathcal{F}_T$, we obtain

$$\inf_{f\in\mathcal{F}_T} \mathbb{E}\left[(f(\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{i}_t)-Q_t^\pi)^2\right] \le \inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \mathbb{E}\left[\left(\tilde{f}(\boldsymbol{h}_t,\boldsymbol{a}_t)-Q_t^\pi\right)^2\right]. \tag{45}$$

Substituting Eq. 44 into the right-hand side gives

$$\inf_{f\in\mathcal{F}_T} \mathbb{E}\left[(f(\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{i}_t)-Q_t^\pi)^2\right] \le \inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \mathbb{E}\left[\left(\tilde{f}(\boldsymbol{h}_t,\boldsymbol{a}_t)-\bar{Q}_t^\pi\right)^2\right] + \mathbb{E}\left[\mathrm{Var}\left(Q_t^\pi|\boldsymbol{h}_t,\boldsymbol{a}_t\right)\right]. \tag{46}$$

This concludes the proof. □

Corollary E.2 (Irreducible error floors of the critic w.r.t. Q_t^π)**.** *Under Assumption 3 and the hypothesis class inclusion $\tilde{\mathcal{F}}_T \subseteq \mathcal{F}_T$, for every $0 \le t < T$, we have*

$$\inf_{f\in\mathcal{F}_T} \mathbb{E}\left[(f(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)-Q_t^\pi)^2\right] = 0, \tag{47}$$

and

$$\inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \mathbb{E}\left[(\tilde{f}(\boldsymbol{h}_t,\boldsymbol{a}_t)-Q_t^\pi)^2\right] = \mathbb{E}\left[Var(Q_t^\pi|\boldsymbol{h}_t,\boldsymbol{a}_t)\right]. \tag{48}$$

Consequently,

$$\inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \mathbb{E}\left[(\tilde{f}(\boldsymbol{h}_t,\boldsymbol{a}_t)-Q_t^\pi)^2\right] - \inf_{f\in\mathcal{F}_T} \mathbb{E}\left[(f(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)-Q_t^\pi)^2\right] = \mathbb{E}\left[Var(Q_t^\pi|\boldsymbol{h}_t,\boldsymbol{a}_t)\right]. \tag{49}$$

Proof. Since $Q_t^\pi \in \mathcal{F}_T$ (cf. Assumption 3), there exists $f^\star \in \mathcal{F}_T$ such that $f^\star(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t) = Q_t^\pi$, which implies

$$\inf_{f\in\mathcal{F}_T} \mathbb{E}\left[(f-Q_t^\pi)^2\right] = 0.$$

For the hypothesis class $\tilde{\mathcal{F}}_T$ of the symmetric critic, the optimal predictor is

$$\tilde{f}^\star(\boldsymbol{h}_t,\boldsymbol{a}_t) = \mathbb{E}[Q_t^\pi|\boldsymbol{h}_t,\boldsymbol{a}_t].$$

Substituting into the mean-squared error decomposition yields

$$\begin{aligned}\inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \mathbb{E}[(\tilde{f}-Q_t^\pi)^2] = \mathbb{E}\left[(\tilde{f}^\star-Q_t^\pi)^2\right] &= \mathbb{E}\left[(\mathbb{E}[Q_t^\pi|\boldsymbol{h}_t,\boldsymbol{a}_t]-Q_t^\pi)^2\right] \\ &= \mathbb{E}\left[\mathrm{Var}(Q_t^\pi|\boldsymbol{h}_t,\boldsymbol{a}_t)\right].\end{aligned}$$

By subtracting the asymmetric critic's risk, we obtain

$$\inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \mathbb{E}\left[(\tilde{f}(\boldsymbol{h}_t,\boldsymbol{a}_t)-Q_t^\pi)^2\right] - \inf_{f\in\mathcal{F}_T} \mathbb{E}\left[(f(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)-Q_t^\pi)^2\right] = \mathbb{E}\left[\mathrm{Var}(Q_t^\pi|\boldsymbol{h}_t,\boldsymbol{a}_t)\right].$$

