Learning to Control Partially Observed Systems with Finite Memory

Semih Cayci* Niao He† R. Srikant‡

Abstract

We consider the reinforcement learning problem for partially observed Markov decision processes (POMDPs) with large or even countably infinite state spaces, where the controller has access to only noisy observations of the underlying controlled Markov chain. We consider a natural actor-critic method that employs a finite internal memory for policy parameterization, and a multi-step temporal difference learning algorithm for policy evaluation. We establish, to the best of our knowledge, the first non-asymptotic global convergence of actor-critic methods for partially observed systems under function approximation. In particular, in addition to the function approximation and statistical errors that also arise in MDPs, we explicitly characterize the error due to the use of finite-state controllers. This additional error is stated in terms of the total variation distance between the traditional belief state in POMDPs and the posterior distribution of the hidden state when using a finite-state controller. Further, we show that this error can be made small in the case of sliding-block controllers by using larger block sizes.

1 Introduction

The class of optimal control problems where the controller has access to only noisy observations of the system state are modeled as partially observed Markov decision processes (POMDPs) [1, 2, 3, 4]. Since the underlying state is only partially known to the controller in POMDPs, the optimal policy depends on the complete history of the system, making the problem highly intractable [4, 5]. To overcome the computational challenges in solving POMDPs, a plethora of model-based and model-free reinforcement learning approaches have been proposed in the literature that incorporate finite memory into the controller, via internal state or quantization of belief state; see e.g., surveys [6, 5, 7].

The actor-critic framework, which combines the benefits of both value-based methods and policy gradient methods, has shown great promise in learning POMDPs in practice [8, 9, 10, 11, 12]. These methods offer more flexibility in controlling the bias-variance tradeoff. However, theoretical analyses of the convergence rates and optimality properties of these POMDP solvers seem largely absent, particularly in the interesting case of function approximation for large state-action-observation spaces. In this paper, we provide new results to this fundamental problem.

1.1 Main Contributions

Our main contributions include the following.

- **NAC for POMDPs:** We consider a model-free natural actor-critic (NAC) method for POMDPs:
  - the actor employs an internal state representation as a form of memory to resolve the *curse of history*, and performs efficient natural policy gradient update,
  - the critic employs a multi-step temporal difference learning algorithm to obtain the value functions by eliminating the impact of partial observability.

*CSL, University of Illinois at Urbana-Champaign, scayci@illinois.edu
†Department of Computer Science, ETH Zurich, niao.he@inf.ethz.ch
‡ECE and CSL, University of Illinois at Urbana-Champaign, rsrikant@illinois.edu

- **Near-optimality of NAC for POMDPs:** We establish, to our best knowledge, the first finite-time performance bounds of actor-critic-type methods for POMDPs, with explicit characterization of the convergence rate, function approximation error, and inference error due to the partial observability. Our result indicates that the proposed NAC method achieves near-optimality with the class of finite-state controllers with $\widetilde{O}(1/\epsilon^6)$ sample complexity in the function approximation regime, up to the compatible function approximation error and inference error.

- **Memory-Inference Error Tradeoff:** We further consider NAC with sliding-block controllers, and provide explicit bounds on the tradeoff between memory complexity and inference error. Notably, under mild ergodicity conditions on the underlying MDP to satisfy filter stability, the inference error decays at a geometric rate in the block-length, which implies tractability and effectiveness of the proposed NAC approach.

1.2 Related Work

PG/NPG/NAC for MDPs: Policy gradient methods have been extensively investigated for fully-observed MDPs [13, 14, 15, 16, 17, 18, 19]. As they rely on the perfect state observation in MDPs, they do not address the problem of partial observability that we consider in this paper.

Policy Evaluation for POMDPs: Policy evaluation methods have been considered in [20, 21] for average-reward POMDPs for the specific class of reactive (memoryless) policies in the tabular (small-state) case. In [22], TD(1) was adapted for finite-state controllers. In our work, as part of the natural actor-critic framework, we study policy evaluation with function approximation for large-scale POMDPs, and employ a multi-step TD learning algorithm to obtain the value function under partial observability. Multi-step TD learning was considered in [23, 24] for MDPs. For POMDPs, finite-step TD learning methods do not converge to the true value functions for policy evaluation unlike MDPs. However, we show in this paper that the error can be controlled by employing a multi-step TD learning algorithm, which leads to a tradeoff between sample complexity and accuracy. In particular, we prove the effectiveness of multi-step TD learning for POMDPs by providing a non-asymptotic analysis, which yields explicit sample complexity bounds and exhibits the impact of partial observability and function approximation.

Model-free RL for POMDPs: Actor-critic methods were first mentioned as a potential solution method for POMDPs in [7, 22] and empirically studied in several papers, e.g., [12, 9, 10, 11]. However, none of these works provides non-asymptotic convergence guarantees or optimality properties. In [25], Q-learning with sliding-block controllers was considered and analyzed for the tabular case. In our paper, inspired by [25, 26], we consider and analyze a direct policy optimization scheme with finite-state controllers and function approximation for large (potentially infinite) state spaces.

1.3 Notation

For a sequence of elements $\{s_k : k \geq 0\}$ in a set S, the vector $(s_i, s_{i+1}, \ldots, s_j)$ for any $i \leq j$ is denoted by s_i^j. Let s^n denote s_0^n. We denote the cardinality of a countable set S by $|\mathsf{S}|$. For a countable set S, $\rho(\mathsf{S})$ denotes the simplex over S: $\rho(\mathsf{S}) = \{v \in \mathbb{R}^{|\mathsf{S}|} : v_i \geq 0,\ \forall i \in \mathsf{S},\ \ \sum_{i\in\mathsf{S}} v_i = 1\}$. For $\xi \in \rho(\mathsf{S})$, $\mathsf{supp}(\xi) = \{s \in \mathsf{S} : \xi(s) > 0\}$ denotes the support set of ξ. For $v \in \mathbb{R}^{|\mathsf{S}|}$ and $\xi \in \rho(\mathsf{S})$, we denote the weighted-ℓ_2 norm as: $\|v\|_\xi = \sqrt{\sum_{i\in\mathsf{S}} \xi(i)\,|v(i)|^2}$. For $d \geq 1$, $\mathcal{B}_d(x, R) = \{z \in \mathbb{R}^d : \|x - z\|_2 \leq R\}$ denotes the ℓ_2-ball with center x and radius R. For a matrix $A \in \mathbb{R}^{m\times n}$, $A^\dagger$ denotes its Moore-Penrose inverse. For a random variable $X \in \mathsf{X}$, $\sigma(X)$ denotes its σ-field.

2 POMDPs and Finite-State Controllers

We consider a discrete-time dynamical system with a countable state space X, and finite control space U. The system state $\{x_k : k \geq 0\}$ is a time-homogenous controlled Markov chain, which evolves according to $x_{k+1} \sim \mathcal{P}(\cdot|x_k, u_k),\ \ k = 0, 1, \ldots,$ where $\mathcal{P}$ is the transition kernel, and $u_k \in \mathsf{U}$ is the control at time k. The system state is available to the controller only through a (noisy) discrete memoryless observation channel: $y_k \sim \Phi(\cdot|x_k),\ \ k =$

$0, 1, \ldots$, where $y_k \in \mathsf{Y}$ is the observation and $\Phi(\cdot|x) \in \rho(\mathsf{Y})$ is the observation channel for any $x \in \mathsf{X}$. The channel is memoryless in the following sense:

$$\mathbb{P}(y^k = \bar{y}^k | x^k = \bar{x}^k) = \prod_{i=0}^{k} \Phi(\bar{y}_i | \bar{x}_i),$$

for any $(\bar{x}^k, \bar{y}^k)$ and $k \geq 0$. The information available to the controller at time k is the following:

$$h_k = (h_{k-1}, y_k, u_{k-1}), \tag{1}$$

with an initial history $h_0 \in \mathsf{H}$. An admissible policy $\pi = (\mu_0, \mu_1, \ldots)$ is a sequence of mappings $\mu_k : \mathsf{H} \times \mathsf{Y}^k \times \mathsf{U}^k \to \rho(\mathsf{U})$ for all $k \geq 0$. The class of all admissible policies is denoted by Π_A. Applying control $u \in \mathsf{U}$ at state $x \in \mathsf{X}$ yields a reward $r(x, u) \in [0, r_{max})$.

2.1 Value Function for POMDPs

For a given admissible policy $\pi \in \Pi_A$, the value function is defined as the γ-discounted reward given the prior distribution b_0 (associated with the initial knowledge h_0):

$$V^\pi(h_0) = \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k r(x_k, u_k) \Big| x_0 \sim b_0(\cdot, h_0)\Big]. \tag{2}$$

Similarly, we define the Q-function under π as follows:

$$Q^\pi(h_0, u_0) = \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k r(x_k, u_k) \Big| x_0 \sim b_0(\cdot, h_0), u_0\Big].$$

The advantage function is defined as follows:

$$A^\pi(h_0, u_0) = Q^\pi(h_0, u_0) - V^\pi(h_0). \tag{3}$$

The ultimate objective is to find the optimal policy over the class of admissible policies that maximizes the discounted reward given an initial distribution $\xi \in \rho(\mathsf{H})$, namely,

$$\max_{\pi \in \Pi_A} \; V^\pi(\xi) := \mathbb{E}_{h_0 \sim \xi}[V^\pi(h_0)]. \tag{4}$$

Note that the optimal controller for a POMDP bases its decisions on $h_k \in \mathsf{H} \times \mathsf{Y}^k \times \mathsf{U}^k$, thus an exponentially growing memory over time is required for policy optimization, which is known as the *curse of history*.

2.2 Bayes Filtering and Belief State Formulation

It is well known that the POMDP objective can be reformulated as a perfect state information problem (MDP) by enlarging the state space [3, 4, 5]. Let

$$b_k(x, h) = \mathbb{P}(x_k = x | h_k = h), \tag{5}$$

be the belief state at time $k \geq 0$, which we omit the dependence on h to simplify the notation. The belief can be computed in a recursive way by the following filtering transformation:

$$b_k(x) = \frac{\sum_{x'} b_{k-1}(x') \mathcal{P}(x|x', u_{k-1}) \Phi(y_k|x)}{\sum_{x', x''} b_{k-1}(x') \mathcal{P}(x''|x', u_{k-1}) \Phi(y_k|x'')}, \tag{6}$$

$$= F(b_{k-1}, y_k, u_{k-1})(x), \tag{7}$$

which follows from the Bayes theorem [5]. We denote $k \geq 0$ successive applications of the filtering transformation F as follows:

$$F^{(k)}(b_0, y_1^k, u_0^{k-1})(x) = b_k(x). \tag{8}$$

For any $u \in \mathsf{U}$, let $\tilde{r}(b_k, u) = \sum_{x \in \mathsf{X}} b_k(x) r(x, u)$. Then, the problem reduces to a fully observable MDP where (b_k, u_k) forms a controlled Markov chain, and action u_k at belief state b_k yields a reward $\tilde{r}(b_k, u)$ [4]. Therefore, the techniques for MDPs can, in theory, be applied to solve the POMDP problem (4). On the other hand, the belief $b_k \in \rho(\mathsf{U})$ is continuous-valued, which makes the policy search problem highly intractable. This constitutes the main challenge in RL for POMDPs [7, 5].

2.3 Finite-State Controllers for POMDPs

In order for tractability, a finite-memory controller that employ internal states to summarize the history, is widely used for POMDPs [26, 22, 21, 20, 7]. In this paper, we will mainly focus on this tractable class of admissible policies and investigate its performance guarantees.

Let Z be a finite set, and consider a stochastic process $\{z_k : k \geq 0\}$ over Z, that evolves according to the following transition:

$$z_{k+1} \sim \varphi(\cdot | z_k, y_k, u_k), \tag{9}$$

where φ is a transition kernel. Consider a subclass of admissible policies $\pi = (\mu_0, \mu_1, \ldots)$ such that the policy at time k bases its decision on the latest observation y_k and the variable z_k, i.e., μ_k is $\sigma(y_k, z_k)$-measurable for all $k \geq 0$. We denote this policy subclass by $\Pi_{F,\mathsf{Z},\varphi}$, and call them *finite-state controllers* (FSC). As such, z_k is called the *internal state* of the controller. The initial knowledge of the controller about the system, $h_0 \in \mathsf{H}$, is the vector $(y_0, z_0) \in \mathsf{Y} \times \mathsf{Z}$. Hence, for the specific case of finite-state controllers, $\mathsf{H} = \mathsf{Y} \times \mathsf{Z}$.

The goal in this paper is to learn the optimal FSC for a given internal state space Z, which is closest to the optimal value function (4).

Definition 1 (Optimal FSC)**.** *For any given Z and $\xi \in \rho(\mathsf{H})$, the optimal FSC is defined as follows:*

$$\pi^* = \arg \max_{\pi \in \Pi_{F,\mathsf{Z},\varphi}} V^\pi(\xi), \tag{10}$$

where the optimization is over all transition kernels φ and policies $\pi \in \Pi_{F,\mathsf{Z},\varphi}$.

Example 1 (Sliding-Block Controllers)**.** *An important class of FSC is sliding-block controllers (SBC) [27, 28, 25], which was shown to achieve good practical performance, particularly in combination with stochastic policies (which we aim to learn in this paper by using the NAC framework) [29]. For a given block-length $n > 0$, the internal state is defined as follows:*

$$z_k = (y_{k-n}^{k-1}, u_{k-n}^{k-1}) \in \mathsf{Y}^n \times \mathsf{U}^n.$$

For $n = 0$, the internal state is null, thus the controller bases its decisions at time k only on the last observation y_k, which is called reactive or memoryless policies [20, 22]. For sliding-block controllers, the initial internal state is $h_0 = (y_0, z_0) = (y_{-n}^0, u_{-n}^{-1}) \in \mathsf{Y}^{n+1} \times \mathsf{U}^n$, thus $\mathsf{H} = \mathsf{Y}^{n+1} \times \mathsf{U}^n$. In Appendix A.1, we exemplify how to sample h_0 by following a given exploratory policy.

