

---

# The Importance of Non-Markovianity in Maximum State Entropy Exploration

---

Mirco Mutti<sup>\*12</sup> Riccardo De Santi<sup>\*13</sup> Marcello Restelli<sup>1</sup>

## Abstract

In the maximum state entropy exploration framework, an agent interacts with a reward-free environment to learn a policy that maximizes the entropy of the expected state visitations it is inducing. Hazan et al. (2019) noted that the class of Markovian stochastic policies is sufficient for the maximum state entropy objective, and exploiting non-Markovianity is generally considered pointless in this setting.

In this paper, we argue that non-Markovianity is instead paramount for maximum state entropy exploration in a finite-sample regime. Especially, we recast the objective to target the expected entropy of the induced state visitations in a single trial. Then, we show that the class of non-Markovian deterministic policies is sufficient for the introduced objective, while Markovian policies suffer non-zero regret in general. However, we prove that the problem of finding an optimal non-Markovian policy is NP-hard. Despite this negative result, we discuss avenues to address the problem in a tractable way in future works.

## 1. Introduction

Several recent works have addressed *Maximum State Entropy* (MSE) exploration (Hazan et al., 2019; Tarbouriech & Lazaric, 2019; Lee et al., 2019; Mutti & Restelli, 2020; Mutti et al., 2021b;a; Zhang et al., 2020a; Guo et al., 2021; Liu & Abbeel, 2021b;a; Seo et al., 2021; Yarats et al., 2021) as a pre-training objective for online Reinforcement Learning (RL) (Sutton & Barto, 2018). In this line of work, an agent interacts with a reward-free environment (Jin et al., 2020) in order to learn a general exploration strategy. The aim of this strategy is to improve the sample efficiency of

any RL task that could be specified over the same environment afterwards, serving as an exploratory initialization to standard learning techniques, such as Q-learning (Watkins & Dayan, 1992) or policy gradient (Peters & Schaal, 2008). To learn this strategy, the agent maximizes an entropic measure of the state distribution induced by its behavior over the environment, effectively targeting a uniform exploration of the state space. In tabular domains, optimizing this kind of MSE objective is known to be provably efficient (Hazan et al., 2019; Zhang et al., 2020b), whereas the obtained exploratory strategy lead to outstanding empirical results in continuous and high-dimensional domains as well, (e.g., Mutti et al., 2021b; Liu & Abbeel, 2021b).

All of the existing works pursuing a MSE objective solely focus on Markovian exploration strategies, in which each decision is conditioned on the current state of the environment rather than the full history of interactions. This choice is common in RL, as it is well-known that an optimal deterministic Markovian strategy maximizes the usual cumulative sum of rewards objective (Puterman, 2014). Hazan et al. (2019, Lemma 3.3) note that the class of Markovian strategies is *sufficient* for the standard MSE objective as well. Indeed, a carefully constructed Markovian strategy is able to induce the same state distribution of any history-based (non-Markovian) one by exploiting randomization. This result does not hold only for asymptotic state distributions (Puterman, 2014), but also for state distributions that are marginalized over a finite horizon. Thus, there is little incentive to consider more complicated strategies as they are not providing any benefit for the entropy objective.

However, the intuition suggests that exploiting the history of the interactions is useful when the agent’s goal is to uniformly explore the environment. To this point, let us consider an illustrative example in which the agent finds itself in the middle of a two-rooms domain (as depicted in Figure 1), having a budget of interactions that is just enough to visit every state within a single episode. It is easy to see that an optimal Markovian strategy for the MSE objective would randomize between going left and right in the initial position, and then would follow the optimal route within a room, finally ending in the initial position again. An episode either results in visiting the left room

---

<sup>\*</sup>Equal contribution <sup>1</sup>Politecnico di Milano, Milan, Italy <sup>2</sup>Università di Bologna, Bologna, Italy <sup>3</sup>ETH Zurich, Zurich, Switzerland. Correspondence to: Mirco Mutti <mirco.mutti@polimi.it>.

In Section 3, we extend known results (Puterman, 2014) to show that the class of Markovian strategies is sufficient for any infinite-samples MSE objective. Then, in Section 4, we propose a novel finite-sample MSE objective and a corresponding regret formulation. Especially, we prove that the class of non-Markovian strategies is sufficient for the introduced objective, whereas the optimal Markovian strategy suffers a non-zero regret. However, in Section 5, we show that the problem of finding an optimal non-Markovian strategy for the finite-sample MSE objective is NP-hard. Despite the hardness result, we comment some potential options to address the problem in a tractable way (Appendix C). The proofs of the theorems can be found in Appendix A.

**Markov Decision Processes** A CMP  $\mathcal{M}$  paired with a reward function  $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is called a Markov Decision Process (MDP) (Puterman, 2014)  $\mathcal{M}^R := \mathcal{M} \cup R$ . We denote with  $R(s, a)$  the expected immediate reward, and with  $R(h) = \sum_{t \in [T]} R(s_t, a_t)$  the return over the

horizon  $T$ . the *average return* of a policy  $\pi$  over the MDP  $\mathcal{M}^R$  is defined as  $\mathcal{J}_{\mathcal{M}^R}(\pi) = \mathbb{E}_{h \sim p_T^\pi}[R(h)]$ , and  $\pi_{\mathcal{J}}^* \in \arg \max_{\pi \in \Pi} \mathcal{J}_{\mathcal{M}^R}(\pi)$  is called an optimal policy.

**Extended MDP** The problem of finding an optimal non-Markovian policy with history-length  $T$  in an MDP  $\mathcal{M}^R$ , i.e.,  $\pi_{\text{NM}}^* \in \arg \max_{\pi \in \Pi_{\text{NM}}} \mathcal{J}_{\mathcal{M}^R}(\pi)$ , can be reformulated as the one of finding an optimal Markovian policy  $\pi_{\text{M}}^* \in \arg \max_{\pi \in \Pi_{\text{M}}} \mathcal{J}_{\widetilde{\mathcal{M}}_T^R}(\pi)$  in an extended MDP  $\widetilde{\mathcal{M}}_T^R := (\widetilde{\mathcal{S}}, \widetilde{\mathcal{A}}, \widetilde{P}, \widetilde{R}, \widetilde{\mu})$ , in which  $\widetilde{\mathcal{S}} \subseteq \mathcal{H}_{[T]} = \mathcal{H}_1 \cup \dots \cup \mathcal{H}_T$ , and  $\widetilde{s} := (\widetilde{s}_0, \dots, \widetilde{s}_{-1})$  corresponds to a history in  $\mathcal{M}^R$  of length  $|\widetilde{s}|$ ,  $\widetilde{\mathcal{A}} = \mathcal{A}$ ,  $\widetilde{P}(\widetilde{s}'|\widetilde{s}, \widetilde{a}) = P(s' = \widetilde{s}'_{-1}|s = \widetilde{s}_{-1}, a = \widetilde{a})$ ,  $\widetilde{R}(\widetilde{s}, \widetilde{a}) = R(s = \widetilde{s}_{-1}, a = \widetilde{a})$ , and  $\widetilde{\mu}(\widetilde{s}) = \mu(s = \widetilde{s})$  for any  $\widetilde{s} \in \widetilde{\mathcal{S}}$  of unit length.