This concludes the proof. □

Corollary E.3 (Trajectory-level irreducible error floors of the critic)**.** *Assume $Q_t^\pi \in \mathcal{F}_T$ for all $0 \le t < T$ and $\tilde{\mathcal{F}}_T$ consists of measurable functions of $(\boldsymbol{h}_t,\boldsymbol{a}_t)$. Then, the trajectory-level risks satisfy $\inf_{f\in\mathcal{F}_T} \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left[(f_t(\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{i}_t)-Q_t^\pi)^2\right] = 0$, and for any $\tilde{f}_t \in \tilde{\mathcal{F}}_T$,*

$$\sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left[\left(\tilde{f}_t(\boldsymbol{h}_t,\boldsymbol{a}_t)-Q_t^\pi\right)^2\right] = \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left[\left(\tilde{f}_t(\boldsymbol{h}_t,\boldsymbol{a}_t)-\bar{Q}_t^\pi\right)^2\right] + \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left[\mathrm{Var}\left(Q_t^\pi|\boldsymbol{h}_t,\boldsymbol{a}_t\right)\right].$$

Consequently,

$$\inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left[\left(\tilde{f}_t(\boldsymbol{h}_t,\boldsymbol{a}_t)-Q_t^\pi\right)^2\right] \ge \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left[\mathrm{Var}\left(Q_t^\pi|\boldsymbol{h}_t,\boldsymbol{a}_t\right)\right] = \sum_{t=0}^{T-1} \gamma^t \varepsilon_{alias,t}. \tag{50}$$

Proof. Since $Q_t^\pi \in \mathcal{F}_T$ for all $0 \le t < T$ (cf. Assumption 3), there exists $f^\star \in \mathcal{F}_T$ such that

$$f_t^\star(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t) = Q_t^\pi,$$

which implies

$$\inf_{f \in \mathcal{F}_T} \mathcal{R}_T^\pi(f) = 0.$$

For the symmetric class, applying the pointwise decomposition (cf. Lemma 5.1) at each time step yields, for any $\tilde{f}_t \in \tilde{\mathcal{F}}_T$,

$$\mathbb{E}\left[(\tilde{f}_t(\boldsymbol{h}_t, \boldsymbol{a}_t) - Q_t^\pi)^2\right] = \mathbb{E}\left[(\tilde{f}_t(\boldsymbol{h}_t, \boldsymbol{a}_t) - \bar{Q}_t^\pi)^2\right] + \mathbb{E}\left[\mathrm{Var}(Q_t^\pi | \boldsymbol{h}_t, \boldsymbol{a}_t)\right].$$

Multiplying by γ^t and summing over $0 \le t < T$ gives

$$\mathcal{R}_T^\pi(\tilde{f}) = \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left[(\tilde{f}_t - \bar{Q}_t^\pi)^2\right] + \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left[\mathrm{Var}(Q_t^\pi | \boldsymbol{h}_t, \boldsymbol{a}_t)\right].$$

Since the second term is independent of $\tilde{f}$ and the first term is non-negative, taking the infimum over $\tilde{f} \in \tilde{\mathcal{F}}_T$ yields

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}_T} \mathcal{R}_T^\pi(\tilde{f}) \ge \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left[\mathrm{Var}(Q_t^\pi | \boldsymbol{h}_t, \boldsymbol{a}_t)\right].$$

This completes the proof. □

E.4 Comparison of Asymmetric and Symmetric Finite-Time Bounds for Rec-TD

Comparing the explicit bounds for informed Rec-TD in the proof of Theorem 5.1 and the corresponding result for the symmetric critic in (Cayci & Eryilmaz, 2025) (cf. Theorem F.1 in Appendix F.1), excluding the explicit information-aliasing contributions, gives

$$\begin{aligned}\Delta L(K, m_F) = \frac{||\nu||_2^2 - ||\tilde{\nu}||_2^2}{\eta_{\text{td}}(1-\gamma)K} + \frac{\eta_{\text{td}}}{(1-\gamma)^3}(\delta_{\max}^2 (L_T')^2 - \tilde{\delta}_{\max}^2 (L_T^F)^2) \\ + \frac{1}{(1-\gamma)^2\sqrt{m_F}}(c_T \delta_{\max} - \tilde{c}_T \tilde{\delta}_{\max}),\end{aligned} \tag{51}$$

where $\eta_{\text{td}} > 0$ denotes the step size, and $\delta_{\max}, L_T', L_T^F, c_T$ are defined in Lemma E.2-E.3 in Appendix E.6-E.8 and in the proof of Theorem 5.1. We further have

$$\tilde{\delta}_{\max} = r_{\max} + 2L_T^F |\tilde{\rho}_{\vartheta^F}|_2, \tag{52}$$

and

$$\tilde{c}_T = 2(\beta_T^F)^2 |\tilde{\rho}_{\vartheta^F}|_2^2 + (\varrho_2 (\Lambda_t^F)^2 + \varrho_1 \chi_t^F)|\tilde{\rho}_{\vartheta^F}|_2^2 + 2||\tilde{\nu}||_2 \sqrt{1 + \varrho_0^2} \cdot p_T(\alpha \varrho_1) \tag{53}$$

for the symmetric critic.