Remark 1. *For SBC with block-length n, the memory complexity is $O(|\mathsf{Y}|^{n+1} \cdot |\mathsf{U}|^{n+1})$, which increases exponentially in n if one considers a rich policy class that approaches to Π_A. Hence, the block-length n leads to a tradeoff between optimality and (memory) complexity.*

Remark 2 (Near-optimality of FSCs)**.** *It was shown in [26] and [25] that the class of finite-state controllers achieves near-optimality in solving* (4) *as $|\mathsf{Z}|$ increases. The goal of this paper is to learn the best FSC π^* by using samples for any given Z.*

Algorithm 1: NAC-FSC

Input: Ψ: feature set, T: time-horizon, N: number of SGD steps, ζ, η: step-sizes, R: projection radius.
Initialization: $\theta_0 = 0$; \\ max-entropy policy
for $t = 0$ ***to*** $T-1$ **do**
  Obtain $\overline{Q}_T^{\pi_t}$ and $\overline{V}_T^{\pi_t}$ by using m-step TD learning (Alg. 2)
  Initialize: $w_t(0) = 0$
  for $k = 0$ ***to*** $N-1$ **do**
    \\ SGD steps
    Obtain $(y_k, z_k) \sim \mathsf{d}_\xi^{\pi_t}$ and $u_k \sim \pi_t(\cdot|y_k, z_k)$
    $w_t(k+1) = \arg\min_{w \in \mathcal{B}_d(0,R)} \|w_t(k) - \zeta \cdot \nabla \mathcal{L}_t(w_t(k), y_k, z_k, u_k) - w\|_2$
  end
  $\theta_{t+1} = \theta_t + \eta \frac{1}{N} \sum_{k<N} w_t(k)$ \\ policy update
end

3 Natural Actor-Critic for POMDPs

In this section, we develop a natural actor-critic (NAC) framework in order to find the optimal policy within the class of FSCs.

We consider softmax policy parameterization for FSCs with linear function approximation. Specifically, for a given set of feature vectors $\Psi = \{\psi(y, z, u) \in \mathbb{R}^d : (y, z, u) \in \mathsf{Y} \times \mathsf{Z} \times \mathsf{U}\}$, let

$$\pi_\theta(u|y, z) = \frac{\exp\left(\theta^\top \psi(y, z, u)\right)}{\sum_{u' \in \mathsf{U}} \exp\left(\theta^\top \psi(y, z, u')\right)}, \quad (u, y, z) \in \mathsf{U} \times \mathsf{Y} \times \mathsf{Z} \tag{11}$$

Under the observation and internal state pair $(y, z) \in \mathsf{Y} \times \mathsf{Z}$, the controller makes a randomized decision $u \sim \pi_\theta(\cdot|y, z)$.

We make the following assumption for the sampling process, which is standard in policy gradient methods [14, 30]. For any $(y, z) \in \mathsf{Y} \times \mathsf{Z}$ and a given $h_0 = (y_0, z_0)$, denote $\mathsf{d}_{h_0}^\pi(y, z) = (1-\gamma) \sum_{k=0}^\infty \gamma^k \cdot \mathbb{P}^\pi((y_k, z_k) = (y, z)|h_0)$. We define the state-action visitation distribution as $\mathsf{d}_\xi^\pi(y, z) = \mathbb{E}_{(y_0, z_0) \sim \xi} \mathsf{d}_{(y_0, z_0)}^\pi(y, z)$. For sampling from the distribution d_ξ^π, see Remark 6.

Assumption 1 (Sampling oracle)**.** *We assume that the controller is able to obtain an independent sample from* $h_0 \sim \mathsf{d}_\xi^\pi$ *at any time.*

The proposed NAC algorithm, which is summarized in Algorithm 1, works as follows. We initialize the policy optimization at the max-entropy policy by setting $\theta_0 = 0$. At iteration $t \geq 0$, the policy parameter is denoted by θ_t, and the corresponding policy is $\pi_t := \pi_{\theta_t}$. NAC algorithm consists of the following steps at iteration t:

Step 1: (Critic) Obtain the state-action value function $\bar{Q}_K^{\pi_t}$ by using multi-step TD learning as described in Algorithm 2, which is sufficient to compute:

$$\bar{A}_K^{\pi_t}(y, z, u) = \bar{Q}_K^{\pi_t}(y, z, u) - \bar{V}_K^{\pi_t}(y, z), \tag{12}$$

where $\bar{V}_K^{\pi_t}(y, z) = \sum_{u \in \mathsf{U}} \pi_t(u|y, z) \bar{Q}_K^{\pi_t}(y, z, u)$, for a given number of critic steps per iteration K.

Remark 3 (Perceptual aliasing and multi-step TD learning)**.** *The main challenge in estimating Q^π by using temporal difference (TD) learning methods is the so-called* perceptual aliasing *phenomenon [6, 20], which refers to receiving the same observation $y \in \mathsf{Y}$ for two different (hidden) states $x, x' \in \mathsf{X}$ with non-zero probability due to the noisy observation channel Φ. Consequently, TD(0) does not converge to the true value of Q^π [20]. We employ multi-step TD learning to overcome perceptual aliasing. The impact of perceptual aliasing and multi-step lookahead is explicitly characterized in Theorem 1.*

Algorithm 2: m-step TD learning for POMDPs

Input: Ψ: feature set, m: memory size, α: step-size, K: time-horizon, R: proj. radius

for $t = 0$ ***to*** $K-1$ **do**

$h_0(t) = (y_0(t), z_0(t)) \sim \mathsf{d}_\xi^{\pi_t}$,

$x_0(t) \sim b_0(\cdot, h_0(t))$,

for $k = 0$ *to* $m-1$ **do**

$u_k(t) \sim \pi(\cdot|y_k(t), z_k(t))$,

$x_{k+1}(t) \sim \mathcal{P}(\cdot|x_k(t), u_k(t))$,

$y_{k+1}(t) \sim \Phi(\cdot|x_{k+1}(t))$,

$z_{k+1}(t) = \varphi(\cdot|z_k(t), y_k(t), u_k(t))$,

Observe $r_k(t) = r(x_k(t), u_k(t))$.

Compute semi-gradient:

$$g_t = \Big(\sum_{k=0}^{m-1} \gamma^k r_k(t) - \gamma^m \widehat{Q}_{t,m}^\pi - \widehat{Q}_{t,0}^\pi\Big)\psi_0(t),$$

where $\psi_k(t) = \psi(y_k(t), z_k(t), u_k(t))$ and $\widehat{Q}_{t,k}^\pi = \langle \beta_t, \psi_k(t)\rangle$.

Update with projection:

$$\beta_{t+1} = \operatorname*{arg\,min}_{\beta \in \mathcal{B}_d(0,R)} \|\beta_t + \alpha g_t - \beta\|_2,$$

Return $\overline{Q}_K^\pi(\cdot) = \langle \frac{1}{K}\sum_{t<K} \beta_t, \psi(\cdot)\rangle$.

Step 2: (Actor) Given $\bar{A}_K^{\pi_t}$, let $\mathfrak{H}_t$ be the σ-field generated by the samples used in the policy optimization up to (excluding) iteration t, and in the computation of $\bar{Q}_K^{\pi_t}$. Then, we aim to solve the following optimization problem:

$$\min_{w \in \mathcal{B}_d(0,R)} \mathbb{E}[\mathcal{L}_t(w, y, z, u)|\mathfrak{H}_t] := \mathbb{E}\Big[\Big(\langle \nabla \log \pi_t(u|y,z), w\rangle - \bar{A}_K^{\pi_t}(y,z,u)\Big)^2\Big|\mathfrak{H}_t\Big], \tag{13}$$

which would approximate the natural gradient ascent update $G_t^\dagger \nabla V^{\pi_t}$ where G_t is the Fisher information matrix under π_t [31]. In order to solve (13), we initialize $w_t(0) = 0$ and utilize the following stochastic gradient descent (SGD) steps with ℓ_2-projection:

$$w_t(k+1) = \arg\min_{w \in \mathcal{B}_d(0,R)} \|w_t(k) - \zeta \cdot \nabla_w \mathcal{L}_t(w_t(k), y_k, z_k, u_k) - w\|_2. \tag{14}$$

After N steps of SGD iterations (14), the policy update is performed as $\theta_{t+1} = \theta_t + \eta \cdot \frac{1}{N}\sum_{k=0}^{N-1} w_t(k)$.

4 Convergence of Natural Actor-Critic for POMDPs

In this section, we will provide finite-time performance bounds for the proposed NAC algorithm, and identify the impacts of partial observability, function approximation and internal state representation on the global optimality of this method.

Without loss of generality, we assume that

$$\|\psi(y,z,u)\|_2 \leq 1,$$

for all $(y,z,u) \in \mathsf{Y} \times \mathsf{Z} \times \mathsf{U}$. Given a projection radius $R > 0$, the function space defined by Ψ is

$$\mathcal{F}_\Psi^R = \{q : q(\cdot) = \langle \beta, \psi(\cdot)\rangle, \beta \in \mathcal{B}_d(0,R)\}. \tag{15}$$

For any given $\pi \in \Pi_{F,\mathsf{Z},\varphi}$, $\mathsf{d} \in \rho(\mathsf{Y} \times \mathsf{Z})$ and $R > 0$, let $(\mathsf{d} \circ \pi)(y,z,u) = \mathsf{d}(y,z)\pi(u|y,z)$, and

$$\beta_\pi = \operatorname*{arg\,min}_{\beta \in \mathcal{B}_d(0,R)} \|Q^\pi(\cdot) - \langle \beta, \psi(\cdot)\rangle\|_{\mathsf{d}\circ\pi}^2. \tag{16}$$

Also, let

$$\ell_{2,\xi}(f, \mathcal{F}_{\Psi}^R) = \min_{v \in \mathcal{F}_{\Psi}^R} \|f - v\|_{\xi}, \tag{17}$$

be the projection error under weighted-ℓ_2 norm with $\xi \in \rho(\mathsf{Y} \times \mathsf{Z} \times \mathsf{U})$.

4.1 Convergence of the Critic

In the following, we present finite-time performance bounds for the m-step TD learning algorithm for policy evaluation for any given FSC $\pi \in \Pi_{F,\mathsf{Z},\varphi}$ in the partial information setting.

Theorem 1 (Convergence of m-step TD learning)**.** *For any $\pi \in \Pi_{F,\mathsf{Z},\varphi}$ and $m \geq 1$, we have the following bound under Algorithm 2 with the step-size $\alpha = \frac{1}{\sqrt{K}}$ and given projection radius $R > 0$:*

$$\sqrt{\mathbb{E}\|Q^\pi - \overline{Q}_K^\pi\|^2_{\mathsf{d}_\xi^\pi \circ \pi}} \leq \sqrt{\frac{\|\beta_0 - \beta_\pi\|_2^2 + M^2(\gamma, R)}{K^{1/2}(1-\gamma^m)}} + \frac{\ell_{2,\mathsf{d}_\xi^\pi \circ \pi}(Q^\pi, \mathcal{F}_\Psi^R)}{1-\gamma^m} + \epsilon_{\mathtt{PA}}^\pi(\gamma, m, R),$$

where $\overline{Q}_K^\pi(\cdot) = \langle \frac{1}{K}\sum_{t<K} \beta_t, \psi(\cdot)\rangle$ is the output of Algorithm 2,

$$M(\gamma, R) = \frac{r_{max}(1-\gamma^m)}{1-\gamma} + (1+\gamma^m)R,$$

and

$$\begin{aligned}\epsilon_{\mathtt{PA}}^\pi(\gamma, m, R) = \Big(R + \frac{r_{max}}{1-\gamma}\Big)\sqrt{\frac{2\gamma^m d_{TV}(\tilde{\mathsf{d}}_{m,\xi}^\pi \circ \pi, \mathsf{d}_\xi^\pi \circ \pi)}{1-\gamma^m}} \\ + O\Big(\frac{\gamma^m}{1-\gamma}\Big)\Big\|\mathbb{E}\Big[\sum_{k=0}^\infty \gamma^{km} d_{TV}\big(b_0(\cdot, I_{(k+1)m}), b_{(k+1)m}(\cdot, h_{(k+1)m})\big)\Big| I_0 = \cdot\Big]\Big\|_{\mathsf{d}_\xi^\pi \circ \pi},\end{aligned} \tag{18}$$

where $I_k = (y_k, z_k)$, $\tilde{\mathsf{d}}_{m,\xi}^\pi(y, z) := \sum_{y_0, z_0} \mathbb{P}\Big((y_m, z_m) = (y, z)\Big| y_0, z_0\Big)\mathsf{d}_\xi^\pi(y_0, z_0)$.

The proof of Theorem 1 is given in Appendix A.

Remark 4. *From Theorem 1, we have the following observations.*

1. *The term $\epsilon_{\mathtt{PA}}(\gamma, m, R)$ is unique to POMDPs and will not appear in the case of (fully observable) MDPs. Specifically, it quantifies the impact of perceptual aliasing on policy evaluation in the partially observed setting. Since $d_{TV}(p, q) \in [0, 1]$ for any $p, q \in \rho(\mathsf{X})$, the error term decays at a rate $O(\gamma^{m/2})$, thus incorporating m-step lookahead resolves perceptual aliasing for POMDPs.*

2. *In the specific case of (fully observable) MDPs, the second term in $\epsilon_{\mathtt{PA}}(\gamma, m, R)$ is 0. If ξ is the stationary distribution of the underlying controlled Markov chain, we have $\tilde{\mathsf{d}}_{m,\xi}^\pi = \mathsf{d}_\xi^\pi = \xi$, which eliminates the first term also. Thus, by setting $m = 1$, we recover the bounds for TD(0) in the iid setting for MDPs [32, 33].*

3. *At each iteration, m samples are used by m-step TD learning. Thus, the sample complexity of m-step TD learning is mK. As such, there is a tradeoff between accuracy in and sample complexity in policy evaluation.*

4. *For any given $n \geq 0$, the required m to achieve a given target error ϵ is $O\left(\log_{1/\gamma}\left(\frac{1}{(1-\gamma)\epsilon}\right)\right)$, which implies that a small memory m suffices for discounted-reward problems with $\gamma \in (0, 1)$. Hence, the sample complexity to achieve a target error $\epsilon > 0$ is $\widetilde{O}\Big(\frac{1}{(1-\gamma)^2\epsilon^4}\Big)$.*

In the following subsection, we provide finite-time convergence bounds for the NAC-FSC.