### 3. Infinite Samples: Non-Markovianity Does Not Matter

Previous works pursuing maximum state entropy exploration of a CMP consider an objective function of the kind:

$$\mathcal{E}_\infty(\pi) := E(d^\pi(\cdot)) = - \mathbb{E}_{s \sim d^\pi} [\log d^\pi(s)], \quad (1)$$

where  $d^\pi(\cdot)$  is either a stationary state distribution (Mutti & Restelli, 2020), a discounted state distribution (Hazan et al., 2019; Tarbouriech & Lazaric, 2019), or a marginal state distribution (Lee et al., 2019; Mutti et al., 2021b). Hazan et al. (2019, Lemma 3.3) prove that the set of (randomized) Markovian policies  $\Pi_{\text{M}}$  is indeed sufficient for  $\mathcal{E}_\infty$  defined over asymptotic state distributions. In the following theorem and corollary, we build upon common MDP results (Puterman, 2014) to show that  $\Pi_{\text{M}}$  suffices for  $\mathcal{E}_\infty$  defined over (non-asymptotic) marginal state distributions as well.

**Theorem 3.1.** *Let  $x \in \{\infty, \gamma, T\}$ , and let  $\mathcal{D}_{\text{NM}}^x = \{d_x^\pi(\cdot) : \pi \in \Pi_{\text{NM}}\}$ ,  $\mathcal{D}_{\text{NS}}^x = \{d_x^\pi(\cdot) : \pi \in \Pi_{\text{NS}}\}$ ,  $\mathcal{D}_{\text{M}}^x = \{d_x^\pi(\cdot) : \pi \in \Pi_{\text{M}}\}$  the corresponding sets of distributions. We can prove that (i) the sets of stationary state distributions are equivalent  $\mathcal{D}_{\text{NM}}^\infty \equiv \mathcal{D}_{\text{NS}}^\infty \equiv \mathcal{D}_{\text{M}}^\infty$ , (ii) the sets of discounted state distributions are equivalent  $\mathcal{D}_{\text{NM}}^\gamma \equiv \mathcal{D}_{\text{NS}}^\gamma \equiv \mathcal{D}_{\text{M}}^\gamma$  for any  $\gamma$ , (iii) the sets of marginal state distributions are equivalent  $\mathcal{D}_{\text{NM}}^T \equiv \mathcal{D}_{\text{NS}}^T \equiv \mathcal{D}_{\text{M}}^T$  for any  $T$ .*

From the equivalence of the sets of induced distributions, it is straightforward to derive the following corollary.

**Corollary 3.2.** *For every CMP  $\mathcal{M}$ , there exists a Markovian policy  $\pi^* \in \Pi_{\text{M}}$  such that  $\pi^* \in \arg \max_{\pi \in \Pi} \mathcal{E}_\infty(\pi)$ .*

As a consequence of Corollary 3.2, there is little incentive to consider non-Markovian (or non-stationary) policies when optimizing objective (1). In this paper, we show that this underlying infinite-sample regime is the reason why the benefit of non-Markovianity, albeit backed up by intuition, does not matter. Instead, we propose a relevant finite-sample entropy objective in which non-Markovianity is crucial.

### 4. Finite Samples: Non-Markovianity Matters

In this section, we reformulate the typical maximum state entropy exploration objective of a CMP (1) to account for a finite-sample regime. Crucially, we consider the expected entropy of the state visitation frequency rather than the entropy of the expected state visitation frequency, i.e.,

$$\mathcal{E}(\pi) := \mathbb{E}_{h \sim p_T^\pi} [E(d_h(\cdot))] = - \mathbb{E}_{h \sim p_T^\pi} \mathbb{E}_{s \sim d_h} [\log d_h(s)]. \quad (2)$$

We note that  $\mathcal{E}(\pi) \leq \mathcal{E}_\infty(\pi)$  for any  $\pi \in \Pi$ , which is trivial by the concavity of the entropy function and the Jensen's inequality. Whereas (2) is ultimately an expectation as it is (1), the entropy is not computed over the infinite-sample state distribution  $d_T^\pi(\cdot)$  but its finite-sample realization  $d_h(\cdot)$ . Thus, to maximize  $\mathcal{E}(\pi)$  we have to find a policy inducing high-entropy state visits within a single trajectory rather than high-entropy state visits over infinitely many trajectories. Crucially, while Markovian policies are as powerful as any other policy class in terms of induced state distributions (Theorem 3.1), this is no longer true when looking at induced trajectory distributions  $p_T^\pi$ . Indeed, we show that non-Markovianity provides a superior policy class for objective (2). First, we define a performance measure to formally assess this benefit, which we call *regret-to-go*.<sup>1</sup>

**Definition 4.1** (Expected Regret-to-go). *Consider a policy  $\pi \in \Pi$  interacting with a CMP over  $T$  steps. We define the expected regret-to-go  $\mathcal{R}_{T-t}$  at step  $t$  (i.e., from step  $t$  onwards) of  $\pi$  as*

$$\mathcal{R}_{T-t}(\pi) = \max_{\pi^* \in \Pi} \mathbb{E}_{h_{T-t}^* \sim p_{T-t}^{\pi^*}} [E(d_{h_{T-t}^*}(\cdot))] - \mathbb{E}_{h_{T-t} \sim p_{T-t}^\pi} [E(d_{h_{T-t}}(\cdot))],$$

where  $h_T^* = (h_t^*, h_{T-t}^*)$ ,  $h_T = (h_t, h_{T-t})$  are concatenations of the  $(T-t)$ -step trajectories  $h_{T-t}^*$ ,  $h_{T-t}$  (starting from  $s_{t, h_t^*}$ ) and the  $t$ -step optimal trajectory  $h_t^* \sim p^{\pi^*}$  respectively. The term  $R_T(\pi)$  denotes the expected regret-to-go of a  $T$ -step trajectory  $h_T$  starting from  $s \sim \mu$ .

The intuition behind the regret-to-go is quite simple. Suppose to have drawn a zero-regret trajectory  $h_t^*$  upon step  $t$ . If we take the subsequent action with the (possibly sub-optimal) policy  $\pi$ , by how much would we decrease (in expectation) the entropy of the state visits  $E(d_{h_T}(\cdot))$  w.r.t. an optimal policy  $\pi^*$ ? In particular, we would like to know how limiting the policy  $\pi$  to a specific policy class would affect the expected regret-to-go and the value of  $\mathcal{E}(\pi)$  we could achieve. The following lemma shows that an optimal non-Markovian policy suffers zero expected regret-to-go.

<sup>1</sup>Note that the entropy function does not enjoy additivity, thus we cannot adopt the usual expected cumulative regret formulation.

**Lemma 4.2.** *For every CMP  $\mathcal{M}$ , there exists a deterministic non-Markovian policy  $\pi_{\text{NM}} \in \Pi_{\text{NM}}^{\text{D}}$  such that  $\pi_{\text{NM}} \in \arg \max_{\pi \in \Pi_{\text{NM}}} \mathcal{E}(\pi)$ , which suffers expected regret-to-go  $\mathcal{R}_{T-t}(\pi_{\text{NM}}) = 0, \forall t \in [T]$ .*