The trade-off between improved approximation error and additional complexity is inherently finite-time. As K and m_F increase, the optimization and finite-width contributions to $\Delta L(K, m_F)$ decrease. In the limiting regime $K, m_F \to \infty$, these complexity terms vanish, so under the function-class inclusion and realizability assumptions, any strictly positive aliasing reduction yields a strict improvement in the corresponding asymptotic guarantee. At finite K and m_F, however, information gain alone does not guarantee a smaller error bound.

Moreover, we find that, since informed critic operates on the concatenated representation $\boldsymbol{z}_t = \boldsymbol{f}_t \oplus \boldsymbol{i}_t$, the NTK-based Lipschitz and smoothness bounds introduce a dependence on $\sup_t \|\boldsymbol{i}_t\|_2$. This quantity reflects the scale of the input rather than its informativeness and enters the finite-time constants through $\delta_{\max}$, L_T', and c_T. Consequently, normalizing the privileged features removes an otherwise arbitrary source of scale dependence and makes the complexity comparison primarily reflect the structural cost of the larger hypothesis class.

Proof of Proposition 5.1.

Proof. By Corollary E.3, we have

$$\inf_{\tilde{f}\in\tilde{\mathcal{F}}_T} \mathbb{E}^{\pi}\big[\sum_{t=0}^{T-1}\gamma^t\big(\tilde{f}(\boldsymbol{h}_t,\boldsymbol{a}_t)-Q_t^{\pi}\big)^2\big] = \varepsilon_{\text{app}}^{F}(\tilde{\mathcal{F}}_T)+\sum_{t=0}^{T-1}\gamma^t\varepsilon_{\text{alias},t},$$

for the symmetric critic's approximation error with respect to Q_t^{π}. Under the assumed realizability of $\bar{Q}_{k,t}^{\pi}$ in $\tilde{\mathcal{F}}_T$, for iteration $k\in\mathbb{Z}_{\geq 0}$, the first term vanishes, so the additional approximation error contribution caused by the coarser information set is $\sum_{t<T}\gamma^t\varepsilon_{\text{alias},k,t}$.

If the realizability assumption of the finite-time theorem is relaxed, this approximation error enters the bound with a coefficient of order $\frac{1}{1-\gamma}$. We denote the corresponding, generally unspecified, coefficient by $\frac{C_{\text{app}}}{1-\gamma}$, where C_{app} is determined by the constants hidden in the non-realizable extension of the bound.

At the terminal step, the symmetric terminal error satisfies $\bar{\omega}_{k,T}^2 = \tilde{\omega}_{k,T}^2+\varepsilon_{\text{alias},k,T}$, and hence its terminal contribution differs from the corresponding contribution based on $\tilde{\omega}_{k,T}^2$ by $\frac{\gamma^T}{(1-\gamma)K}\sum_{k<K}\varepsilon_{\text{alias},k,T}$.

The remaining differences between the informed and symmetric finite-time bounds are collected in $\Delta L(K,m_F)$.

Consequently, after canceling common terms, the informed bound is smaller whenever

$$\frac{C_{\text{app}}}{1-\gamma}\sum_{t<T}\gamma^t\varepsilon_{\text{alias},k,t}+\frac{\gamma^T}{(1-\gamma)K}\sum_{k<K}\varepsilon_{\text{alias},k,T} > \Delta L(K,m_F)+\frac{\gamma^T}{(1-\gamma)K}\sum_{k<K}\left(\omega_{k,T}^2-\tilde{\omega}_{k,T}^2\right).$$

Multiplying both sides by the positive factor $(1-\gamma)$ yields the stated sufficient condition.