4.2 Finite-Time Bounds for NAC-FSC

First, we consider the performance of the natural actor-critic for a general finite-state controller, and characterize the function approximation error, statistical error and inference error.

Definition 2. *For a given set of feature vectors Ψ and projection radius $R > 0$, let*

$$\bar{\ell}_{\mathsf{C}}(R, \Psi) = \sup_{t\geq 0} \mathbb{E}[\ell_{2,\mathsf{d}_\xi^{\pi_t}\circ\pi_t}(Q^{\pi_t}, \mathcal{F}_\Psi^R)|\mathfrak{H}_t],$$

be the compatible function approximation error, where $\mathfrak{H}_t$ is the σ-field generated by all samples used in policy optimization until (and excluding) iteration t.

Definition 2 characterizes the representation power of the function approximation used in policy parameterization and policy evaluation.

Assumption 2 (Concentrability coefficient)**.** *Let*

$$C_t = \mathbb{E}\Big[\mathbb{E}\Big[\Big(\frac{(\mathsf{d}_\xi^{\pi^*}\circ\pi^*)(y,z,u)}{(\mathsf{d}_\xi^{\pi_t}\circ\pi_t)(y,z,u)}\Big)^2\Big|\mathfrak{H}_t\Big]\Big], \tag{19}$$

where the inner expectation is taken over $(y,z)\sim\mathsf{d}_\xi^{\pi_t}$ and $u\sim\pi_t(\cdot|y,z)$. We assume there exists $\bar{C}_\infty < \infty$ such that

$$\sup_{t\geq 0} C_t \leq \bar{C}_\infty.$$

Note that $\bar{C}_\infty$ quantifies the impact of distributional shift in policy optimization, and Assumption 2 is standard in policy gradient methods for MDPs [33, 14]. On the other hand, since $\mathsf{d}_\xi^\pi\circ\pi$ depends on (Z,φ), and π^* may use a different transition kernel for the internal state on Z, Assumption 2 imposes a regularity condition for the internal state representation characterized by (Z,φ).

Theorem 2 (Convergence of NAC-FSC)**.** *Consider the finite-state natural actor-critic with internal state (Z,φ). Then, Algorithm 1 with step-sizes $\alpha = \frac{1}{\sqrt{K}}$, $\zeta = \frac{R\sqrt{1-\gamma}}{\sqrt{2Nr_{max}}}$ and $\eta = \frac{1}{\sqrt{T}}$ achieves the following bound:*

$$(1-\gamma)\min_{t<T}\mathbb{E}[V^{\pi^*}(\xi) - V^{\pi_t}(\xi)] \leq \frac{\log|\mathsf{U}| + R^2}{\sqrt{T}} + 8\bar{C}_\infty\Big(\epsilon_{\mathsf{C}}^m(K,R) + \epsilon_{\mathsf{A}}(N,R)\Big) + 2r_{max}\Gamma^{\pi^*}(\xi),$$

where

$$\epsilon_{\mathsf{C}}^m(K,R) = \sqrt{\frac{\|\beta_0-\beta_\pi\|_2^2 + M^2(\gamma,R)}{K^{1/2}(1-\gamma^m)}} + \frac{\bar{\ell}_{\mathsf{C}}(R,\Psi)}{1-\gamma^m} + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\epsilon_{\mathtt{PA}}^{\pi_t}(\gamma,m,R)],$$

is the error in the critic due to using sample-based estimation of the state-action value functions obtained by the m-step TD learning algorithm,

$$\epsilon_{\mathsf{A}}(N,R) = \sqrt{\frac{2-\gamma}{1-\gamma}\cdot\frac{R\cdot r_{max}}{\sqrt{N}}} + \bar{\ell}_{\mathsf{C}}(R,\Psi),$$

is the error in the actor due the compatible function approximation error and using SGD, and

$$\Gamma^{\pi^*}(\xi) = \mathbb{E}^{\pi^*}\Big[\sum_{k=0}^\infty \gamma^k d_{TV}\Big(b_k(\cdot,h_k), b_0(\cdot,I_k)\Big)\Big|I_0\sim\xi\Big],$$

with $I_k = (y_k,z_k)$ is the inference error due to using the internal state representation (Z,φ) to address partial observability.

The proof of Theorem 2 is given in Appendix B.

The error bound characterized in Theorem 2 can be decomposed into three parts.

- *Inference error:* The inference error at stage k is $d_{TV}\Big(b_k(\cdot,h_k),b_0(\cdot,I_k)\Big)$, where h_k is the complete history up to time k (see (1)) and $I_k=(y_k,z_k)$ is the information used by the controller. If the internal state $\{z_k : k\geq 0\}$ with (Z,φ) provides a good "temporal" approximation, i.e., summarizes the history properly, then the inference error is small. The error is due to employing an internal state to compress the history. For sliding-block controllers, we will expand this discussion in the following subsection.

- *Error in actor*: This corresponds to the combination of statistical error due to using SGD for policy update, and using a function approximation scheme for policy parameterization. In order to achieve ϵ-optimality up to a function approximation error, which depends on the expressive power of Ψ, one should choose $N=O(\frac{1}{\epsilon^4})$.

- *Error in critic*: This corresponds to the error in the critic in every stage of the policy optimization. Note that, by (18), $\frac{1}{T}\sum_t \mathbb{E}[\epsilon_{\mathtt{PA}}^{\pi_t}(\gamma,m,R)]=\exp(-\Omega(m))$. Therefore, in order to achieve ϵ-optimality, one has to choose $K=O(\frac{1}{\epsilon^4})$, and apply m-step TD learning with $m=O\big(\log_{1/\gamma}(1/\epsilon)\big)$ to control the error due to perceptual aliasing in policy evaluation (see Section 3).

4.3 Memory-Inference Error Tradeoff for Sliding-Block Controllers

The choice of internal state representation determines a tradeoff between memory complexity and the inference error Γ^{π^*}. For the special class of sliding-block controllers with block-length $n\geq 0$, we can explicitly characterize this tradeoff under the following conditions.

Condition 1 (Persistence of excitation under π^*)**.** *There exist* $\alpha\in(0,1)$ *and* $\bar{\mu}\in\rho(\mathsf{U})$ *such that*

$$\alpha\cdot\bar{\mu}(u)\leq\pi^*(u|y,z)\leq\frac{1}{\alpha}\cdot\bar{\mu}(u), \tag{20}$$

for all $(u,y,z)\in\mathsf{U}\times\mathsf{Y}\times\mathsf{Z}$.

Condition 1 implies that $\mathsf{supp}(\pi^*(\cdot|y,z))$ is the same for all $(y,z)\in\mathsf{Y}\times\mathsf{Z}$, which holds if π^* satisfies the persistence of excitation condition. Note that, unlike MDPs, there may be only strictly non-deterministic policies for POMDPs [20]. Furthermore, if one employs entropy regularization within the NAC framework, which is commonly employed in practice, π^* automatically satisfies Condition 1 [34, 35, 12].

Condition 2 (Minorization-majorization)**.** *There exist* $\epsilon_0\in(0,1)$*,* $m_0\geq 1$ *and* $\nu\in\rho(\mathsf{X}\times\mathsf{Y}^{m_0}\times\mathsf{U}^{m_0})$ *such that the following holds:*

$$\begin{aligned}\epsilon_0\cdot\nu(x_{m_0},y_1^{m_0},u^{m_0-1})&\leq\bar{P}_{m_0}(x_{m_0},y_1^{m_0},u^{m_0-1}|x_0),\\&\leq\frac{1}{\epsilon_0}\cdot\nu(x_{m_0},y_1^{m_0},u^{m_0-1}),\end{aligned}$$

for any $x_{m_0},x_0\in\mathsf{X}$ *and* $(y_1^{m_0},u^{m_0-1})$ *where*

$$\bar{P}_m(x_m,y_1^m,u^{m-1}|x_0)=\sum_{x_1^{m-1}}\prod_{j=0}^{m-1}\bar{\mu}(u_j)P(x_{j+1}|x_j,u_j)\Phi(y_{j+1}|x_{j+1}).$$

Condition 2 is an ergodicity condition, and one of the implications is that every hidden state is visited within a finite time interval. This is akin to the standard ergodicity conditions in [36, 37] for MDPs, but it is stronger than them because of the complications due to partial observability. For hidden Markov chains (HMCs), along with a non-degeneracy condition on Φ, Condition 2 implies filter stability for any finite X,Y when $\{x_k : k\geq 0\}$ is irreducible and aperiodic [38]. For further discussion on Conditions 1-2 and an alternative ergodicity condition, see Appendix C.

The following result characterizes the tradeoff between the inference error and the memory complexity.

Proposition 1 (Memory-performance tradeoff)**.** *Under Conditions 1-2, we have the following result:*

$$\Gamma^{\pi^*}(\xi) \leq \frac{1}{(1-\gamma)} \cdot O\Big((1-\alpha^{2m_0-2}\cdot\epsilon_0^2)^{\lfloor \frac{n}{m_0} \rfloor}\Big), \tag{21}$$

for any $n \geq 1$.

Proposition 1 is an extension of Theorem 5.4 in [38] for hidden Markov chains to the case of POMDPs, which accounts for the control (see the following discussion for details). We provide a detailed proof in Appendix C.

Proposition 1 implies that the inference error for NAC-SB of block-length $n \geq 0$ decays at a rate $e^{-\Omega(n/m_0)}$ under Condition 2. Hence, a target inference error ϵ requires a memory complexity of $O(m_0 \log(1/\epsilon))$ where m_0 is specified in Condition 2. In order to gain intuition about this result, note that we can write

$$\begin{aligned} b_0(x, I_k) &= F^{(n)}\Big(b_{-n}(\cdot, y_{k-n}), y_{k-n+1}^k, u_{k-n}^{k-1}\Big)(x), \\ b_k(x, h_k) &= F^{(n)}\Big(b_{k-n}(\cdot, h_{k-n}), y_{k-n+1}^k, u_{k-n}^{k-1}\Big)(x), \end{aligned} \tag{22}$$

where $F^{(n)}$ is the n-step Bayes filter (see Section 2.2), $z_k = (y_{k-n}^{k-1}, u_{k-n}^{k-1})$ and $I_k = (y_k, z_k)$. Note that the prior b_{-n} is time-invariant and deterministic (see (41)), whereas b_{k-n} is $\sigma(h_{k-n})$-measurable. Thus, the inference error at time $k \geq 0$ is the total-variation distance between the probability measures in (22), which start from two different prior distributions $b_{-n}(\cdot, y_{k-n})$ and $b_{k-n}(\cdot, h_{k-n})$, and are updated by using the same samples $(y_{k-n+1}^k, u_{k-n}^{k-1}) \in \mathsf{Y}^n \times \mathsf{U}^n$ which are obtained under π^*. Proposition 1 implies that, if the underlying Markov chain is ergodic in the sense of Condition 2, different priors are forgotten at a geometric rate in n, similar to the case of HMCs [39, 38, 40]. In the case of finite-state POMDPs, for tabular Q-learning, a different characterization of the inference error was presented in [25] under an assumption on the Dobrushin coefficient. In this work, we prove bounds on the inference error in general (potentially countably infinite) state-observation spaces within the natural actor-critic framework under different conditions. A part of these results are inspired by the connection between the inference error and the notion of filter stability which was observed in [25]. However, one difference is that, in our analysis, we were not able to directly use existing filter stability results for HMCs (e.g., [38, 39]) for our problem which deals with POMDPs. The reason is that, while in the case of HMCs, the current observation is only a function of the current hidden state, the current control action is potentially a function of all past observations and control actions. Therefore, in Appendix C, we extend the filter stability results in [38] for HMCs to the case of POMDPs.

Finally, we make the following remark about the total sample complexity.

Remark 5 (Sample Complexity)**.** *Note that the FSC-NAC is a two-timescale policy search, which consists of T outer loops, and N stages of SGD and K stages of m-step TD learning in each loop. By Theorem 2, the sample complexity of each inner loops is $\widetilde{O}(1/\epsilon^4)$, and $T = O(1/\epsilon^2)$. By Proposition 1, if $n = O(m_0 \log(1/\epsilon))$, then we can achieve an optimality gap $\epsilon + \bar{C}_\infty O(\bar{\ell}_{\mathsf{C}}(R, \Psi))$ with sample complexity $\widetilde{O}(1/\epsilon^6)$.*

5 Conclusion

In this paper, we proposed a natural actor-critic method for POMDPs, which employs an internal state for memory compression, and a multi-step TD learning algorithm for the critic. We established bounds on the sample complexity and memory complexity of the proposed NAC method. An important open question is the investigation of optimizing the internal state representation (e.g., by using a recurrent neural networks) to achieve a better sample complexity and inference error.

References

[1] E. J. Sondik, *The optimal control of partially observable Markov processes.* Stanford University, 1971.

[2] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable markov processes over a finite horizon," *Operations research*, vol. 21, no. 5, pp. 1071–1088, 1973.

[3] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

[4] D. Bertsekas, *Dynamic programming and optimal control: Volume I*. Athena scientific, 2012, vol. 1.

[5] V. Krishnamurthy, *Partially observed Markov decision processes*. Cambridge university press, 2016.

[6] G. Shani, J. Pineau, and R. Kaplow, "A survey of point-based pomdp solvers," *Autonomous Agents and Multi-Agent Systems*, vol. 27, no. 1, pp. 1–51, 2013.