*Proof.* The result  $\mathcal{R}_{T-t}(\pi_{\text{NM}}) = 0$  is straightforward by noting that the set of non-Markovian policies  $\Pi_{\text{NM}}$  with arbitrary history-length is as powerful as the general set of policies  $\Pi$ . To show that there exists a deterministic  $\pi_{\text{NM}}$ , we consider the extended MDP  $\widetilde{\mathcal{M}}_T^R$  obtained from the CMP  $\mathcal{M}$  as in Section 2, in which the extended reward function is  $\widetilde{R}(\widetilde{s}, \widetilde{a}) = \text{Entropy}(d_{\widetilde{s}}(\cdot))$  for every  $\widetilde{a} \in \widetilde{\mathcal{A}}$  and every  $\widetilde{s} \in \widetilde{\mathcal{S}}$  such that  $|\widetilde{s}| = T$ , and  $\widetilde{R}(\widetilde{s}, \widetilde{a}) = 0$  otherwise. Since a Markovian policy  $\widetilde{\pi}_{\text{M}} \in \Pi_{\text{M}}^{\text{D}}$  on  $\widetilde{\mathcal{M}}_T^R$  can be mapped to a non-Markovian policy  $\pi_{\text{NM}} \in \Pi_{\text{NM}}^{\text{D}}$  on  $\mathcal{M}$ , and it is well-known (Puterman, 2014) that for any MDP there exists an optimal deterministic Markovian policy, we have that  $\widetilde{\pi}_{\text{M}} \in \arg \max_{\pi \in \Pi_{\text{M}}} \mathcal{J}_{\widetilde{\mathcal{M}}_T^R}(\pi)$  implies  $\pi_{\text{NM}} \in \arg \max_{\pi \in \Pi_{\text{NM}}} \mathcal{E}(\pi)$ .  $\square$

Whereas Lemma 4.2 ensures that the set of non-Markovian policies  $\Pi_{\text{NM}}$  is sufficient for the objective (2), we would like to know if it is also necessary. Especially, we aim to assess whether there exist CMPs in which a Markovian policy  $\pi \in \Pi_{\text{M}}$  would suffer non-zero regret-to-go. First, it is worth showing that Markovian policies can rely on randomization to optimize objective (2).

**Lemma 4.3.** *Let  $\pi_{\text{NM}} \in \Pi_{\text{NM}}^{\text{D}}$  be a non-Markovian policy such that  $\pi_{\text{NM}} \in \arg \max_{\pi \in \Pi} \mathcal{E}(\pi)$  on a CMP  $\mathcal{M}$ . The variance of an optimal Markovian policy  $\pi_{\text{M}} \in \arg \max_{\pi \in \Pi_{\text{M}}} \mathcal{E}(\pi)$  is given by*

$$\text{Var}_{a \sim \pi_{\text{M}}(s)} [a] = \text{Var}_{hs \sim p_T^{\pi_{\text{NM}}}} [\pi_{\text{NM}}(hs)], \quad \forall s \in \mathcal{S},$$

where  $hs \in \mathcal{H}_{[T]}$  is any history with final state  $s$ .

*Proof Sketch.* We can prove the result through the Law of Total Variance (LoTV) (Bertsekas & Tsitsiklis, 2002), which gives  $\text{Var}_{a \sim \pi_{\text{M}}(s)} [a] = \mathbb{E}_{hs \sim p_T^{\pi_{\text{NM}}}} [\text{Var}_{a \sim \pi_{\text{NM}}(hs)} [a]] + \text{Var}_{hs \sim p_T^{\pi_{\text{NM}}}} [\mathbb{E}_{a \sim \pi_{\text{NM}}(hs)} [a]]$ ,  $\forall s \in \mathcal{S}$ . Then, exploiting the determinism of  $\pi_{\text{NM}}$  (through Lemma 4.2), it is straightforward to see that  $\mathbb{E}_{hs \sim p_T^{\pi_{\text{NM}}}} [\text{Var}_{a \sim \pi_{\text{NM}}(hs)} [a]] = 0$  and that  $\text{Var}_{hs \sim p_T^{\pi_{\text{NM}}}} [\mathbb{E}_{a \sim \pi_{\text{NM}}(hs)} [a]] = \text{Var}_{hs \sim p_T^{\pi_{\text{NM}}}} [\pi_{\text{NM}}(hs)]$ , which concludes the proof.  $\square$

Especially, Lemma 4.3 shows that, whenever the optimal strategy for objective (2) (i.e.,  $\pi_{\text{NM}}$ ) requires to adapt its decision in a state  $s$  according to the history that led to it ( $hs$ ), an optimal Markovian policy for the same objective (i.e.,  $\pi_{\text{M}}$ ) must be randomized. This is crucial to prove the main result of this section, which establishes a lower  $\underline{\mathcal{R}}_{T-t}$  and an upper  $\overline{\mathcal{R}}_{T-t}$  bound to the expected regret-to-go of any Markovian policy that optimizes objective (2).

**Theorem 4.4.** *Let  $\pi_{\text{M}} \in \Pi_{\text{M}}$  be a Markovian policy  $\pi_{\text{M}} \in \arg \max_{\pi \in \Pi_{\text{M}}} \mathcal{E}(\pi)$  on a CMP  $\mathcal{M}$ . Then, for any  $t \in [T]$ , it holds  $\underline{\mathcal{R}}_{T-t}(\pi_{\text{M}}) \leq \mathcal{R}_{T-t}(\pi_{\text{M}}) \leq \overline{\mathcal{R}}_{T-t}(\pi_{\text{M}})$  such that*

$$\begin{aligned} \underline{\mathcal{R}}_{T-t}(\pi_{\text{M}}) &= \frac{E_{\max} - E_{\max,2}}{\pi_{\text{M}}(a_{\text{NM}}|s_t)} \text{Var}_{hs_t \sim p_T^{\pi_{\text{NM}}}} [\pi_{\text{NM}}(hs_t)], \\ \overline{\mathcal{R}}_{T-t}(\pi_{\text{M}}) &= \frac{E_{\max} - E_{\min,t}}{\pi_{\text{M}}(a_{\text{NM}}|s_t)} \text{Var}_{hs_t \sim p_T^{\pi_{\text{NM}}}} [\pi_{\text{NM}}(hs_t)], \end{aligned}$$

where  $\pi_{\text{NM}} \in \arg \max_{\pi \in \Pi_{\text{NM}}^{\text{D}}} \mathcal{E}(\pi)$ ,  $a_{\text{NM}} = \pi_{\text{NM}}(h_t^*)$ , and

$$\begin{aligned} E_{\max} &= \max_{\pi^* \in \Pi} \mathbb{E}_{h_{T-t}^* \sim p_T^{\pi^*}} [E(d_{h_{T-t}^*}(\cdot))], \\ E_{\min,t} &= \min_{h \in \mathcal{H}_{T-t}} E(d_{(h_t^*, h)}(\cdot)), \\ E_{\max,2} &= \max_{h \in \mathcal{H}_{T-t} \setminus \mathcal{H}_{T-t}^*} E(d_{(h_t^*, h)}(\cdot)) \\ \text{s.t. } \mathcal{H}_{T-t}^* &= \arg \max_{h \in \mathcal{H}_{T-t}} E(d_{(h_t^*, h)}(\cdot)). \end{aligned}$$

*Proof Sketch.* The crucial idea to derive lower and upper bounds to the regret-to-go is to consider the impact of a sub-optimal action in the best-case and the worst-case CMP respectively (see Lemma A.1, A.2). This gives  $\mathcal{R}_{T-t}(\pi_{\text{M}}) \geq E_{\max} - \pi_{\text{M}}(a_{\text{NM}}|s_t)E_{\max} - (1 - \pi_{\text{M}}(a_{\text{NM}}|s_t))E_{\max,2}$  and  $\mathcal{R}_{T-t}(\pi_{\text{M}}) \leq E_{\max} - \pi_{\text{M}}(a_{\text{NM}}|s_t)E_{\max} - (1 - \pi_{\text{M}}(a_{\text{NM}}|s_t))E_{\min,t}$ . Then, by combining the variance of the Bernoulli distribution that controls the event of taking a sub-optimal action together with Lemma 4.3, we get  $\text{Var}_{a \sim \pi_{\text{NM}}(s_t)} [a] = \pi_{\text{M}}(a_{\text{NM}}|s_t)(1 - \pi_{\text{M}}(a_{\text{NM}}|s_t)) = \text{Var}_{hs_t \sim p_T^{\pi_{\text{NM}}}} [\pi_{\text{NM}}(hs_t)]$ , which concludes the proof.  $\square$