This concludes the proof. □

E.5 Decomposition of the Actor's Approximation Error

Proposition E.2 (Decomposition of the actor's approximation error)**.** *Let*

$$\varepsilon_{app,n}^{D} := \inf_{\boldsymbol{q}\in\Omega_{\rho,m_G}} \mathbb{E}_{\boldsymbol{s}_0\sim P}^{\pi_{\theta_n}}\left[\sum_{t=0}^{T-1}\gamma^t\left|\nabla^{\top}D_t(\boldsymbol{h}_t,\boldsymbol{a}_t;\theta_0)\boldsymbol{q}-Q_t^{\pi_{\theta_n}}(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)\right|^2\right]$$

and

$$\bar{Q}_t^{\pi_{\theta_n}}(\boldsymbol{h}_t,\boldsymbol{a}_t) := \mathbb{E}\left[Q_t^{\pi_{\theta_n}}(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)\,\middle|\,\boldsymbol{h}_t,\boldsymbol{a}_t\right].$$

Then the approximation error decomposes as

$$\varepsilon_{app,n}^{D} = \tilde{\varepsilon}_{app,n}^{D}+\varepsilon_{info,n},$$

where

$$\tilde{\varepsilon}_{app,n}^{D} := \inf_{\boldsymbol{q}\in\Omega_{\rho,m_G}} \mathbb{E}_{\boldsymbol{s}_0\sim P}^{\pi_{\theta_n}}\left[\sum_{t=0}^{T-1}\gamma^t\left|\nabla^{\top}D_t(\boldsymbol{h}_t,\boldsymbol{a}_t;\theta_0)\boldsymbol{q}-\bar{Q}_t^{\pi_{\theta_n}}(\boldsymbol{h}_t,\boldsymbol{a}_t)\right|^2\right],$$

$$\varepsilon_{info,n} := \sum_{t=0}^{T-1}\gamma^t\mathbb{E}_{\boldsymbol{s}_0\sim P}^{\pi_{\theta_n}}\Big[\mathrm{Var}(Q_t^{\pi_{\theta_n}}(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)|\boldsymbol{h}_t,\boldsymbol{a}_t)\Big].$$

Proof. Since the actor features depend only on $(\boldsymbol{h}_t,\boldsymbol{a}_t)$, $(\nabla^{\top}D_t(\boldsymbol{h}_t,\boldsymbol{a}_t;\theta_0)\boldsymbol{q})$ is measurable with respect to $(\boldsymbol{h}_t,\boldsymbol{a}_t)$. Let $Q_t := Q_t^{\pi_{\theta_n}}(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)$ and

$$\bar{Q}_t := \bar{Q}_t^{\pi_{\theta_n}}(\boldsymbol{h}_t,\boldsymbol{a}_t) = \mathbb{E}\left[Q_t^{\pi_{\theta_n}}(\boldsymbol{h}_t,\boldsymbol{i}_t,\boldsymbol{a}_t)|\boldsymbol{h}_t,\boldsymbol{a}_t\right].$$

For any $\boldsymbol{q}$, we have

$$|\nabla^{\top}D_t(\boldsymbol{h}_t,\boldsymbol{a}_t;\theta_0)\boldsymbol{q}-Q_t|^2 = \left|\nabla^{\top}D_t(\boldsymbol{h}_t,\boldsymbol{a}_t;\theta_0)\boldsymbol{q}-\bar{Q}_t+\bar{Q}_t-Q_t\right|^2.$$

Expanding the square yields

$$|\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - Q_t|^2 = |\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - \bar{Q}_t|^2 + |\bar{Q}_t - Q_t|^2 + 2(\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - \bar{Q}_t)(\bar{Q}_t - Q_t).$$

Taking conditional expectation given $(\boldsymbol{h}_t, \boldsymbol{a}_t)$, the cross term vanishes:

$$\mathbb{E}\big[(\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - \bar{Q}_t)(\bar{Q}_t - Q_t)|\boldsymbol{h}_t, \boldsymbol{a}_t\big] = (\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - \bar{Q}_t)\mathbb{E}\big[\bar{Q}_t - Q_t|\boldsymbol{h}_t, \boldsymbol{a}_t\big] = 0,$$

as $\bar{Q}_t$ is the conditional expectation of Q_t.

Therefore

$$\mathbb{E}\big[|\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - Q_t|^2\big|\boldsymbol{h}_t, \boldsymbol{a}_t\big] = |\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - \bar{Q}_t|^2 + \mathrm{Var}(Q_t|\boldsymbol{h}_t, \boldsymbol{a}_t).$$

Taking expectation over trajectories gives

$$\mathbb{E}|\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - Q_t|^2 = \mathbb{E}|\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - \bar{Q}_t|^2 + \mathbb{E}[\mathrm{Var}(Q_t|\boldsymbol{h}_t, \boldsymbol{a}_t)]\,.$$