[7] K. P. Murphy, "A survey of pomdp solution techniques," *environment*, vol. 2, p. X3, 2000.

[8] H. Yu, "Approximate solution methods for partially observable markov and semi-markov decision processes," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

[9] S. Srinivasan, M. Lanctot, V. Zambaldi, J. Pérolat, K. Tuyls, R. Munos, and M. Bowling, "Actor-critic policy optimization in partially observable multiagent environments," *arXiv preprint arXiv:1810.09026*, 2018.

[10] F. Jurčíček, B. Thomson, and S. Young, "Natural actor and belief critic: Reinforcement algorithm for learning parameters of dialogue systems modelled as pomdps," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 3, pp. 1–26, 2011.

[11] D. Xu and Q. Liu, "Acis: An improved actor-critic method for pomdps with internal state," in *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2015, pp. 369–376.

[12] A. Lee, A. Nagabandi, P. Abbeel, and S. Levine, "Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[13] V. R. Konda and J. N. Tsitsiklis, "Onactor-critic algorithms," *SIAM journal on Control and Optimization*, vol. 42, no. 4, pp. 1143–1166, 2003.

[14] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *Journal of Machine Learning Research*, vol. 22, no. 98, pp. 1–76, 2021.

[15] J. Bhandari and D. Russo, "Global optimality guarantees for policy gradient methods," *arXiv preprint arXiv:1906.01786*, 2019.

[16] G. Lan, "Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes," *arXiv preprint arXiv:2102.00135*, 2021.

[17] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6820–6829.

[18] T. Xu, Z. Wang, and Y. Liang, "Improving sample complexity bounds for (natural) actor-critic algorithms," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4358–4369, 2020.

[19] S. Khodadadian, T. T. Doan, S. T. Maguluri, and J. Romberg, "Finite sample analysis of two-time-scale natural actor-critic algorithm," *arXiv preprint arXiv:2101.10506*, 2021.

[20] S. P. Singh, T. Jaakkola, and M. I. Jordan, "Learning without state-estimation in partially observable markovian decision processes," in *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 284–292.

[21] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *Journal of Artificial Intelligence Research*, vol. 15, pp. 319–350, 2001.

[22] H. Yu, "A function approximation approach to estimation of policy gradient for pomdp with structured policies," in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, 2005, pp. 642–649.

[23] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[24] Z. Chen, S. T. Maguluri, S. Shakkottai, and K. Shanmugam, "A lyapunov theory for finite-sample guarantees of asynchronous q-learning and td-learning variants," *arXiv preprint arXiv:2102.01567*, 2021.

[25] A. D. Kara and S. Yuksel, "Convergence of finite memory q-learning for pomdps and near optimality of learned policies under filter stability," *arXiv preprint arXiv:2103.12158*, 2021.

[26] H. Yu and D. P. Bertsekas, "On near optimality of the set of finite-state controllers for average cost pomdp," *Mathematics of Operations Research*, vol. 33, no. 1, pp. 1–11, 2008.

[27] J. Loch and S. P. Singh, "Using eligibility traces to find the best memoryless policy in partially observable markov decision processes." in *ICML*, vol. 98, 1998, pp. 323–331.

[28] J. Sung, J. K. Salisbury, and A. Saxena, "Learning to represent haptic feedback for partially-observable tasks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2802–2809.

[29] J. Williams and S. Singh, "Experimental results on learning stochastic memoryless policies for partially observable markov decision processes," *Advances in Neural Information Processing Systems*, vol. 11, 1998.

[30] L. Wang, Q. Cai, Z. Yang, and Z. Wang, "Neural policy gradient methods: Global optimality and rates of convergence," *arXiv preprint arXiv:1909.01150*, 2019.

[31] S. M. Kakade, "A natural policy gradient," *Advances in neural information processing systems*, vol. 14, 2001.

[32] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Conference on learning theory*. PMLR, 2018, pp. 1691–1692.

[33] Q. Cai, Z. Yang, J. D. Lee, and Z. Wang, "Neural temporal-difference learning converges to global optima," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[34] L. Shani, Y. Efroni, and S. Mannor, "Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5668–5675.

[35] S. Cayci, N. He, and R. Srikant, "Linear convergence of entropy-regularized natural policy gradient with linear function approximation," *arXiv preprint arXiv:2106.04096*, 2021.

[36] R. Ortner, "Regret bounds for reinforcement learning via markov chain concentration," *Journal of Artificial Intelligence Research*, vol. 67, pp. 115–128, 2020.

[37] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo, "Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games," *arXiv preprint arXiv:2102.04540*, 2021.

[38] R. van Handel, "Hidden markov models," *Lecture notes*, 2018. [Online]. Available: https://web.math.princeton.edu/~rvan/orf557/hmm080728.pdf

[39] O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*, ser. Springer series in statistics. New York: Springer, 2005.

[40] C. McDonald and S. Yüksel, "Exponential filter stability via dobrushin's coefficient," *Electronic Communications in Probability*, vol. 25, pp. 1–13, 2020.

[41] J. Tsitsiklis and B. Van Roy, "Analysis of temporal-diffference learning with function approximation," *Advances in neural information processing systems*, vol. 9, 1996.

[42] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[43] H. L. Royden and P. Fitzpatrick, *Real analysis*. Macmillan New York, 1988, vol. 32.

[44] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

A Analysis of m-Step TD Learning for POMDPs

Definition 3 (Fixed point). *Under an FSC $\pi \in \Pi_{F,\mathsf{Z},\varphi}$, for any $(y_0, z_0, u_0) \in \mathsf{Y} \times \mathsf{Z} \times \mathsf{U}$, let Q_*^π be the fixed point of the following equation:*

$$Q(y_0, z_0, u_0) = \mathbb{E}^\pi \Big[\sum_{k=0}^{m-1} \gamma^k r(x_k, u_k) + \gamma^m Q(y_m, z_m, u_m) \Big| y_0, z_0, u_0 \Big]. \tag{23}$$

Proof of Theorem 1. For any $t \geq 0$ and $k \in \{0, 1, \dots, m\}$, let

$$\psi_k(t) := \psi(y_k(t), z_k(t), u_k(t)),$$

and

$$g_t = \Big(\sum_{k=0}^{m-1} \gamma^k r(x_k(t), u_k(t)) + \gamma^m \widehat{Q}_{t,m}^\pi - \widehat{Q}_{t,0}^\pi \Big) \psi_0(t), \tag{24}$$

be the stochastic semi-gradient, where

$$\widehat{Q}_{t,k}^\pi = \langle \beta_t, \psi_k(t) \rangle,$$

is the value function estimate at time t. Let

$$\tilde{\beta}_{t+1} = \beta_t + \alpha \cdot g_t,$$

which implies that

$$\beta_t = \mathfrak{P}_{\mathcal{B}_d(0,R)}\{\tilde{\beta}_t\},$$

where $\mathfrak{P}_{\mathcal{C}}\{s\} = \arg\min_{s' \in \mathcal{C}} \|s - s'\|_2$ is the projection operator onto $\mathcal{C} \subset \mathbb{R}^d$. Similar to the analysis of TD(0) with function approximation for MDPs [41, 32, 33], we consider the following Lyapunov function:

$$L(\beta) = \|\beta - \beta_\pi\|_2^2, \tag{25}$$

where $\beta_\pi = \underset{\beta \in \mathcal{B}_d(0,R)}{\arg\min} \|Q^\pi(\cdot) - \langle \beta, \psi(\cdot) \rangle\|^2_{\xi^\mu_{\bar{\pi},n}}$ is the optimal approximator defined in (16). Since $\mathcal{B}_d(0,R)$ is a convex subset of $\mathbb{R}^d$ and $\mathfrak{P}_{\mathcal{C}}$ is non-expansive for convex $\mathcal{C}$, we have:

$$\begin{aligned} L(\beta_{t+1}) &\leq \|\tilde{\beta}_{t+1} - \beta_\pi\|_2^2, \\ &= \|\beta_t + \alpha g_t - \beta_\pi\|_2^2, \\ &= L(\beta_t) + 2\alpha \langle g_t, \beta_t - \beta_\pi \rangle + \alpha^2 \|g_t\|_2^2, \end{aligned}$$

for any $t \geq 0$. Since $\sup_{(y,z) \in \mathsf{Y} \times \mathsf{Z}} \|\psi(y,z)\|_2 \leq 1$ and $\|\beta_t\|_2 \leq R$ due to projection, we have:

$$\sup_{t \geq 0} \|g_t\|_2 \leq r_{max} \frac{1 - \gamma^m}{1 - \gamma} + (1 + \gamma^m) R = M(\gamma, R),$$

with probability 1. For any $t \geq 0$, let

$$\begin{aligned} \mathfrak{G}_t &= \sigma(y_k(\tau), z_k(\tau), u_k(\tau), k \leq m, \tau \leq t), \\ \mathfrak{F}_t &= \sigma(y_0(t), z_0(t), u_0(t)), \\ \mathbb{E}_t[\cdot] &= \mathbb{E}[\cdot | \mathfrak{G}_{t-1}]. \end{aligned}$$

Then, the Lyapunov drift is as follows:

$$\mathbb{E}_t[L(\beta_{t+1}) - L(\beta_t) | \mathfrak{F}_t] \leq 2\alpha \mathbb{E}_t[\langle g_t, \beta_t - \beta_\pi \rangle | \mathfrak{F}_t] + \alpha^2 M^2(\gamma, R). \tag{26}$$

Now, we focus on the term $\mathbb{E}_t[\langle g_t, \beta_t - \beta_\pi\rangle|\mathfrak{F}_t]$. Let $\tilde{Q}_*^\pi(\cdot) = \langle \beta_\pi, \psi(\cdot)\rangle$. Then, we obtain the following identity:

$$\mathbb{E}_t[\langle g_t, \beta_t - \beta_\pi\rangle|\mathfrak{F}_t] = \Big(\mathbb{E}_t[\sum_{k=0}^{m-1}\gamma^k r(x_k(t), u_k(t)) + \gamma^m \widehat{Q}_{t,m}^\pi|\mathfrak{F}_t] - \widehat{Q}_{t,0}^\pi\Big)\langle \beta_t - \beta_\pi, \psi_0(t)\rangle. \tag{27}$$

From Definition 3, under Assumption 1, we have:

$$\mathbb{E}_t[\sum_{k=0}^{m-1}\gamma^k r(x_k(t), u_k(t))|\mathfrak{F}_t] = Q_*^\pi(y_0(t), z_0(t), u_0(t)) - \gamma^m \mathbb{E}[Q_*^\pi(y_m(t), z_m(t), u_m(t))|\mathfrak{G}_t]. \tag{28}$$

For notational convenience, we use (y_0, z_0, u_0) instead of $(y_0(t), z_0(t), u_0(t))$, below. Substituting (28) into (27), and expanding the multiplicative terms, we obtain the following:

$$\begin{aligned}\mathbb{E}_t[\langle g_t, \beta_t - \beta_\pi\rangle|\mathfrak{F}_t] = &-\underbrace{\Big(Q_*^\pi(y_0, z_0, u_0) - \widehat{Q}_t^\pi(y_0, z_0, u_0)\Big)^2}_{(i)} \\ &+ \underbrace{\Big(Q_*^\pi(y_0, z_0, u_0) - \widehat{Q}_t^\pi(y_0, z_0, u_0)\Big)\cdot\Big(Q_*^\pi(y_0, z_0, u_0) - \tilde{Q}_*^\pi(y_0, z_0, u_0)\Big)}_{(ii)} \\ &\underbrace{-\gamma^m \mathbb{E}_t[\widehat{Q}_t^\pi(y_m, z_m, u_m) - Q_*^\pi(y_m, z_m, u_m)|\mathfrak{F}_t]\cdot\Big(Q_*^\pi(y_0, z_0, u_0) - \widehat{Q}_t^\pi(y_0, z_0, u_0)\Big)}_{(iii)} \\ &\underbrace{+\gamma^m \mathbb{E}_t[\widehat{Q}_t^\pi(y_m, z_m, u_m) - Q_*^\pi(y_m, z_m, u_m)|\mathfrak{F}_t]\cdot(Q_*^\pi(y_0, z_0, u_0) - \tilde{Q}_*^\pi(y_0, z_0, u_0)\Big)}_{(iv)}.\end{aligned}$$

First, we take expectation of the terms (i)-(iv) above over $\mathfrak{F}_t = \sigma(y_0(t), z_0(t), u_0(t))$. (i) is equal to $-\|Q_*^\pi - \widehat{Q}_t^\pi\|^2_{\mathsf{d}_\xi^\pi\circ\pi}$. For (ii), by Cauchy-Schwarz inequality:

$$\begin{aligned}\mathbb{E}_t\Big[\Big(Q_*^\pi(y_0, z_0, u_0) - \widehat{Q}_t^\pi(y_0, z_0, u_0)\Big)\cdot\Big(Q_*^\pi(y_0, z_0, u_0) - \tilde{Q}_*^\pi(y_0, z_0, u_0)\Big)\Big] \\ \leq \|Q_*^\pi - \widehat{Q}_t^\pi\|_{\mathsf{d}_\xi^\pi\circ\pi}\cdot\|Q_*^\pi - \tilde{Q}^\pi\|_{\mathsf{d}_\xi^\pi\circ\pi}.\end{aligned}$$

Note that:

$$\begin{aligned}\mathbb{E}_t[\big(\widehat{Q}_t^\pi(y_m, z_m, u_m) - Q_*^\pi(y_m, z_m, u_m)\big)^2|\mathfrak{F}_t] = \sum_{y_m, z_m, u_m}\underbrace{\mathsf{d}_\xi^\pi(y_m, z_m)\pi(u_m|y_m, z_m)}_{(\mathsf{d}_\xi^\pi\circ\pi)(y_m, z_m, u_m)} \\ \times\big(\widehat{Q}_t^\pi(y_m, z_m, u_m) - Q_*^\pi(y_m, z_m, u_m)\big)^2,\end{aligned}$$