Theorem 4.4 basically states that, whenever the Markovian policy ( $\pi_{\text{M}}$ ) has to randomize its strategy at the step  $t$ , it will suffer non-zero regret-to-go under the assumption that  $\mathcal{M}$  admits a unique optimal action  $a_{\text{NM}}$  in  $s_t$ .<sup>2</sup> Clearly, the value of this regret is related to the probability of taking a sub-optimal action (through  $1/\pi_{\text{M}}(a_{\text{NM}}|s_t)$ ) and the variance of the optimal strategy in  $s_t$  over the possible histories (through  $\text{Var}_{hs_t \sim p_T^{\pi_{\text{NM}}}} [\pi_{\text{NM}}(hs_t)]$ ). To compute the regret-to-go exactly, one should have access to the full CMP  $\mathcal{M}$ . Instead, we provide upper and lower bounds to the regret that only depend on the length of the remaining interaction  $T - t$ . While at the final step  $t = T$  a policy cannot suffer any regret, the earlier it pulls a sub-optimal action in the interaction process, the more it might incur in a bad entropy over this trajectory. The factor  $E_{\max} - E_{\min,t}$  in the upper bound quantifies how badly it can go in the worst possible CMP, in which the agent never recovers from a sub-optimal action. Instead,  $E_{\max} - E_{\max,2}$  in the lower bound quantifies the regret caused by a sub-optimal action in the best-case CMP, in which the negative impact of the sub-optimal action

<sup>2</sup>Note that this assumption could be easily removed by partitioning the action space in  $s_t$  as  $\mathcal{A}(s_t) = \mathcal{A}_{\text{opt}}(s_t) \cup \mathcal{A}_{\text{sub-opt}}(s_t)$ .

is minimized. Note that, whenever the optimal decision in  $s_t$  does not depend on the history that lead to it ( $h_{s_t}$ ), the policy  $\pi_M$  can act deterministically ( $\pi_M(a_{NM}|s_t) = 1$ ) and the regret-to-go, its lower bound, and its upper bound are simultaneously zero (through  $\text{Var}_{h_{s_t} \sim p_T^{\pi_{NM}}}[\pi_{NM}(h_{s_t})] = 0$ ).

Finally, although the objective (2) is non-additive across time steps, we can still define a notion of *pseudo-instantaneous regret* by comparing the regret-to-go of two subsequent time steps. In the following, we provide the definition of this expected pseudo-instantaneous regret along with lower and upper bounds.

**Definition 4.5** (Expected Pseudo-Instantaneous Regret). *Consider a policy  $\pi \in \Pi$  interacting with a CMP over  $T$  steps. We define the expected pseudo-instantaneous regret of  $\pi$  at  $t$  as  $r_t(\pi) := \max(0, \mathcal{R}_{T-t}(\pi) - \mathcal{R}_{T-t-1}(\pi))$ .*

**Corollary 4.6.** *Let  $\pi_M \in \Pi_M$  be a Markovian policy such that  $\pi_M \in \arg \max_{\pi \in \Pi_M} \mathcal{E}(\pi)$  on a CMP  $\mathcal{M}$ . Then, for any  $t \in [T]$ , it holds  $\underline{r}_t(\pi_M) \leq r_t(\pi_M) \leq \bar{r}_t(\pi_M)$  such that*

$$\begin{aligned} \underline{r}_t(\pi_M) &= \max\left(0, E_{\max}(\mathcal{V}_t(\pi_M) - \mathcal{V}_{t+1}(\pi_M)) \right. \\ &\quad \left. - E_{\max,2}\mathcal{V}_t(\pi_M) + E_{\min,t+1}\mathcal{V}_{t+1}(\pi_M)\right), \\ \bar{r}_t(\pi_M) &= \max\left(0, E_{\max}(\mathcal{V}_t(\pi_M) - \mathcal{V}_{t+1}(\pi_M)) \right. \\ &\quad \left. - E_{\min,t}\mathcal{V}_t(\pi_M) + E_{\max,2}\mathcal{V}_{t+1}(\pi_M)\right), \end{aligned}$$

where

$$\mathcal{V}_t(\pi_M) := \frac{1}{\pi_M(a_{NM}|s_t)} \text{Var}_{h_{s_t} \sim p_T^{\pi_{NM}}}[\pi_{NM}(h_{s_t})].$$

## 5. Complexity Analysis

It is now worth considering how hard it is to optimize the objective (2) within the class of non-Markovian policies. Especially, we aim at characterizing the complexity of

$$\Psi_0 := \underset{\pi \in \Pi_{NM}}{\text{maximize}} \mathcal{E}(\pi),$$

defined over a CMP  $\mathcal{M}$ . First, we recall that  $\Psi_0$  can be rewritten as the problem of finding a reward-maximizing Markovian policy, i.e.,  $\tilde{\pi}_M \in \arg \max_{\pi \in \Pi_M} \mathcal{J}_{\tilde{\mathcal{M}}_T^R}(\pi)$ , over a convenient extended MDP  $\tilde{\mathcal{M}}_T^R$  obtained from CMP  $\mathcal{M}$  (see the proof of Lemma 4.2 for details). We call this problem  $\tilde{\Psi}_0$  and we note that  $\tilde{\Psi}_0 \in P$ , as the problem of finding a reward-maximizing Markovian policy is well-known to be in  $P$  for any MDP (Papadimitriou & Tsitsiklis, 1987). However, the following lemmas shows that it does not exist a many-to-one reduction from  $\Psi_0$  to  $\tilde{\Psi}_0$ , and that  $\Psi \notin NP$ .

**Lemma 5.1.** *A reduction  $\Psi_0 \leq_m \tilde{\Psi}_0$  does not exist.*

**Lemma 5.2.**  $\Psi_0 \notin NP$ .

As a direct consequence,  $\Psi_0 \notin NP$ -complete. We can now prove the main theorem of this section, which shows that  $\Psi_0$  is NP-hard under the common assumption that  $P \neq NP$ .

**Theorem 5.3.**  $\Psi_0 \in NP$ -hard.

*Proof Sketch.* To prove that  $\Psi_0 \in NP$ -hard, it is sufficient to show that there exists a problem  $\Psi_c \in NP$ -complete so that  $\Psi_c \leq_p \Psi_0$ . We show this by reducing 3SAT, which is a well-known NP-complete problem, to  $\Psi_0$ . To derive the reduction we consider two intermediate problems, namely  $\Psi_1$  and  $\Psi_2$ . Especially, we aim to show that the following chain of reductions holds