Summing over t with weights γ^t yields

$$\mathbb{E}^{\pi_{\theta_n}}_{\boldsymbol{s}_0\sim P}\left[\sum_{t=0}^{T-1}\gamma^t|\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - Q_t|^2\right] = \mathbb{E}^{\pi_{\theta_n}}_{\boldsymbol{s}_0\sim P}\left[\sum_{t=0}^{T-1}\gamma^t|\nabla^\top D_t(\boldsymbol{h}_t, \boldsymbol{a}_t; \theta_0)\boldsymbol{q} - \bar{Q}_t|^2\right] + \sum_{t=0}^{T-1}\gamma^t\mathbb{E}\Big[\mathrm{Var}(Q_t|\boldsymbol{h}_t, \boldsymbol{a}_t)\Big].$$

The second term does not depend on $\boldsymbol{q}$, hence taking the infimum over $\boldsymbol{q} \in \Omega_{\rho,m_G}$ only affects the first term. This yields

$$\varepsilon^D_{\text{app},n} = \tilde{\varepsilon}^D_{\text{app},n} + \varepsilon_{\text{info},n}.$$

This concludes the proof. □

E.6 Local Continuity of the Informed Critic

For proving Theorem 5.1, we need to derive Lipschitzness and smoothness for the parameters of the informed critic. As this critic is composed of an IndRNN and a learnable linear readout, we extend the results given by Lemma B.1 in (Cayci & Eryilmaz, 2025) to the composed neural architecture of the informed 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 E.2 (Local continuity of the informed 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^F_t-Lipschitz continuous with $L^F_t = \left(\varrho_0^2 + 1\right)\varrho_1^2 \cdot p_t^2\left(\alpha_{m_F}\varrho_1\right)$,*
- $\vartheta^F_{j,:} \mapsto F_{t_j}\left(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F\right)$ *is β^F_t-smooth with $\beta^F_t = \mathcal{O}\left(d_x \cdot p_t\left(\alpha_{m_F}\varrho_1\right) \cdot q_t\left(\alpha_{m_F}\varrho_1\right)\right)$,*
- $\vartheta^F_{j,:} \mapsto \Gamma_{t_j}\left(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F\right)$ *is Λ^F_t-Lipschitz continuous with $\Lambda^F_t = \sqrt{2}\left(\varrho_0 + 1 + \alpha_{m_F} L^F_t\right)$,*
- $\vartheta^F_{j,:} \mapsto \Gamma_{t_j}\left(\boldsymbol{h}_t, \boldsymbol{a}_t; \vartheta^F\right)$ *is χ^F_t-smooth with $\chi^F_t = \sqrt{2}\left(L^F_t + \alpha_{m_F}\beta^F_t\right)$,*
- $b_l \mapsto \Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \boldsymbol{b})$ *is L^Φ_t-Lipschitz continuous with*

$$L^\Phi_t = \tfrac{1}{\sqrt{d_z}}|(\boldsymbol{z}_t)_l| \leq \tfrac{1}{\sqrt{d_z}}\left(L^F_t \cdot \|\boldsymbol{\rho}_{\vartheta^F}\|_2 + 1\right),$$

- $b_l \mapsto \Phi_t(\boldsymbol{h}_t, \boldsymbol{a}_t, \boldsymbol{i}_t; \boldsymbol{b})$ *is β^Φ_t-smooth with $\beta^\Phi_t = 0$*

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

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

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

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

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

E.7 Approximation Error between NTRF and NTK in the Informed Critic

Furthermore, we extend the external result outlined in Lemma B.2 in (Cayci & Eryilmaz, 2025) to the composed neural architecture of the informed critic.

Lemma E.3 (Approximation error between NTRF and NTK in the informed critic)**.** *Let* $\phi^\star \in \mathcal{F}$ *with the transportation mapping* $v \in \mathcal{M}_F$, *and let*

$$\bar{\vartheta}_{j,:} = \vartheta_{0_{j,:}} + \frac{1}{\sqrt{m_F}}\zeta_j v\left(\vartheta_{0_{j,:}}\right), \quad j \in \{1,\ldots,d+m_F+d_i-1\}, \tag{54}$$

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

$$\Phi_t^{Lin}(\cdot;\vartheta) := \nabla_\vartheta \Phi_t(F_t^{Lin}(\cdot;\vartheta_0^F),\cdot;\boldsymbol{b}_0)\cdot(\vartheta-\vartheta_0). \tag{55}$$