We have $\sup_{y,z,u} Q^\pi(y,z,u) \leq \frac{r_{max}}{1-\gamma}$, $\sup_{y,z,u}\widetilde{Q}^\pi(y,z,u) \leq R$, and $\sup_{t\geq 0}\max\{|\widehat{Q}_{t,0}^\pi|, |\widehat{Q}_{t,m}^\pi|\} \leq R$ with probability 1 since $\beta_\pi \in \mathcal{B}_d(0,R)$ and $\beta_t \in \mathcal{B}_d(0,R)$ for all $t\geq 0$. Therefore:

$$\begin{aligned}\sqrt{\mathbb{E}_t[\big(\widehat{Q}_t^\pi(y_m, z_m, u_m) - Q_*^\pi(y_m, z_m, u_m)\big)^2|\mathfrak{F}_t]} \leq \|\widehat{Q}_t^\pi - Q_*^\pi\|_{\mathsf{d}_\xi^\pi\circ\pi} \\ + \Big(R + \frac{r_{max}}{1-\gamma}\Big)\sqrt{d_{TV}(\tilde{\mathsf{d}}_{m,\xi}^\pi\circ\pi, \mathsf{d}_\xi^\pi\circ\pi)},\end{aligned} \tag{29}$$

Hence, using (29) in (iii) and (iv), then taking expectation over $\mathfrak{G}_{t-1}$, we obtain:

$$\begin{aligned}\mathbb{E}[\langle g_t, \beta_t - \beta_\pi\rangle] \leq -(1-\gamma^m)\mathbb{E}\|\widehat{Q}_t^\pi - Q_*^\pi\|^2_{\mathsf{d}_\xi^\pi\circ\pi} + (1+\gamma^m)\mathbb{E}[\|\widehat{Q}_t^\pi - Q_*^\pi\|_{\mathsf{d}_\xi^\pi\circ\pi}]\|\tilde{Q}_*^\pi - Q_*^\pi\|_{\mathsf{d}_\xi^\pi\circ\pi} \\ + 2\gamma^m\Big(R + \frac{r_{max}}{1-\gamma}\Big)^2\sqrt{d_{TV}(\tilde{\mathsf{d}}_{m,\xi}^\pi\circ\pi, \mathsf{d}_\xi^\pi\circ\pi)}.\end{aligned} \tag{30}$$

Let $\delta_{t,\pi}^2 = \mathbb{E}\|Q_*^\pi - \widehat{Q}_t^\pi\|_\xi^2$, and recall that $\ell_{2,\xi}(f, \mathcal{F}_\Psi^R) = \min_{\beta \in \mathcal{B}_d(0,R)} \|f(\cdot) - \langle \beta, \psi(\cdot)\|_\xi$. For notational convenience, let

$$\ell_\pi^* = \frac{1+\gamma^m}{2(1-\gamma^m)} \cdot \ell_{2,\mathsf{d}_\xi^\pi}(Q^\pi, \mathcal{F}_\Psi^R).$$

Then, (30) can be written as follows:

$$\mathbb{E}[[\langle g_t, \beta_t - \beta_\pi\rangle] \leq -(1-\gamma^m)\Big(\delta_{t,\pi}^2 - 2\ell_\pi^* \cdot \delta_{t,\pi}\Big) + 2\gamma^m\Big(R + \frac{r_{max}}{1-\gamma}\Big)^2 \sqrt{d_{TV}(\tilde{\mathsf{d}}_{m,\xi}^\pi \circ \pi, \mathsf{d}_\xi^\pi \circ \pi)}. \quad (31)$$

Taking expectation of (26), and using the bound (31), we have the following expected drift result:

$$\begin{aligned}\mathbb{E}[L(\beta_{t+1}) - L(\beta_t)] &\leq -2\alpha(1-\gamma^m)(\delta_{t,\pi} - \ell_\pi^*)^2 \\ &+ 2\alpha(1-\gamma^m)(\ell_\pi^*)^2 + 4\alpha\gamma^m\Big(R + \frac{r_{max}}{1-\gamma}\Big)^2 \sqrt{d_{TV}(\tilde{\mathsf{d}}_{m,\xi}^\pi \circ \pi, \mathsf{d}_\xi^\pi \circ \pi)}\Big) + \alpha^2 M^2(\gamma, R). \quad (32)\end{aligned}$$

Telescoping sum over $t = 0, 1, \ldots, K-1$ yields the following:

$$\begin{aligned}\mathbb{E}[L(\beta_K) - L(\beta_0)] &\leq -2\alpha(1-\gamma^m)\sum_{t<K}\big(\delta_{t,\pi} - \ell_\pi^*\big)^2 \\ &+ 2\alpha K\Big((1-\gamma^m)(\ell_\pi^*)^2 + 2\gamma^m\Big(R + \frac{r_{max}}{1-\gamma}\Big)^2\sqrt{d_{TV}(\tilde{\mathsf{d}}_{m,\xi}^\pi \circ \pi, \mathsf{d}_\xi^\pi \circ \pi)}\Big) + \alpha^2 K M^2(\gamma, R), \quad (33)\end{aligned}$$

Note that $L(\beta_K) \geq 0$ and $L(\beta_0) = \|\beta_0 - \beta_\pi\|_2^2$, which implies that:

$$\begin{aligned}\frac{1}{K}\sum_{t=0}^{K-1}\Big(\delta_{t,\pi} - \ell_\pi^*\Big)^2 \leq \frac{1}{2\alpha K}\|\beta_0 - \beta_\pi\|_2^2 + (\ell_\pi^*)^2 + 2\gamma^m\Big(R + \frac{r_{max}}{1-\gamma}\Big)^2\sqrt{d_{TV}(\tilde{\mathsf{d}}_{m,\xi}^\pi \circ \pi, \mathsf{d}_\xi^\pi \circ \pi)} \\ + \alpha M^2(\gamma, R).\end{aligned}$$

First, by Jensen's inequality,

$$\frac{1}{K}\sum_{t=0}^{K-1}\delta_{t,\pi} \leq \frac{\|\beta_0 - \beta_\pi\|_2}{\sqrt{2\alpha K}} + 2\ell_\pi^* + \sqrt{2\gamma^m\Big(R + \frac{r_{max}}{1-\gamma}\Big)^2\sqrt{d_{TV}(\tilde{\mathsf{d}}_{m,\xi}^\pi \circ \pi, \mathsf{d}_\xi^\pi \circ \pi)}} + \sqrt{\alpha}M(\gamma, R). \quad (34)$$

We have:

$$\sqrt{\mathbb{E}\|Q_*^\pi - \overline{Q}_K^\pi\|_{\mathsf{d}_\xi^\pi \circ \pi}^2} \leq \frac{1}{K}\sum_{t=0}^{K-1}\delta_{t,\pi}. \quad (35)$$

where the equality holds from the linearity of $\widehat{Q}_t^\pi$ in β_t, and the inequality follows from Jensen's inequality. Hence, we have:

$$\begin{aligned}\sqrt{\mathbb{E}\|Q_*^\pi - \overline{Q}_K^\pi\|_{\mathsf{d}_\xi^\pi \circ \pi}^2} \leq \frac{\|\beta_0 - \beta_\pi\|_2}{\sqrt{2\alpha K}} + 2\ell_\pi^* + \sqrt{2\gamma^m\Big(R + \frac{r_{max}}{1-\gamma}\Big)^2\sqrt{d_{TV}(\tilde{\mathsf{d}}_{m,\xi}^\pi \circ \pi, \mathsf{d}_\xi^\pi \circ \pi)}} \\ + \sqrt{\alpha}M(\gamma, R). \quad (36)\end{aligned}$$

In order to bound $\sqrt{\mathbb{E}\|Q^\pi - Q_*^\pi\|_{\mathsf{d}_\xi^\pi \circ \pi}^2}$, we use the following lemma, which extends the analysis in [25] to multi-step TD learning, to characterize the fixed point Q_*^π.

Lemma 1. *Let $\pi \in \Pi_{F,\mathsf{Z},\varphi}$ be an FSC. Let*

$$\bar{r}_m(x, y, z, u) = \mathbb{E}^\pi\Big[\sum_{k=0}^{m-1}\gamma^k r(x_k, u_k)\Big| x_0 = x, y_0 = y, z_0 = z, u_0 = u\Big]. \quad (37)$$

For any $m \geq 1$,

$$Q_*^\pi(y_0, z_0, u_0) = \mathbb{E}^\pi\Big[\sum_{t=0}^\infty \gamma^{tm} \sum_{x_{tm}\in\mathsf{X}} b_0(x_{tm}, I_{tm})\bar{r}_m(x_{tm}, y_{tm}, z_{tm}, u_{tm})\Big|y_0, z_0, u_0\Big], \tag{38}$$

$$Q^\pi(y_0, z_0, u_0) = \mathbb{E}^\pi\Big[\sum_{t=0}^\infty \gamma^{tm} \sum_{x_{tm}\in\mathsf{X}} b_{tm}(x_{tm}, h_{tm})\bar{r}_m(x_{tm}, y_{tm}, z_{tm}, u_{tm})\Big|y_0, z_0, u_0\Big], \tag{39}$$

for any $m \geq 1$ *and* $(y_0, z_0, u_0) \in \mathsf{Y} \times \mathsf{Z} \times \mathsf{U}$, *where* $I_k = (y_k, z_k)$. *Consequently,*

$$\Big|Q_*^\pi(y_0, z_0, u_0) - Q^\pi(y_0, z_0, u_0)\Big| \leq \frac{2r_{max}(1-\gamma^m)\gamma^m}{1-\gamma} \times \mathbb{E}^\pi\Big[\sum_{t=0}^\infty \gamma^{tm} d_{TV}\big(b_0(\cdot, I_{(t+1)m}), b_{(t+1)m}(\cdot, h_{(t+1)m})\big)\Big|y_0, z_0, u_0\Big].$$

We use Lemma 1 to bound $\sqrt{\mathbb{E}\|Q^\pi - Q_*^\pi\|^2_{\mathsf{d}_\xi^\pi \circ \pi}}$ by using the definition of the weighted-ℓ_2 norm. Thus, using this bound and (36) with triangle inequality, and substituting the step-size $\alpha = 1/\sqrt{K}$ conclude the proof. □

Proof of Lemma 1. Since $\{(x_k, y_k, z_k, u_k) : k \geq 0\}$ forms a Markov chain under an FSC, we have:

$$\mathbb{E}^\pi\Big[\sum_{k=0}^{m-1} \gamma^k r(x_{k+tm}, u_{k+tm})\Big|x_{tm} = x, y_{tm} = y, z_{tm} = z, u_{tm} = u, h_{tm-1}\Big] = \bar{r}_m(x, y, z, u), \tag{40}$$

for any $m \geq 1, t \geq 0$, $(x, y, z, u) \in \mathsf{X} \times \mathsf{Y} \times \mathsf{Z} \times \mathsf{U}$. By using the tower property of conditional expectation, the identities for Q_*^π and Q^π follow. For the second part of the proof, we simply use the triangle inequality in conjunction with the fact that $h_0 = I_0$, and the following upper bound:

$$\sup_{x,y,z,u} |\bar{r}_m(x, y, z, u)| \leq \frac{r_{max}(1-\gamma^m)}{1-\gamma}.$$

Hence, we conclude the proof. □

A.1 Sampling h_0 for SBC

For a given block-length n, the system starts from time $-n$ with $x_{-n} \sim \vartheta$, $y_{-n} \sim \Phi(\cdot|y_{-n})$, and obtains an n-step trajectory h_0 by following a given exploratory policy $\tilde{\pi} \in \Pi_A$:

$$\begin{aligned} x_{k+1} &\sim \mathcal{P}(\cdot|x_k, u_k), \\ y_{k+1} &\sim \Phi(\cdot|x_{k+1}), \\ u_{k+1} &\sim \tilde{\pi}(\cdot|y_{-n}^{k+1}, u_{-n}^k), \end{aligned}$$

for $k \in [-n, 0)$ with $u_{-n} \sim \tilde{\pi}(\cdot|y_{-n})$. By using this trajectory, the controller obtains $h_0 = (y_0, z_0) = (y_{-n}^0, u_{-n}^{-1})$, which yields the following prior:

$$b_0 = F^{(n)}\left(b_{-n}(\cdot, y_{-n}), y_{-n+1}^0, u_{-n}^{-1}\right),$$

where

$$b_{-n}(\cdot|y) = \frac{\vartheta(\cdot)\Phi(y|\cdot)}{\sum_{x\in\mathsf{X}} \vartheta(x)\Phi(y|x)}. \tag{41}$$

The initial history (y_0, z_0) is random with the distribution ξ, which can be explicitly specified by using ϑ, $\tilde{\pi}$, $\mathcal{P}$ and Φ:

$$\xi(y_0, z_0) = \sum_{x_{-n}^0\in\mathsf{X}^{n+1}} \vartheta(x_{-n})\Phi(y_{-n}|x_{-n})\mathsf{p}(x_{-n+1}^0, y_{-n+1}^0, u_{-n}^{-1}; x_{-n}, y_{-n}), \tag{42}$$

where

$$\mathsf{p}(x_{-n+1}^{0}, y_{-n+1}^{0}, u_{-n}^{-1}; x_{-n}, y_{-n}) = \prod_{j=-n}^{-1} \tilde{\pi}(u_j | y_{-n}^{j}, u_{-n}^{j-1}) \mathcal{P}(x_{j+1} | x_j, u_j) \Phi(y_{j+1} | x_{j+1}).$$

Remark 6 (Sampling from d_{ξ}^{π})**.** *We can obtain samples from the discounted observation-internal state visitation distribution d_{ξ}^{π} by using an initial sample from ξ obtained via the above scheme in conjunction with the sequential sampler for state visitation distributions (see Algorithm 1 in [14] and [13]).*

B Convergence of FSC-NAC: Proof of Theorem 2

We start with an important lemma for the proof of Theorem 2.