$$\Psi_0 \geq_m \Psi_1 \geq_p \Psi_2 \geq_p 3SAT.$$

First, we define  $\Psi_1$  and we prove that  $\Psi_0 \geq_m \Psi_1$ . Informally,  $\Psi_1$  is the problem of finding a reward-maximizing Markovian policy  $\pi_M \in \Pi_M$  w.r.t. the entropy objective (2) encoded through a reward function in a convenient POMDP  $\tilde{\mathcal{M}}_\Omega^R$ . We can build  $\tilde{\mathcal{M}}_\Omega^R$  from the CMP  $\mathcal{M}$  similarly as the extended MDP  $\tilde{\mathcal{M}}_T^R$  (see Section 2 and the proof of Lemma 4.2 for details), except that the agent only access the observation space  $\tilde{\Omega}$  instead of the extended state space  $\tilde{\mathcal{S}}$ . In particular, we define  $\tilde{\Omega} = \mathcal{S}$  (note that  $\mathcal{S}$  is the state space of the original CMP  $\mathcal{M}$ ), and  $\tilde{O}(\tilde{o}|\tilde{s}) = \tilde{s}_{-1}$ . Then, the reduction  $\Psi_0 \geq_m \Psi_1$  works as follows. We denote as  $\mathcal{I}_{\Psi_i}$  the set of possible instances of problem  $\Psi_i$ . We show that  $\Psi_0$  is harder than  $\Psi_1$  by defining the polynomial-time functions  $\psi$  and  $\phi$  such that any instance of  $\Psi_1$  can be rewritten through  $\psi$  as an instance of  $\Psi_0$ , and a solution  $\pi_{NM}^* \in \Pi_{NM}$  for  $\Psi_0$  can be converted through  $\phi$  into a solution  $\pi_M^* \in \Pi_M$  for the original instance of  $\Psi_1$ . The function  $\psi$  sets  $\mathcal{S} = \tilde{\Omega}$  and derives the transition model of  $\mathcal{M}$  from the one of  $\tilde{\mathcal{M}}_\Omega^R$ , while  $\phi$  converts the optimal solution of  $\Psi_0$  by computing  $\pi_M^* = \sum_{ho \in \mathcal{H}_o} p^{\pi_{NM}^*}(ho) \pi_{NM}^*(a|ho)$ , where  $\mathcal{H}_o$  stands for the set of histories  $h \in \mathcal{H}_{[T]}$  ending in the observation  $o \in \Omega$ . Thus, we have that  $\Psi_0 \geq_m \Psi_1$  holds. We now define  $\Psi_2$  as the policy existence problem w.r.t. the problem statement of  $\Psi_1$ . Hence,  $\Psi_2$  is the problem of determining whether the value of a reward-maximizing Markovian policy  $\pi_M^* \in \arg \max_{\pi \in \Pi_M} \mathcal{J}_{\tilde{\mathcal{M}}_\Omega^R}(\pi)$  is greater than 0. Since computing an optimal policy in POMDPs is in general harder than the relative policy existence problem (Lusena et al., 2001, Section 3), we have that  $\Psi_1 \geq_p \Psi_2$ . For the last reduction, i.e.,  $\Psi_2 \geq_p 3SAT$ , we extend the proof of Theorem 4.13 in (Mundhenk et al., 2000), which states that the policy existence problem for POMDPs is NP-complete. In particular, we show that this holds within the restricted class of POMDPs defined in  $\Psi_1$ . Since the chain  $\Psi_0 \geq_m \Psi_1 \geq_p \Psi_2 \geq_p 3SAT$  holds, we have that  $\Psi_0 \geq_p 3SAT$ . Moreover, since  $3SAT \in NP$ -complete and  $\Psi_0 \notin NP$  (thanks to Lemma 5.2), we conclude that  $\Psi_0 \in NP$ -hard.  $\square$

## References

- Bertsekas, D. P. and Tsitsiklis, J. N. *Introduction to probability*. Athena Scientific Belmont, MA, 2002.
- Deisenroth, M. P., Neumann, G., Peters, J., et al. A survey on policy search for robotics. *Foundations and trends in Robotics*, 2013.
- Guo, Z. D., Azar, M. G., Saade, A., Thakoor, S., Piot, B., Pires, B. A., Valko, M., Mesnard, T., Lattimore, T., and Munos, R. Geometric entropic exploration. *arXiv preprint arXiv:2101.02055*, 2021.
- Hallak, A., Di Castro, D., and Mannor, S. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.
- Hazan, E., Kakade, S., Singh, K., and Van Soest, A. Provably efficient maximum entropy exploration. In *Proceedings of the International Conference on Machine Learning*, 2019.
- Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 1997.
- Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2020.
- Kocsis, L. and Szepesvári, C. Bandit based monte-carlo planning. In *European conference on machine learning*, 2006.
- Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. RL for latent mdps: Regret guarantees and a lower bound. *arXiv preprint arXiv:2102.04939*, 2021.
- Lee, L., Eysenbach, B., Parisotto, E., Xing, E., Levine, S., and Salakhutdinov, R. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.
- Liu, H. and Abbeel, P. Aps: Active pretraining with successor features. In *Proceedings of the International Conference on Machine Learning*, 2021a.
- Liu, H. and Abbeel, P. Behavior from the void: Unsupervised active pre-training. *arXiv preprint arXiv:2103.04551*, 2021b.
- Lusena, C., Goldsmith, J., and Mundhenk, M. Nonapproximability results for partially observable markov decision processes. *J. Artif. Int. Res.*, 2001.
- Mundhenk, M., Goldsmith, J., Lusena, C., and Allender, E. Complexity of finite-horizon markov decision process problems. *Journal of the ACM (JACM)*, 2000.
- Mutti, M. and Restelli, M. An intrinsically-motivated approach for learning highly exploring and fast mixing policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- Mutti, M., Mancassola, M., and Restelli, M. Learning to explore a class of multiple reward-free environments. In *Self-Supervision for Reinforcement Learning Workshop - ICLR 2021*, 2021a. URL <https://openreview.net/forum?id=xFcniDGBYKH>.
- Mutti, M., Pratissoli, L., and Restelli, M. Task-agnostic exploration via policy gradient of a non-parametric state entropy estimate. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021b.
- Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of markov decision processes. *Mathematics of operations research*, 1987.
- Peters, J. and Schaal, S. Reinforcement learning of motor skills with policy gradients. *Neural networks*, 2008.
- Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- Seo, Y., Chen, L., Shin, J., Lee, H., Abbeel, P., and Lee, K. State entropy maximization with random encoders for efficient exploration. In *Proceedings of the International Conference on Machine Learning*, 2021.
- Strehl, A. L. and Littman, M. L. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 2008.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.
- Tarbouriech, J. and Lazaric, A. Active exploration in markov decision processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2019.
- Watkins, C. J. and Dayan, P. Q-learning. *Machine learning*, 1992.
- Yarats, D., Fergus, R., Lazaric, A., and Pinto, L. Reinforcement learning with prototypical representations. In *Proceedings of the International Conference on Machine Learning*, 2021.
- Zhang, C., Cai, Y., Huang, L., and Li, J. Exploration by maximizing rényi entropy for zero-shot meta rl. *arXiv preprint arXiv:2006.06193*, 2020a.
- Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, 2020b.