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*

$$\mathbb{E}_{\vartheta_0}\left[\mathbb{E}^\pi\left[\left(\phi_t^\star\left(\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{i}_t\right) - \Phi_t^{Lin}\left(\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{i}_t;\bar{\vartheta}\right)\right)^2\right]\right] \leq \frac{2\|\boldsymbol{\nu}\|_2^2(1+\varrho_0^2)p_t^2(\alpha\varrho_1)}{m_F}. \tag{56}$$

Remark E.1. In the informed critic, $\boldsymbol{z}_t$ is correlated with the RNN hidden state $\boldsymbol{f}_t$ (that is, $\boldsymbol{z}_t = \boldsymbol{f}_t \oplus \boldsymbol{i}_t$), so the NTK contains nonzero cross-block entries. This correlation is implicitly captured in the linearization $\Phi_t^{\text{Lin}}(\cdot;\vartheta)$ and affects the effective feature space and statistical complexity.

E.8 Non-stationary Bellman Equation in the Asymmetric Setting

We extend the non-stationary Bellman equation (cf. Proposition B.3 in (Cayci & Eryilmaz, 2025)) to the asymmetric setting:

Proposition E.3 (Non-stationary Bellman equation in the 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* $t \in \mathbb{Z}_{\geq 0}$.

F External Results

This section provides a summary of external results necessary for analyzing the asymmetric variant of Rec-NAC. Section F.1 and Section F.2 contain the main theoretical results of Cayci & Eryilmaz (2025) for symmetric Rec-TD (Theorem F.1) and symmetric Rec-NPG (Theorem F.2), respectively.

F.1 Finite-time Bound for Symmetric Rec-TD

Theorem F.1 (Finite-time bound for Rec-TD; Theorem 5.4 in (Cayci & Eryilmaz, 2025)). *Assume for the symmetric history Q-function that $\{Q_t^\pi : t \in \mathbb{Z}_{\geq 0}\} \in \tilde{\mathcal{F}}$ with a transportation mapping $\tilde{v} \in \tilde{\mathcal{M}}$ such that*

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

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:

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

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

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

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

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

is a uniformly bounded sequence for $t, k \in \mathbb{Z}_{\geq 0}$.

Furthermore, the loss at average-iterate Rec-TD with $\bar{\theta}_K := \frac{1}{K}\sum_{k=0}^{K-1}\theta_k$ admits the same upper bound as the regret upper bound in Eq. 57, up to a multiplicative factor of 10.

F.2 Error Bound on the Best Iterate for Symmetric Rec-NPG

Theorem F.2 (Error bound on the best-iterate for symmetric Rec-NPG; Theorem 6.3 in (Cayci & Eryilmaz, 2025)). *Assume that $\mathbb{P}_T^{\pi^\star,P} \ll \mathbb{P}_T^{\pi_{\theta_n},P}, n < N$, and let*

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

be the concentrability coefficient.

Rec-NPG after $N \in \mathbb{N}$ steps with step-size $\eta_{npg} = \frac{1}{\sqrt{N}}$ with projection radius $\boldsymbol{\rho}_G \in \mathbb{R}_{>0}^2$ yields

$$\min_{0\leq n<N}\mathbb{E}_0\left[V^{\pi^\star}(P) - V^{\pi_{\theta_n}}(P)\right] \lesssim \frac{\log|\mathcal{A}|}{(1-\gamma)\sqrt{N}} + \frac{\|\tilde{\boldsymbol{\rho}}\|_2^2}{1-\gamma}\frac{\left|\left(\varrho_1(\alpha + \frac{\tilde{\rho}_w}{\sqrt{m_G}})\right)\right|^k}{m_G^{\frac{1}{4}}} + \frac{\gamma^T r_\infty}{(1-\gamma)^2} + \frac{\sqrt{\Xi}}{N\sqrt{1-\gamma}}\sum_{n=0}^{N-1}\mathbb{E}_0\big(\tilde{\varepsilon}_{cfa}^T(\theta_n, \boldsymbol{q}_n)\big)^{\frac{1}{2}}$$

where $\mathbb{E}_0$ is the conditional expectation given the symmetric random initialization $(\theta_0, \boldsymbol{c}_0)$, and

$$\tilde{\varepsilon}_{cfa}^T(\theta, \boldsymbol{q}) := \sum_{t=0}^{T-1}\gamma^t|\nabla^\top \log\pi_t^\theta(\boldsymbol{a}_t|\boldsymbol{h}_t)\boldsymbol{q} - A_t^{\pi^\theta}(\boldsymbol{h}_t, \boldsymbol{a}_t)|^2. \tag{58}$$