Lemma 2. *For any given (Z, φ), transition kernel for the internal state φ', admissible policy $\pi' \in \Pi_A$ and FSC $\pi \in \Pi_{F,\mathsf{Z},\varphi}$, we have the following bound:*

$$V^{\pi'}(y_0, z_0) - V^{\pi}(y_0, z_0) \geq \mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} A^{\pi}(y_k, z_k, u_k) | y_0, z_0\Big] - \frac{2r_{max}}{1-\gamma} \cdot \Gamma^{\pi'}(y_0, z_0), \tag{43}$$

where

$$A^{\pi}(y, z, u) = Q^{\pi}(y, z, u) - V^{\pi}(y, z),$$

and

$$\Gamma^{\pi'}(y_0, z_0) = \mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} \gamma^k d_{TV}\Big(b_k(\cdot, h_k), b_0(\cdot, I_k)\Big)\Big| y_0, z_0\Big],$$

under $\pi' \in \Pi_A$ and internal state transition dynamics φ'.

As a corollary to Lemma 2, for an FSC π', we have the following result.

Corollary 1 (Performance Difference Lemma for FSCs)**.** *Consider (Z, φ') and (Z, φ), and two FSCs, $\pi' \in \Pi_{F,\mathsf{Z},\varphi'}$ and $\pi \in \Pi_{F,\mathsf{Z},\varphi}$. Then,*

$$V^{\pi'}(y_0, z_0) - V^{\pi}(y_0, z_0) \geq \frac{1}{1-\gamma} \sum_{y \in \mathsf{Y}} \sum_{z \in \mathsf{Z}} \sum_{u \in \mathsf{U}} \mathsf{d}_{y_0, z_0}^{\pi'}(y, z) \pi'(u | y, z) A^{\pi}(y, z, u) - \frac{2r_{max}}{1-\gamma} \cdot \Gamma^{\pi'}(y_0, z_0), \tag{44}$$

where $\mathsf{d}_{y_0,z_0}^{\pi'}(y, z) = (1-\gamma)\sum_{k=0}^{\infty} \mathbb{P}^{\pi'}\big((y_k, z_k) = (y, z) | y_0, z_0\big)$ is the discounted observation-internal state visitation distribution under $\pi' \in \Pi_{F,\mathsf{Z},\varphi'}$.

For the case of (fully observable) MDPs, Corollary 1 reduces to the well-known performance difference lemma proposed in [42]. In Lemma 2 and Corollary 1, inspired by the analyses in [42] and [25], we establish the performance difference results for POMDPs, which characterize the impact of partial observability for finite-state controllers.

Proof. We have the following identity from the definition:

$$\begin{aligned}
V^{\pi'}(y_0, z_0) - V^{\pi}(y_0, z_0) &= \mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} \gamma^k r(x_k, u_k) \Big| y_0, z_0\Big] - V^{\pi}(y_0, z_0), \\
&= \mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} \gamma^k \Big(r(x_k, u_k) + V^{\pi}(y_k, z_k) - V^{\pi}(y_k, z_k)\Big) \Big| y_0, z_0\Big] - V^{\pi}(I_0), \\
&\overset{(a)}{=} \mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} \gamma^k \Big(r(x_k, u_k) + \gamma V^{\pi}(y_{k+1}, z_{k+1}) - V^{\pi}(y_k, z_k)\Big) \Big| y_0, z_0\Big],
\end{aligned}$$

where (a) holds since

$$\mathbb{E}[\sum_{k=0} \gamma^k V^\pi(y_k, z_k)|y_0, z_0] = V^\pi(y_0, z_0) + \gamma \cdot \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k V^\pi(y_{k+1}, z_{k+1})|y_0, z_0].$$

Since the Bayes-filtered value function V^π is not the fixed point of a Bellman equation due to POMDP dynamics, we decompose (a) into two parts as follows:

$$\begin{aligned} V^{\pi'}(y_0, z_0) - V^\pi(y_0, z_0) = &\underbrace{\mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} \gamma^k \Big(r(x_k, u_k) + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1}) - V^\pi(y_k, z_k)\Big)\Big| y_0, z_0\Big]}_{(i)}, \\ &+ \underbrace{\gamma \cdot \mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} \gamma^k \Big(V^\pi(y_{k+1}, z_{k+1}) - V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1})\Big| y_0, z_0\Big]}_{(ii)}. \end{aligned} \tag{45}$$

In what follows, we will bound both summands on the right-hand side of the above identity.

Bounding (i) **in** (45)**:** Since $0 \leq \inf_{x \in \mathsf{X}, u \in \mathsf{U}} r(x, u) \leq \sup_{x \in \mathsf{X}, u \in \mathsf{U}} r(x, u) \leq r_{max} < \infty$, we have the following inequality for any $K > 0$ almost surely:

$$\Big|\sum_{k=0}^{K} \gamma^k \Big(r(x_k, u_k) + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1}) - V^\pi(y_k, z_k)\Big)\Big| \leq \frac{2r_{max}}{(1-\gamma)^2} < \infty. \tag{46}$$

By Lebesgue's dominated convergence theorem [43], this implies that we can expand (i) in (45) as follows:

$$\begin{aligned} &\mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} \gamma^k \Big(r(x_k, u_k) + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1}) - V^\pi(y_k, z_k)\Big)\Big| y_0, z_0\Big] \\ &\qquad = \sum_{k=0}^{\infty} \gamma^k \mathbb{E}^{\pi'}\Big[r(x_k, u_k) + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1}) - V^\pi(y_k, z_k)\Big| y_0, z_0\Big]. \end{aligned} \tag{47}$$

For any $k \geq 0$, by the law of iterated expectation, we can write the following:

$$\begin{aligned} &\mathbb{E}^{\pi'}[r(x_k, u_k) + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1}) - V^\pi(y_k, z_k)|y_0, z_0] \\ &\qquad = \mathbb{E}\Big[\mathbb{E}^{\pi'}[r(x_k, u_k) + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1})|h_k, z_k] - V^\pi(y_k, z_k)\Big| y_0, z_0\Big]. \end{aligned} \tag{48}$$

To bound $\mathbb{E}^{\pi'}[r(x_k, u_k) + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1})|h_k, z_k]$, note that $\{(x_k, y_k, z_k, u_k) : k \geq 0\}$ forms a Markov chain and b_k is sufficient statistics for x_k given (h_k, z_k). Hence,

$$\mathbb{E}^{\pi'}[r(x_k, u_k) + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1})|h_k, z_k] = \sum_{x_k, u_k} b_k(x_k, h_k)\pi'(u_k|h_k) Q_0^\pi(x_k, y_k, z_k, u_k),$$

where

$$\begin{aligned} Q_0^\pi(x_0, y_0, z_0, u_0) &= \mathbb{E}^\pi[\sum_{k=0}^{\infty} \gamma^k r(x_k, u_k)|x_0, y_0, z_0, u_0], \\ &= \mathbb{E}^\pi\Big[r(x_0, u_0) + \gamma V_0^\pi(x_1, y_1, z_1)\Big| x_0, y_0, z_0, u_0\Big]. \end{aligned} \tag{49}$$

Also, by the definition of the Bayes-filtered Q-function, we have the following identity:

$$Q^\pi(y, z, u) = \sum_x b_0(x, (y, z)) Q_0^\pi(x, y, z, u),$$

where b_0 is the conditional probability distribution of x_0 given (y_0, z_0). Hence, we obtain:

$$\mathbb{E}^{\pi'}[r_k + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1})|h_k, z_k] = \sum_{u_k \in \mathsf{U}} \pi'(u_k|h_k) Q^\pi(y_k, z_k, u_k) \\ + \sum_{x_k, u_k} \pi'(u_k|h_k)\Big(b_k(x_k, h_k) - b_0(x_k, I_k)\Big) Q_0^\pi(x_k, y_k, z_k, u_k), \quad (50)$$

where we used the last two identity in the expansion. Note that

$$\Big|Q_0^\pi(x_k, y_k, z_k, u_k)\Big| \leq \frac{2r_{max}}{1-\gamma}, \text{ a.s.,}$$

and the factor $b_k(x_k, h_k) - b_0(x_k, I_k)$ does not depend on u_k. Thus, we can write:

$$\mathbb{E}^{\pi'}[r_k + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1})|h_k, z_k] \geq \sum_{u_k \in \mathsf{U}} \pi'(u_k|h_k) Q^\pi(y_k, z_k, u_k) \\ - \frac{r_{max}}{1-\gamma} d_{TV}\big(b_k(\cdot, h_k), b_0(\cdot, I_k)\big), \quad (51)$$

since $\|\mu - \nu\|_1 = 2d_{TV}(\mu, \nu)$ for any $\mu, \nu \in \rho(\mathsf{X})$. Substituting (51) into (48), we obtain:

$$\mathbb{E}^{\pi'}[r(x_k, u_k) + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1}) - V^\pi(y_k, z_k)|y_0, z_0] \\ \geq \mathbb{E}^{\pi'}[A^\pi(y_k, z_k, u_k) - \frac{r_{max}}{1-\gamma} d_{TV}\big(b_k(\cdot, h_k), b_0(\cdot, I_k)\big)|y_0, z_0], \quad (52)$$

where $A^\pi(y, z, u) = Q^\pi(y, z, u) - V^\pi(y, z)$. By using the dominated convergence theorem again on (52), we conclude that

$$\mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} \gamma^k \Big(r(x_k, u_k) + \gamma V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1}) - V^\pi(y_k, z_k)\Big)\Big|y_0, z_0\Big] \\ \geq \mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} A^\pi(y_k, z_k, u_k)\Big|y_0, z_0\Big] - \frac{r_{max}}{1-\gamma} \cdot \Gamma^{\pi'}(y_0, z_0). \quad (53)$$

Bounding (ii) **in** (45)**:** By following identical steps as we used in bounding (50), we obtain the following:

$$\gamma \cdot \mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} \gamma^k \Big(V^\pi(y_{k+1}, z_{k+1}) - V_0^\pi(x_{k+1}, y_{k+1}, z_{k+1})\Big|y_0, z_0\Big] \geq -\frac{r_{max}}{1-\gamma} \cdot \Gamma^{\pi'}(y_0, z_0). \quad (54)$$

Hence, we conclude that

$$V^{\pi'}(y_0, z_0) - V^\pi(y_0, z_0) \geq \mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} A^\pi(y_k, z_k, u_k)\Big|y_0, z_0\Big] - \frac{2r_{max}}{1-\gamma} \cdot \Gamma^{\pi'}(y_0, z_0).$$

□

Note that (45) holds for any $\pi' \in \Pi_A$ as we used only FSC property of the compared function $\pi \in \Pi_{F, \mathsf{Z}, \varphi}$. If π' is also a finite-state controller (possibly with a different internal state dynamic than π), we have a more useful expression that was stated in Corollary 1

Proof of Corollary 1. We want to find an alternative expression for the discounted sum of the advantage function. Since $\sup_{y,z,u} |A^\pi(y,z,u)| < \infty$ and $\gamma \in (0,1)$, by dominated convergence theorem, we have:

$$\begin{aligned}
&\mathbb{E}^{\pi'}\Big[\sum_{k=0}^{\infty} A^\pi(y_k, z_k, u_k)\Big| y_0, z_0\Big] \\
&\qquad = \sum_{y\in\mathsf{Y}}\sum_{z\in\mathsf{Z}}\sum_{u\in\mathsf{Z}}\sum_{k=0}^{\infty}\gamma^k \mathbb{P}^{\pi'}\Big(y_k = y, z_k = z, u_k = u\Big| y_0, z_0\Big) A^\pi(y,z,u), \\
&\qquad = \sum_{y\in\mathsf{Y}}\sum_{z\in\mathsf{Z}}\sum_{u\in\mathsf{Z}}\sum_{k=0}^{\infty}\gamma^k \mathbb{P}^{\pi'}\Big(y_k = y, z_k = z\Big| y_0, z_0\Big)\pi'(u|y,z) A^\pi(y,z,u),
\end{aligned}$$

where the second line holds since $\pi' \in \Pi_{F,\mathsf{Z},\varphi}$. Recall that

$$\mathsf{d}_{y_0,z_0}^{\pi'}(y,z) = (1-\gamma)\sum_{k=0}^{\infty}\mathbb{P}^{\pi'}(y_k = y, z_k = z | y_0, z_0),$$

by definition. Using this definition along with the dominated convergence theorem, we conclude the proof. □

Proof of Theorem 2. The first part of the proof is based on a Lyapunov drift result, which is an extension of the analysis provided in [14] for natural policy gradient for (fully observable) MDPs. For $\pi \in \Pi_A$, let

$$\begin{aligned}
\Lambda(\pi) &= \sum_{y\in\mathsf{Y}}\sum_{z\in\mathsf{Z}}\mathsf{d}_\xi^{\pi^*}(y,z)\sum_{u\in\mathsf{U}}\pi^*(u|y,z)\log\frac{\pi^*(u|y,z)}{\pi(u|y,z)}, \\
&= \sum_{y\in\mathsf{Y}}\sum_{z\in\mathsf{Z}}\mathsf{d}_\xi^{\pi^*}(y,z) D_{KL}(\pi^*(\cdot|y,z)||\pi(\cdot|y,z))
\end{aligned} \tag{55}$$

be the potential function, where d_ξ^π is the discounted action-observation visitation distribution under π. For any $t \geq 0$, we have the following drift:

$$\Lambda(\pi_{t+1}) - \Lambda(\pi_t) = \sum_{y\in\mathsf{Y}}\sum_{z\in\mathsf{Z}}\mathsf{d}_\xi^{\pi^*}(y,z)\sum_{u\in\mathsf{U}}\pi^*(u|y,z)\log\frac{\pi_t(u|y,z)}{\pi_{t+1}(u|y,z)}. \tag{56}$$