## A. Missing Proofs

### A.1. Proofs of Section 3

**Theorem 3.1.** *Let  $x \in \{\infty, \gamma, T\}$ , and let  $\mathcal{D}_{\text{NM}}^x = \{d_x^\pi(\cdot) : \pi \in \Pi_{\text{NM}}\}$ ,  $\mathcal{D}_{\text{NS}}^x = \{d_x^\pi(\cdot) : \pi \in \Pi_{\text{NS}}\}$ ,  $\mathcal{D}_{\text{M}}^x = \{d_x^\pi(\cdot) : \pi \in \Pi_{\text{M}}\}$  the corresponding sets of distributions. We can prove that (i) the sets of stationary state distributions are equivalent  $\mathcal{D}_{\text{NM}}^\infty \equiv \mathcal{D}_{\text{NS}}^\infty \equiv \mathcal{D}_{\text{M}}^\infty$ , (ii) the sets of discounted state distributions are equivalent  $\mathcal{D}_{\text{NM}}^\gamma \equiv \mathcal{D}_{\text{NS}}^\gamma \equiv \mathcal{D}_{\text{M}}^\gamma$  for any  $\gamma$ , (iii) the sets of marginal state distributions are equivalent  $\mathcal{D}_{\text{NM}}^T \equiv \mathcal{D}_{\text{NS}}^T \equiv \mathcal{D}_{\text{M}}^T$  for any  $T$ .*

*Proof.* First, note that a non-Markovian policy  $\pi \in \Pi_{\text{NM}}$  can always reduce to a non-stationary policy  $\pi \in \Pi_{\text{NS}}$  by conditioning the decision rules on the history length, or to a Markovian policy  $\pi \in \Pi_{\text{M}}$  by conditioning the decision rules on the last history entry. Thus,  $\mathcal{D}_{\text{NM}}^x \supseteq \mathcal{D}_{\text{NS}}^x \supseteq \mathcal{D}_{\text{M}}^x$  is straightforward for any  $x \in \{\infty, \gamma, T\}$ . From the derivations in (Puterman, 2014, Theorem 5.5.1), we have that  $\mathcal{D}_{\text{NS}}^x \supseteq \mathcal{D}_{\text{NM}}^x$  as well. Indeed, for any non-Markovian policy  $\pi \in \Pi_{\text{NM}}$ , we can build a non-stationary policy  $\pi' \in \Pi_{\text{NS}}$  as

$$\pi' = (\pi'_1, \pi'_2, \dots, \pi'_t, \dots), \quad \text{such that } \pi'_t(a|s) = \frac{d_t^\pi(s, a)}{d_t^\pi(s)}, \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}.$$

For  $t = 0$ , we have that  $d_0^\pi(\cdot) = d_0^{\pi'}(\cdot) = \mu(\cdot)$ , which is the initial state distribution. We proceed by induction to show that if  $d_{t-1}^\pi(\cdot) = d_{t-1}^{\pi'}(\cdot)$ , then we have

$$\begin{aligned} d_t^{\pi'}(s) &= \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} d_{t-1}^{\pi'}(s') \pi'_{t-1}(a|s') P(s|s', a) \\ &= \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \frac{d_{t-1}^{\pi'}(s')}{d_{t-1}^\pi(s')} d_{t-1}^\pi(s', a) P(s|s', a) \\ &= \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} d_{t-1}^\pi(s', a) P(s|s', a) = d_t^\pi(s). \end{aligned}$$

Since  $d_t^\pi(s) = d_t^{\pi'}(s)$  holds for any  $t \geq 0$  and  $\forall s \in \mathcal{S}$ , we have  $d_\infty^\pi(\cdot) = d_\infty^{\pi'}(\cdot)$ ,  $d_\gamma^\pi(\cdot) = d_\gamma^{\pi'}(\cdot)$ ,  $d_T^\pi(\cdot) = d_T^{\pi'}(\cdot)$ , and thus  $\mathcal{D}_{\text{NS}}^x \supseteq \mathcal{D}_{\text{NM}}^x$ . Then,  $\mathcal{D}_{\text{NM}}^x \equiv \mathcal{D}_{\text{NS}}^x$  follows.

(i)  $\mathcal{D}_{\text{NM}}^\infty \equiv \mathcal{D}_{\text{NS}}^\infty \equiv \mathcal{D}_{\text{M}}^\infty$

This result can be obtained with the same construction as before. Let us consider a non-Markovian policy  $\pi \in \Pi_{\text{NM}}$  having mixing time  $t_{\text{mix}} := \{t \in \mathbb{N} : \sup_{s \in \mathcal{S}} |d_\infty^\pi(s) - d_t^\pi(s)| \leq \epsilon\}$  for some mixing threshold  $\epsilon$ . Then, we can define a Markovian policy  $\pi_{\text{M}} \in \Pi_{\text{M}}$  as  $\pi_{\text{M}}(a|s) = d_{t_{\text{mix}}}^\pi(s, a) / d_{t_{\text{mix}}}^\pi(s)$ ,  $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$ . If we construct the same non-stationary policy  $\pi'_t = (\pi'_1, \pi'_2, \dots, \pi'_t, \dots) \in \Pi_{\text{NS}}$  as before, i.e.,  $\pi'_t(a|s) = d_t^\pi(s, a) / d_t^\pi(s)$ ,  $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$ , we have that  $d_t^{\pi'}(\cdot) = d_t^\pi(\cdot)$  for any  $t \geq 0$ , and thus also for  $t \geq t_{\text{mix}}$ . This implies that  $\pi, \pi', \pi_{\text{M}}$  all converges to the same stationary state distribution  $d_\infty^\pi(\cdot) = \lim_{t \rightarrow \infty} d_t^\pi(\cdot)$  if we take the limit  $\epsilon \rightarrow 0$  on the mixing threshold. Then, we have  $\mathcal{D}_{\text{NM}}^\infty \supseteq \mathcal{D}_{\text{NS}}^\infty \supseteq \mathcal{D}_{\text{M}}^\infty$  and  $\mathcal{D}_{\text{NM}}^\infty \equiv \mathcal{D}_{\text{NS}}^\infty \equiv \mathcal{D}_{\text{M}}^\infty$ .

(ii)  $\mathcal{D}_{\text{NM}}^\gamma \equiv \mathcal{D}_{\text{NS}}^\gamma \equiv \mathcal{D}_{\text{M}}^\gamma$

We can prove the statement by showing that  $\mathcal{D}_{\text{M}}^\gamma \supseteq \mathcal{D}_{\text{NS}}^\gamma$ . To this purpose, let us define the  $T$ -bounded discounted state distribution as  $d_{\gamma, T}^\pi(\cdot) := \frac{(1-\gamma)}{(1-\gamma^T)} \sum_{t \in [T]} \gamma^t d_t^\pi(\cdot)$ , such that  $\lim_{T \rightarrow \infty} d_{\gamma, T}^\pi(\cdot) = d_\gamma^\pi(\cdot)$ . Then, We consider a general non-stationary policy  $\pi = (\pi_0, \pi_1, \dots, \pi_t, \dots) \in \Pi_{\text{NS}}$ . We can build a Markovian policy  $\pi' \in \Pi_{\text{M}}$  as follows

$$\pi'(a|s) = \frac{\sum_{t=0}^{T-2} \gamma^t d_t^\pi(s) \pi_t(a|s)}{\sum_{t=0}^{T-2} \gamma^t d_t^\pi(s)}, \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}.$$

We prove by induction that such a policy  $\pi'$  induces the same  $T$ -bounded discounted state distribution of  $\pi$  for every  $T > 0$ .

For  $T = 1$  we have  $d_{\gamma,T}^{\pi'}(\cdot) = d_{\gamma,T}^{\pi}(\cdot) = \mu(\cdot)$ . Then, if  $d_{\gamma,T-1}^{\pi'}(\cdot) = d_{\gamma,T-1}^{\pi}(\cdot)$ , for every  $s \in \mathcal{S}$  it follows

$$\begin{aligned}
 d_{\gamma,T}^{\pi'}(s) &= \frac{(1-\gamma)}{(1-\gamma^T)} \sum_{t=0}^{T-1} \gamma^t d_t^{\pi'}(s) \\
 &= \frac{(1-\gamma)}{(1-\gamma^T)} \sum_{t=0}^{T-2} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \gamma^t d_t^{\pi'}(s') \pi'(a'|s') P(s|s', a') \\
 &= \frac{(1-\gamma)}{(1-\gamma^T)} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \frac{\sum_{t=0}^{T-2} \gamma^t d_t^{\pi'}(s')}{\sum_{t=0}^{T-2} \gamma^t d_t^{\pi}(s')} \sum_{t=0}^{T-2} \gamma^t d_t^{\pi}(s') \pi_t(a'|s') P(s|s', a') \\
 &= \frac{(1-\gamma)}{(1-\gamma^T)} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \frac{d_{\gamma,T-1}^{\pi'}(s')}{d_{\gamma,T-1}^{\pi}(s')} \sum_{t=0}^{T-2} \gamma^t d_t^{\pi}(s') \pi_t(a'|s') P(s|s', a') \\
 &= \frac{(1-\gamma)}{(1-\gamma^T)} \sum_{t=0}^{T-2} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \gamma^t d_t^{\pi}(s') \pi_t(a'|s') P(s|s', a') \\
 &= \frac{(1-\gamma)}{(1-\gamma^T)} \sum_{t=0}^{T-1} \gamma^t d_t^{\pi}(s) = d_{\gamma,T}^{\pi}(s).
 \end{aligned}$$

Since the relation holds for any  $T > 0$ , we can take the limit for  $T \rightarrow \infty$  to have  $d_{\gamma}^{\pi'}(\cdot) = d_{\gamma}^{\pi}(\cdot)$ , which gives  $\mathcal{D}_M^{\gamma} \supseteq \mathcal{D}_{NS}^{\gamma}$  and then  $\mathcal{D}_{NM}^{\gamma} \equiv \mathcal{D}_{NS}^{\gamma} \equiv \mathcal{D}_M^{\gamma}$ .

$$(iii) \mathcal{D}_{NM}^T \equiv \mathcal{D}_{NS}^T \equiv \mathcal{D}_M^T$$

We can prove the statement by showing that  $\mathcal{D}_M^T \supseteq \mathcal{D}_{NS}^T$ . We consider a general non-stationary policy  $\pi = (\pi_0, \pi_1, \dots, \pi_t, \dots) \in \Pi_{NS}$ . We can build a Markovian policy  $\pi' \in \Pi_M$  which marginalizes  $\pi$  over the time steps, such that

$$\pi'(a|s) = \frac{\sum_{t=0}^{T-2} d_t^{\pi}(s) \pi_t(a|s)}{\sum_{t=0}^{T-2} d_t^{\pi}(s)}, \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}.$$

We prove the statement by induction. For  $T = 1$  we have  $d_T^{\pi'}(\cdot) = d_T^{\pi}(\cdot) = \mu(\cdot)$ . Then, we can show that if  $d_{T-1}^{\pi'}(\cdot) = d_{T-1}^{\pi}(\cdot)$ , it follows  $d_T^{\pi'}(\cdot) = d_T^{\pi}(\cdot)$ . Especially,

$$\begin{aligned}
 d_T^{\pi'}(s) &= \frac{1}{T} \sum_{t=0}^{T-1} d_t^{\pi'}(s) = \frac{1}{T} \sum_{t=0}^{T-2} \sum_{s' \in \mathcal{S}} d_t^{\pi'}(s') \sum_{a' \in \mathcal{A}} \pi'(a'|s') P(s|s', a') \\
 &= \frac{1}{T} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \frac{\sum_{t=0}^{T-2} d_t^{\pi'}(s')}{\sum_{t=0}^{T-2} d_t^{\pi}(s')} \sum_{t=0}^{T-2} d_t^{\pi}(s') \pi_t(a'|s') P(s|s', a') \\
 &= \frac{1}{T} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \frac{d_{T-1}^{\pi'}(s')}{d_{T-1}^{\pi}(s')} \sum_{t=0}^{T-2} d_t^{\pi}(s') \pi_t(a'|s') P(s|s', a') \\
 &= \frac{1}{T} \sum_{t=0}^{T-2} \sum_{s' \in \mathcal{S}} d_t^{\pi}(s') \sum_{a' \in \mathcal{A}} \pi_t(a'|s') P(s|s', a') = \frac{1}{T} \sum_{t=0}^{T-1} d_t^{\pi}(s) = d_T^{\pi}(s),
 \end{aligned}$$

holds for any  $s \in \mathcal{S}$ , thus  $\mathcal{D}_M^T \supseteq \mathcal{D}_{NS}^T$  and  $\mathcal{D}_{NM}^T \equiv \mathcal{D}_{NS}^T \equiv \mathcal{D}_M^T$  follows.  $\square$

**Corollary 3.2.** For every CMP  $\mathcal{M}$ , there exists a Markovian policy  $\pi^* \in \Pi_M$  such that  $\pi^* \in \arg \max_{\pi \in \Pi} \mathcal{E}_{\infty}(\pi)$ .

*Proof.* The result is straightforward from Theorem 3.1 and noting that the set of non-Markovian policies  $\Pi_{NM}$  with arbitrary history-length is as powerful as the general set of policies  $\Pi$ . Thus, for every policy  $\pi \in \Pi$  there exists a (possibly randomized) policy  $\pi' \in \Pi_M$  inducing the same (stationary, discounted or marginal) state distribution of  $\pi$ , i.e.,  $d^{\pi}(\cdot) = d^{\pi'}(\cdot)$ , which implies  $E(d^{\pi}(\cdot)) = E(d^{\pi'}(\cdot))$ . If it holds for any  $\pi \in \Pi$ , then it holds for  $\pi^* \in \arg \max_{\pi \in \Pi} E(d^{\pi}(\cdot))$ .  $\square$



## A.2. Proofs of Section 4

**Lemma 4.3.** Let  $\pi_{\text{NM}} \in \Pi_{\text{NM}}^{\text{D}}$  be a non-Markovian policy such that  $\pi_{\text{NM}} \in \arg \max_{\pi \in \Pi} \mathcal{E}(\pi)$  on a CMP  $\mathcal{M}$ . The variance of an optimal Markovian policy  $\pi_{\text{M}} \in \arg \max_{\pi \in \Pi_{\text{M}}} \mathcal{E}(\pi)$  is given by

$$\mathbb{V}\text{ar}_{a \sim \pi_{\text{M}}(s)}[a] = \mathbb{V}\text{ar}_{hs \sim p_T^{\pi_{\text{NM}}}}[\pi_{\text{NM}}(hs)], \quad \forall s \in \mathcal{S},$$

where  $hs \in \mathcal{H}_{[T]}$  is any history with final state  $s$ .