Note that

$$\nabla \log \pi_\theta(u|y,z) = \psi(y,z,u) - \sum_{u'\in\mathsf{U}}\pi_\theta(u'|y,z)\psi(y,z,u'),$$

which implies

$$\|\nabla \log \pi_\theta(u|y,z) - \nabla \log \pi_{\theta'}(u|y,z)\|_2 \leq 2,$$

for any $\theta, \theta' \in \mathbb{R}^d$ and u, y, z. Thus, by Taylor's theorem, we have:

$$|\log \pi_{\theta'}(u|y,z) - \log \pi_\theta(u|y,z) - \langle \nabla \log \pi_\theta(u|y,z), \theta' - \theta\rangle| \leq \|\theta - \theta'\|_2^2, \tag{57}$$

for any u, y, z and $\theta, \theta' \in \mathbb{R}^d$. This implies the following [14]:

$$\log\frac{\pi_t(u|y,z)}{\pi_{t+1}(u|y,z)} \leq \eta^2\|\bar{w}_t\|_2^2 - \eta\langle\nabla \log \pi_t(u|y,z), \bar{w}_t\rangle, \tag{58}$$

where

$$\bar{w}_t = \frac{1}{N}\sum_{k<N} w_t(k).$$

Hence,

$$\Lambda(\pi_{t+1}) - \Lambda(\pi_t) \leq -\eta\sum_{y\in\mathsf{Y}}\sum_{u\in\mathsf{U}}\sum_{z\in\mathsf{Z}}\mathsf{d}_\xi^{\pi^*}(y,z)\pi^*(u|y,z)\langle\nabla \log \pi_t(u|y,z), \bar{w}_t\rangle + \eta^2\|\bar{w}_t\|_2^2, \tag{59}$$

which leads to:

$$\Lambda(\pi_{t+1}) - \Lambda(\pi_t) \leq \eta\sqrt{\sum_{y\in\mathsf{Y}}\sum_{u\in\mathsf{U}}\sum_{z\in\mathsf{Z}} \mathsf{d}_\xi^{\pi^*}(y,z)\pi^*(u|y,z)\Big(\langle\nabla\log\pi_t(u|y,z),\bar{w}_t\rangle - A^{\pi_t}(y,z,u)\Big)^2}$$
$$+\eta^2R^2 - \eta\sum_{y\in\mathsf{Y}}\sum_{u\in\mathsf{U}}\sum_{z\in\mathsf{Z}} \mathsf{d}_\xi^{\pi^*}(y,z)\pi^*(u|y,z)A^{\pi_t}(y,z,u), \quad (60)$$

For any $t \leq T$ and $w \in \mathbb{R}^d$, let

$$L_{0,t}(w) = \mathbb{E}\big[\big(\nabla^\top \log\pi_t(u|y,z)w - A^{\pi_t}(y,z,u)\big)^2\big|\mathfrak{H}_t\big], \quad (61)$$
$$\widehat{L}_{0,t}(w) = \mathbb{E}\big[\big(\nabla^\top \log\pi_t(u|y,z)w - \overline{A}_K^{\pi_t}(y,z,u)\big)^2\big|\mathfrak{H}_t\big], \quad (62)$$

where $\mathfrak{H}_t$ is the σ-field generated by all samples used in policy optimization steps (up to and excluding t) and policy evaluation step at iteration t. Also, by Theorem 1 and Jensen's inequality, we have:

$$\mathbb{E}\big[\big(\bar{A}_K^{\pi_t}(y,z,u) - A^{\pi_t}(y,z,u)\big)^2\big|\mathfrak{H}_t'\big] \leq \epsilon_{critic}(t),$$

where $\mathfrak{H}_t'$ is the σ-field generated by all variables in the policy optimization steps before t, and

$$\epsilon_{critic}(t) = 2\Big(\sqrt{\frac{\|\beta_0-\beta_\pi\|_2^2 + M^2(\gamma,R)}{K^{1/2}(1-\gamma^m)}} + \frac{\bar{\ell}_{\mathsf{C}}(R,\Psi)}{1-\gamma^m} + \epsilon_{\mathtt{PA}}^{\pi_t}(\gamma,m,R)\Big).$$

Thus, by using the inequality $(x+y)^2 \leq 2x^2+2y^2,\ x,y\in\mathbb{R}$, we have:

$$\min_w\ \widehat{L}_{0,t}(w) \leq \min_w\ 2L_{0,t}(w) + 2\epsilon_{critic}(t). \quad (63)$$

By Theorem 14.8 in [44], the SGD iterations with the step-size choice ζ yield the following:

$$\widehat{L}_{0,t}(\bar{w}_t) \leq \epsilon_{actor} + \min_w \widehat{L}_{0,t}(w), \quad (64)$$

where

$$\epsilon_{actor} = \frac{2-\gamma}{1-\gamma}\cdot\frac{R\cdot r_{max}}{\sqrt{N}},$$

at each iteration $t \leq T$. Similarly,

$$L_{0,t}(\bar{w}_t) \leq 2\widehat{L}_{0,t}(\bar{w}_t) + 2\epsilon_{critic}(t).$$

Thus, taking expectation over the samples, we obtain:

$$\begin{aligned}\mathbb{E}[L_{0,t}(\bar{w}_t)] &\leq 2\mathbb{E}[\widehat{L}_{0,t}(w_t)] + 2\epsilon_{critic}(t),\\ &\leq 2\min_w \widehat{L}_{0,t}(w) + 2\epsilon_{actor} + 2\epsilon_{critic}(t),\\ &\leq 4\min_w L_{0,t}(w) + 2\epsilon_{actor} + 6\epsilon_{critic}(t),\end{aligned}$$

where the second line follows from the definition of ϵ_{actor} and the last line follows from (63). Since

$$\min_w L_{0,t}(w) \leq 2\bar{\ell}_{\mathsf{C}}(R,\Psi),$$

we have:

$$\sqrt{\mathbb{E}[L_{0,t}(\bar{w}_t)]} \leq 4\Big(\bar{\ell}_{\mathsf{C}}(R,\Psi) + \epsilon_{actor} + \epsilon_{critic}(t)\Big).$$

By taking expectation of the drift inequality (60), using the above inequality and Corollary 1 (performance difference lemma for FSCs), we obtain:

$$\mathbb{E}[\Lambda(\pi_{t+1}) - \Lambda(\pi_t)] \leq 4\eta\bar{C}_\infty\Big(\bar{\ell}_{\mathsf{C}}(R,\Psi) + \epsilon_{actor} + \mathbb{E}\epsilon_{critic}(t)\Big) + \eta^2R^2 - (1-\gamma)\eta\Big(\mathbb{E}\Delta_t - 2r_{max}\Gamma^{\pi^*}(\xi)\Big),$$

where $\Delta_t = V^{\pi^*}(\xi) - V^{\pi_t}(\xi)$. The proof then follows by telescoping sum over $t < T$, re-arranging the terms, and using the step-size choice in the theorem statement. □

C Memory-Inference Error Tradeoff: Proof of Proposition 1

In this section, we will prove Proposition 1. First, we provide an alternative condition for ergodicity.

C.1 Alternative Minorization-Majorization Condition

An alternative condition which leads to a similar bound in Proposition 1 is the following.

Condition 3 (Minorization-majorization)**.** *Under the optimal FSC π^*, there exist $m_0 \geq 1$, $\epsilon_0 \in (0,1)$ and a probability measure $\upsilon \in \rho(\mathsf{X} \times \mathsf{Y} \times \mathsf{Z} \times \mathsf{U})$ such that the following holds:*

$$\epsilon_0 \cdot \upsilon(x_{m_0}, I_{m_0}, u_{m_0}) \leq \mathbb{P}(x_{m_0}, I_{m_0}, u_{m_0} | x_0, I_0, u_0) \leq \frac{1}{\epsilon_0} \cdot \upsilon(x_{m_0}, I_{m_0}, u_{m_0}),$$

for any $(x_k, I_k, u_k) \in \mathsf{X} \times \mathsf{Y} \times \mathsf{Z} \times \mathsf{U}$ for $k \in \{0, m_0\}$ where $I_k = (y_k, z_k)$.

Note that $\{(x_k, y_k, z_k, u_k) : k \geq 0\}$ forms a Markov chain under the Markovian state transition dynamics of the FSC π^*, and $\mathbb{P}(x_m, I_m, u_m | x_0, I_0, u_0)$ is m-step transition matrix for the corresponding Markov chain. As such, Condition 2 is an ergodicity condition on the underlying Markov chain under π^*. Then, we have the following proposition.

Proposition 2 (Memory-performance tradeoff)**.** *Under Condition 3, we have the following result:*

$$\Gamma^{\pi^*}(\xi) \leq \frac{1}{(1-\gamma)} \cdot O\Big(\big(1-\epsilon_0^2\big)^{\lfloor \frac{n}{m_0} \rfloor}\Big), \tag{65}$$

for any $n \geq 1$.

In the next subsection, we will prove Proposition 1.

C.2 Proof of Proposition 1

The proof will follow a similar strategy described in [38]. The main specific challenge in our case is incorporating the control actions into the filter stability results. For hidden Markov chains (HMCs) considered in [38], due to the discrete memoryless observation channel Φ, the observation y_k depends only on x_k. On the other hand, in the case of POMDPs that we consider here, the controller interacts with the environment, and the data obtained from the environment is (y_k, u_{k-1}), where u_k partially depends on the observation history. This necessitates different conditions to establish filter stability, which we establish in this section.

We begin with an important lemma, which will be fundamental in the proof.

Definition 4 (Left multiplication)**.** *Let $\upsilon \in \rho(\mathsf{X})$ be a probability measure, and $K(\cdot|x) \in \rho(\mathsf{X})$ be a transition kernel. Then, we define $\upsilon K \in \rho(\mathsf{X})$ as follows:*

$$(\upsilon K)(x) = \sum_{x' \in \mathsf{X}} \upsilon(x') K(x|x'), \tag{66}$$

for any $x \in \mathsf{X}$.

Lemma 3 (Lemma 5.2 in [38])**.** *Let $\upsilon, \upsilon' \in \rho(\mathsf{X})$ be two probability mass functions on X, and $\{K(\cdot|x) \in \rho(\mathsf{X}) : x \in \mathsf{X}\}$ be a transition kernel. Then,*

1. *Left multiplication operator is 1-Lipschitz with respect to the total-variation distance d_{TV}:*

$$d_{TV}(\upsilon K, \upsilon' K) \leq d_{TV}(\upsilon, \upsilon'),$$

2. *(Minorization) If there exists a fixed $\mu \in \rho(\mathsf{X})$ and $\epsilon_0 \in (0,1)$ such that*

$$K(x|x') \geq \epsilon_0 \cdot \mu(x), \ \ \forall x, x' \in \mathsf{X},$$

then we have a contraction:

$$d_{TV}(\upsilon K, \upsilon' K) \leq (1-\epsilon_0) \cdot d_{TV}(\upsilon, \upsilon'). \tag{67}$$

Main idea: We can show that, for any $k < n$ and $h_n = (h_0, y_1^n, u^{n-1})$, the stochastic process $\{x_k : k \geq 0\}$ has a conditional Markovianity property:

$$\mathbb{P}(x_{k+1}|x^k, h_n) = \mathbb{P}(x_{k+1}|x_k, h_n), \ \ \forall x^{k+1}, h_n.$$

Hence, by following the left multiplication notation, we can express the n-step filtering transformation (8) as follows:

$$\mathbb{P}(x_n = \cdot|h_n) = \mathbb{P}(x_0 = \cdot|h_n) K_{0|n} K_{1|n} \dots K_{\lambda-1|n}, \tag{68}$$

where $n = \lambda m_0$ and

$$K_{\ell|n}(x_{(\ell+1)m_0}|x_{\ell m_0}) = \mathbb{P}(x_{(\ell+1)m_0}|x_{\ell m_0}, h_n).$$

Now, for every $\ell = 0, 1, \dots, \lambda-1$, if the transition kernel $K_{\ell|n}$ satisfies the minorization condition in Lemma 3 for fixed $\mu_{\ell|n} \in \rho(\mathsf{X})$ and $\epsilon_0 \in (0,1)$, then Lemma 3 implies that $\mathbb{P}(x_n = \cdot|h_n)$ for two different prior distributions for x_0 converges to each other in d_{TV} at a geometric rate with exponent $\lambda = n/m_0$. The key part of the proof is to show that Conditions 1-2 suffice to minorize $K_{\ell|n}$ for all $\ell \in [0, \lambda)$.

Definition 5 (Backward variable)**.** *For any $k < n$, let*

$$\beta_{k|n}(x_k, h_k; y_{k+1}^n, u_k^{n-1}) = \mathbb{P}(y_{k+1}^n, u_k^{n-1}|x_k, h_k). \tag{69}$$

Notably, it is straightforward to show that the backward variable $\beta_{k|n}$ satisfies the following recursion:

$$\begin{aligned}\beta_{k|n}&(x_k, h_k; y_{k+1}^n, u_k^{n-1}) \\ &= \sum_{x_{k+1} \in \mathsf{X}} \pi^*(u_k|y_k, z_k) P(x_{k+1}|x_k, u_k) \Phi(y_{k+1}|x_{k+1}) \beta_{k+1|n}(x_{k+1}, h_{k+1}; y_{k+2}^n, u_{k+1}^{n-1}),\end{aligned} \tag{70}$$

with $\beta_{n|n} = 1$. As such, $\beta_{k|n}$ is $\sigma(x_k, h_k)$-measurable, does not depend on x^{k-1} or the prior $b_0(\cdot) = \mathbb{P}(x_0 = \cdot|h_0)$. By using the backward variable, we can establish the following conditional Markovianity property for the hidden state $\{x_k : k \geq 0\}$ in POMDPs.

Lemma 4 (Conditional Markovianity under an FSC)**.** *For any $k < n$, $x^{k+1} \in \mathsf{X}^{k+2}$ and $h_n \in \mathsf{H} \times \mathsf{Y}^n \times \mathsf{U}^n$, we have the following:*

$$\begin{aligned}\mathbb{P}(x_{k+1}|x^k, h_n) &= \mathbb{P}(x_{k+1}|x_k, h_n), &(71)\\ &= \frac{P(x_{k+1}|x_k, u_k)\Phi(y_{k+1}|x_{k+1})\beta_{k+1|n}(x_{k+1}, h_{k+1}; y_{k+2}^n, u_{k+1}^n)}{\sum_{x'_{k+1} \in \mathsf{X}} P(x'_{k+1}|x_k, u_k)\Phi(y_{k+1}|x'_{k+1})\beta_{k+1|n}(x'_{k+1}, h_{k+1}; y_{k+2}^n, u_{k+1}^n)}, &(72)\\ &= \tilde{\kappa}_{k|n}(x_{k+1}|x_k). &(73)\end{aligned}$$

Based on Lemma 4, for any $m_0 \geq 1$, we can establish a conditional version of the Chapman-Kolmogorov equation for POMDPs:

$$\begin{aligned}\mathbb{P}(x_{(\ell+1)m_0}|x_{\ell m_0}, h_n) &= (\tilde{\kappa}_{\ell m_0|n} \tilde{\kappa}_{\ell m_0+1|n} \dots \tilde{\kappa}_{(\ell+1)m_0|n})(x_{(\ell+1)m_0}|x_{\ell m_0}), &(74)\\ &=: \kappa_{\ell|n}^{m_0}(x_{(\ell+1)m_0}|x_{\ell m_0}), &(75)\end{aligned}$$

For a given (potentially random) prior $v_0 = \mathbb{P}(x_0 = \cdot|h_0) \in \rho(\mathsf{X})$, for $n = \lambda m_0$, we can write the following:

$$\mathbb{P}_{v_0}(x_n = \cdot|h_n) = \phi(v_0, h_n)\kappa_{0|n}^{m_0}\kappa_{1|n}^{m_0}\dots\kappa_{\lambda-1|n}^{m_0}, \tag{76}$$

where

$$\begin{aligned}\phi(v_0, h_n)(x) &= \mathbb{P}(x_0 = x|h_n), &(77)\\ &= \frac{v_0(x)\cdot\beta_{0|n}(x, h_0; y_1^n, u^{n-1})}{\sum_{x'\in\mathsf{X}} v_0(x')\cdot\beta_{0|n}(x', h_0; y_1^n, u^{n-1})}, &(78)\end{aligned}$$

is the posterior distribution of x_0 given the observation history $h_n = (h_0, y_1^n, u^{n-1})$. By using this, for two prior distributions $v_0, v_0' \in \rho(\mathsf{X})$, we want to bound $d_{TV}\left(\mathbb{P}_{v_0'}(x_n = \cdot|h_n), \mathbb{P}_{v_0}(x_n = \cdot|h_n)\right)$. In the following, we show that Conditions 1-2 lead to the minorization of the m_0-step transition kernels $\kappa_{\ell|n}^{m_0}$ for $m_0 \geq 1$ specified in Condition 2.

Lemma 5 (Minorization of the smoothing kernel)**.** *Under Conditions 1-2, there exist $\epsilon_0 \in (0,1)$ and a probability measure $\nu_{\ell|n} \in \rho(\mathsf{X})$ for any $\ell \in (0,\lambda)$ such that the following holds:*

$$\kappa_{\ell|n}^{m_0}(x_{(\ell+1)m_0}|x_{\ell m_0}) \geq \alpha^{2m_0-2}\cdot\epsilon_0^2\cdot\nu_{\ell|n}(x_{(\ell+1)m_0}), \tag{79}$$

for all $x_{\ell m_0}, x_{(\ell+1)m_0} \in \rho(\mathsf{X})$ given $h_n \in \mathsf{H}\times\mathsf{Y}^n\times\mathsf{U}^n$.

Proof. First, notice that we have the following:

$$\mathbb{P}(x_{(\ell+1)m_0}|x_{\ell m_0}, h_n) = \frac{\mathbb{P}\big(x_{(\ell+1)m_0}, x_{\ell m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}, y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}, h_{\ell m_0}\big)}{\sum_{x'_{(\ell+1)m_0}\in\mathsf{X}}\mathbb{P}\big(x_{(\ell+1)m_0'}, x_{\ell m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}, y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}, h_{\ell m_0}\big)}, \tag{80}$$

where we decomposed $h_n = (y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}, y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}, h_{\ell m_0})$. We can expand the numerator of (80) as follows:

$$\begin{aligned}\mathbb{P}\big(x_{(\ell+1)m_0}, x_{\ell m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}, y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}, h_{\ell m_0}\big) &= \mathbb{P}\big(x_{\ell m_0}, h_{\ell m_0}\big)\\ \times\mathbb{P}\big(y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}\big|x_{(\ell+1)m_0}, h_{(\ell+1)m_0}\big)\cdot\mathbb{P}(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, &u_{\ell m_0}^{(\ell+1)m_0-1}|x_{\ell m_0}, h_{\ell m_0}), \quad (81)\end{aligned}$$

where the first term on the RHS of the above identity follows from:

$$\mathbb{P}\big(y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}\big|x_{(\ell+1)m_0}, h_{(\ell+1)m_0}\big) = \mathbb{P}\big(y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}\big|x_{(\ell+1)m_0}, x_{\ell m_0}, h_{(\ell+1)m_0}\big). \tag{82}$$

The above identity is true since (i) $u_{(\ell+1)m_0}$ is $\sigma(h_{(\ell+1)m_0})$-measurable, (ii) $\{x_k : k \geq 0\}$ is a controlled Markov chain, (iii) Φ is a discrete memoryless channel. From Definition 5, we observe that

$$\mathbb{P}\big(y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}\big|x_{(\ell+1)m_0}, h_{(\ell+1)m_0}\big) = \beta_{(\ell+1)m_0|n}(x_{(\ell+1)m_0}, h_{(\ell+1)m_0}; y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}). \tag{83}$$

Thus, substituting the above identity and (81) into (80), we obtain the following:

$$\begin{aligned}&\kappa_{\ell|n}^{m_0}(x_{(\ell+1)m_0}|x_{\ell m_0})\\ &\quad= \frac{\mathbb{P}(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}|x_{\ell m_0}, h_{\ell m_0})\mathbb{P}\big(y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}\big|x_{(\ell+1)m_0}, h_{(\ell+1)m_0}\big)}{W(x_{\ell m_0}, h_n)}, \quad (84)\end{aligned}$$

where

$$\begin{aligned}W(x_{\ell m_0}, h_n) = \sum_{x'_{(\ell+1)m_0}}\Big[&\mathbb{P}(x'_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}|x_{\ell m_0}, h_{\ell m_0})\\ &\times\mathbb{P}\big(y_{(\ell+1)m_0+1}^n, u_{(\ell+1)m_0}^{n-1}\big|x'_{(\ell+1)m_0}, h_{(\ell+1)m_0}\big)\Big].\end{aligned}$$

Now, we will use Conditions 1-2 to show that $\mathbb{P}(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}|x_{\ell m_0}, h_{\ell m_0})$ minorizes and majorizes simultaneously, which will let us show that $\kappa_{\ell|n}^{m_0}$ minorizes. First, note that we can perform the following expansion:

$$\begin{aligned}\mathbb{P}(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}|x_{\ell m_0}, h_{\ell m_0})&\\ = \sum_{x_{\ell m_0+1}^{(\ell+1)m_0-1}} \prod_{j=\ell m_0}^{(\ell+1)m_0-1} \pi^*(u_j|y_j, z_j) P(x_{j+1}|x_j, u_j)\Phi(y_{j+1}|x_{j+1}).&\end{aligned} \tag{85}$$

For any probability measure $\mu \in \rho(\mathsf{U})$, let

$$\mathbb{P}^{\mu}(x_{m_0}, y_1^{m_0}, u^{m_0-1}|x_0) = \sum_{x_1^{m_0-1}} \prod_{j=0}^{m_0-1} \mu(u_j) P(x_{j+1}|x_j, u_j)\Phi(y_{j+1}|x_{j+1}). \tag{86}$$

Then, under Condition 1, we have the following inequalities:

$$\begin{aligned}\alpha^{m_0-1}\mathbb{P}^{\bar{\mu}}(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}|x_{\ell m_0}) &\le \mathbb{P}(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}|x_{\ell m_0}, h_{\ell m_0}),\\ &\le \frac{1}{\alpha^{m_0-1}}\mathbb{P}^{\bar{\mu}}(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}|x_{\ell m_0}).\end{aligned} \tag{87}$$

Furthermore, Condition 2 implies that

$$\begin{aligned}\epsilon_0 \cdot \nu\big(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}\big) &\le \mathbb{P}^{\bar{\mu}}(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}|x_{\ell m_0}),\\ &\le \frac{1}{\epsilon_0}\cdot \nu\big(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}\big).\end{aligned} \tag{88}$$

Combining (87) and (88), we obtain the following simultaenous minorization-majorization result:

$$\begin{aligned}\epsilon_0\alpha^{m_0-1}\nu\big(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}\big) &\le \mathbb{P}(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}|x_{\ell m_0}, h_{\ell m_0}),\\ &\le \frac{1}{\epsilon_0\alpha^{m_0-1}}\nu\big(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}\big),\end{aligned} \tag{89}$$

for any $(x_{\ell m_0}, h_{\ell m_0})$. By using the lower bound and upper bound in (89), we obtain the following bound for (84):

$$\kappa_{\ell|n}^{m_0}(x_{(\ell+1)m_0}|x_{\ell m_0}) \ge (\epsilon_0\alpha^{m_0-1})^2 \cdot \nu_{\ell|n}^{m_0}(x_{(\ell+1)m_0}), \tag{90}$$

where

$$\nu_{\ell|n}^{m_0}(x_{(\ell+1)m_0}) = \frac{\nu(x_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1}) \cdot \mathbb{P}\big(y_{(\ell+1)m_0+1}^{n}, u_{(\ell+1)m_0}^{n-1}\big|x_{(\ell+1)m_0}, h_{(\ell+1)m_0}\big)}{\sum_{x'_{(\ell+1)m_0}} \nu(x'_{(\ell+1)m_0}, y_{\ell m_0+1}^{(\ell+1)m_0}, u_{\ell m_0}^{(\ell+1)m_0-1})\mathbb{P}\big(y_{(\ell+1)m_0+1}^{n}, u_{(\ell+1)m_0}^{n-1}\big|x'_{(\ell+1)m_0}, h_{(\ell+1)m_0}\big)}.$$

From the discussion in (82), we can directly deduce that $\nu_{\ell|n}^{m_0}(x_{(\ell+1)m_0})$ does *not* depend on $x_{\ell m_0}$. Hence, we conclude the proof. □

Proof of Proposition 1. Let $\lambda = \lfloor n/m_0 \rfloor$. For a given prior distribution $v_0 \in \rho(\mathsf{X})$, recall the definition of the posterior:

$$\phi(v_0, h_n)(x) = \frac{v_0(x)\cdot\beta_{0|n}(x, h_0; y_1^n, u^{n-1})}{\sum_{x'\in\mathsf{X}} v_0(x')\cdot\beta_{0|n}(x', h_0; y_1^n, u^{n-1})}. \tag{91}$$

By using (66), we can express:

$$F^{(n)}(v_0, y_1^n, u_0^{n-1}) = \phi(v_0, h_n)\kappa_{0|n}^{m_0}\kappa_{1|n}^{m_0}\dots\kappa_{\lambda-1|n}^{m_0}, \tag{92}$$

where $\kappa_{\ell|n}^{m_0}$ is the smoothing kernel in (84). Thus, for two (potentially random) prior distributions $v_0, v_0' \in \rho(\mathsf{X})$, we have:

$$
\begin{aligned}
d_{TV}(F^{(n)}(v_0, y_1^n, &u_0^{n-1}), F^{(n)}(v_0', y_1^n, u_0^{n-1})) \\
&= d_{TV}(\phi(v_0, h_n)\kappa_{0|n}^{m_0} \dots \kappa_{\lambda-1|n}^{m_0}, \phi(v_0', h_n)\kappa_{0|n}^{m_0} \dots \kappa_{\lambda-1|n}^{m_0}), \\
&\leq (1 - (\epsilon_0 \alpha^{m_0-1})^2) d_{TV}(\phi(v_0, h_n)\kappa_{0|n}^{m_0} \dots \kappa_{\lambda-2|n}^{m_0}, \phi(v_0', h_n)\kappa_{0|n}^{m_0} \dots \kappa_{\lambda-2|n}^{m_0}), \\
&\vdots \\
&\leq (1 - (\epsilon_0 \alpha^{m_0-1})^2)^{\lambda} d_{TV}(\phi(v_0, h_n), \phi(v_0', h_n)),
\end{aligned}
$$

where all inequalities are obtained by the successive applications of the contraction result in Lemma 3 and the minorization result for the smoothing kernels $\kappa_{\ell|n}^{m_0}$ proved in Lemma 5. Hence, we conclude that

$$
d_{TV}(F^{(n)}(v_0, y_1^n, u_0^{n-1}), F^{(n)}(v_0', y_1^n, u_0^{n-1})) = O\Big(\big(1 - (\epsilon_0 \alpha^{m_0-1})^2\big)^{\lfloor \frac{n}{m_0} \rfloor}\Big).
$$

Note that v_0, v_0' can be random, and they may depend on the history of the decision process. Hence, the above result implies the following:

$$
d_{TV}(F^{(n)}(b_{-n}(\cdot|y_{k-n}), y_{k-n+1}^k, u_{k-n}^{k-1}), F^{(n)}(b_{k-n}(\cdot|y_{k-n}), y_{k-n+1}^k, u_{k-n}^{k-1}) = O\Big(\big(1 - (\epsilon_0 \alpha^{m_0-1})^2\big)^{\lfloor \frac{n}{m_0} \rfloor}\Big), \quad (93)
$$

for all $k \geq 0$, which concludes the proof. □