*Proof.* Let us derive the variance of a policy  $\pi \in \Pi$  in state  $s \in \mathcal{S}$  through the law of total variance (Bertsekas & Tsitsiklis, 2002). We have

$$\begin{aligned} \mathbb{V}\text{ar}_{\pi}[a|s] &= \mathbb{E}_{\pi}[a^2|s] - \mathbb{E}_{\pi}[a|s]^2 \\ &= \mathbb{E}_h \left[ \mathbb{E}_{\pi}[a^2|s, h] \right] - \mathbb{E}_h \left[ \mathbb{E}_{\pi}[a|s, h] \right]^2 \\ &= \mathbb{E}_h \left[ \mathbb{V}\text{ar}_{\pi}[a|s, h] + \mathbb{E}_{\pi}[a|s, h]^2 \right] - \mathbb{E}_h \left[ \mathbb{E}_{\pi}[a|s, h] \right]^2 \\ &= \mathbb{E}_h \left[ \mathbb{V}\text{ar}_{\pi}[a|s, h] \right] + \mathbb{E}_h \left[ \mathbb{E}_{\pi}[a|s, h]^2 \right] - \mathbb{E}_h \left[ \mathbb{E}_{\pi}[a|s, h] \right]^2 \\ &= \mathbb{E}_h \left[ \mathbb{V}\text{ar}_{\pi}[a|s, h] \right] + \mathbb{V}\text{ar}_h \left[ \mathbb{E}_{\pi}[a|s, h] \right]. \end{aligned}$$

Let  $h \sim p_T^{\pi}$ , so that the variable  $a|s, h$  becomes  $a|hs$  where  $hs = (s_{0,hs}, a_{0,hs}, s_{1,hs}, \dots, s_{t,hs} = s) \in \mathcal{H}_{[T]}$ , to obtain

$$\mathbb{V}\text{ar}_{\pi}[a|s] = \mathbb{E}_{hs \sim p_T^{\pi}} \left[ \mathbb{V}\text{ar}_{\pi}[a|hs] \right] + \mathbb{V}\text{ar}_{hs \sim p_T^{\pi}} \left[ \mathbb{E}_{\pi}[a|hs] \right]. \quad (3)$$

If we take the policy  $\pi$  that optimizes the objective (2), we have that the action  $a$  on the left-hand side of equation (3) is distributed according to  $\pi_{\text{M}} \arg \max_{\pi \in \Pi_{\text{M}}} \mathcal{E}(\pi)$ , while on the right-hand side is distributed according to  $\pi_{\text{NM}} \arg \max_{\pi \in \Pi_{\text{NM}}} \mathcal{E}(\pi)$ , i.e.,

$$\mathbb{V}\text{ar}_{a \sim \pi_{\text{M}}(s)}[a] = \mathbb{E}_{hs \sim p_T^{\pi_{\text{NM}}}} \left[ \mathbb{V}\text{ar}_{a \sim \pi_{\text{NM}}(hs)}[a] \right] + \mathbb{V}\text{ar}_{hs \sim p_T^{\pi_{\text{NM}}}} \left[ \mathbb{E}_{a \sim \pi_{\text{NM}}(hs)}[a] \right]. \quad (4)$$

From Lemma 4.2, we know that there exists a deterministic optimal non-Markovian policy  $\pi_{\text{NM}} \in \Pi_{\text{NM}}^{\text{D}}$ , which gives  $\mathbb{E}_{hs \sim p_T^{\pi_{\text{NM}}}} [\mathbb{V}\text{ar}_{a \sim \pi_{\text{NM}}(hs)}[a]] = 0$  in (4) and concludes the proof.  $\square$

**Lemma A.1 (Best-Case CMP).** Let  $h_t^*$  be a zero-regret trajectory of  $t$ -steps. Taking a sub-optimal action  $a_t \in \mathcal{A} \setminus \pi_{\text{NM}}(h_t^*)$  at step  $t$  in the best-case CMP  $\mathcal{M}$  gives a final entropy

$$E_{\text{max},2} = \max_{h \in \mathcal{H}_{T-t} \setminus \mathcal{H}_{T-t}^*} \text{Entropy}(d_{(h_t^*, h)}(\cdot)) \quad \text{s.t. } \mathcal{H}_{T-t}^* = \arg \max_{h \in \mathcal{H}_{T-t}} \text{Entropy}(d_{(h_t^*, h)}(\cdot))$$

where the maximum is attained by

$$\bar{h}_{T-t} = (s_{\text{max},2}, h_{T-t-1}^* \in \mathcal{H}_{T-t}^*) \in \arg \max_{h \in \mathcal{H}_{T-t} \setminus \mathcal{H}_{T-t}^*} \text{Entropy}(d_{(h_t^*, h)}(\cdot)),$$

in which  $s_{\text{max},2}$  is any state that is the second-closest to a uniform entry in  $d_{(h_t^*, \bar{h}_{T-t})}$ .

*Proof.* The best-case CMP  $\bar{\mathcal{M}}$  is designed such that taking a sub-optimal action  $a_t \in \mathcal{A} \setminus \pi_{\text{NM}}(h_t^*)$  minimally decrease the final entropy. Especially, instead of reaching at step  $t+1$  an optimal state  $s_{\text{max}}$ , i.e., a state that maximally balances the state visits of the final trajectory, the agent is drawn to the second-to-optimal state  $s_{\text{max},2}$ , from which it gets back on track on the optimal trajectory for the remaining steps. Note that visiting  $s_{\text{max},2}$  cannot lead to the optimal final entropy, achieved when  $s_{\text{max}}$  is visited at step  $t+1$ , due to the sub-optimality of action  $a_t$ .  $\square$

**Lemma A.2** (Worst-Case CMP). *Let  $h_t^*$  be a zero-regret trajectory of  $t$ -steps. Taking a sub-optimal action  $a_t \in \mathcal{A} \setminus \pi_{\text{NM}}(h_t^*)$  at step  $t$  in the worst-case CMP  $\underline{\mathcal{M}}$  gives a final entropy*

$$E_{\min,t} = \min_{h \in \mathcal{H}_{T-t}} \text{Entropy}(d_{h_t^*,h}(\cdot)),$$

where the minimum is attained by

$$\underline{h}_{T-t} = \left( s_i \in \arg \max_{s \in \mathcal{S}} d_{h_t^*}(s) \right)_{i=t+1}^T \in \arg \min_{h \in \mathcal{H}_{T-t}} \text{Entropy}(d_{h_t^*,h}(\cdot)).$$

*Proof.* The worst-case CMP  $\underline{\mathcal{M}}$  is designed such that the agent cannot recover from a sub-optimal action  $a_t \in \mathcal{A} \setminus \pi_{\text{NM}}(h_t^*)$  as it is absorbed by a worst-case state given the trajectory  $h_t^*$ . A worst-case state is one that maximizes the visitation frequency in  $h_t^*$ , i.e.,  $\underline{s} \in \arg \max_{s \in \mathcal{S}} d_{h_t^*}(s)$ , so that the visitation frequency becomes increasingly unbalanced. A sub-optimal action at the first step in  $\underline{\mathcal{M}}$  leads to  $T - 1$  visits to the initial state  $s_0 \sim \mu$ , and the final entropy is zero.  $\square$

**Theorem 4.4.** *Let  $\pi_{\text{M}} \in \Pi_{\text{M}}$  be a Markovian policy  $\pi_{\text{M}} \in \arg \max_{\pi \in \Pi_{\text{M}}} \mathcal{E}(\pi)$  on a CMP  $\mathcal{M}$ . Then, for any  $t \in [T]$ , it holds  $\underline{\mathcal{R}}_{T-t}(\pi_{\text{M}}) \leq \mathcal{R}_{T-t}(\pi_{\text{M}}) \leq \overline{\mathcal{R}}_{T-t}(\pi_{\text{M}})$  such that*

$$\begin{aligned} \underline{\mathcal{R}}_{T-t}(\pi_{\text{M}}) &= \frac{E_{\max} - E_{\max,2}}{\pi_{\text{M}}(a_{\text{NM}}|s_t)} \mathbb{V}_{h s_t \sim p_T^{\pi_{\text{NM}}}} [\pi_{\text{NM}}(h s_t)], \\ \overline{\mathcal{R}}_{T-t}(\pi_{\text{M}}) &= \frac{E_{\max} - E_{\min,t}}{\pi_{\text{M}}(a_{\text{NM}}|s_t)} \mathbb{V}_{h s_t \sim p_T^{\pi_{\text{NM}}}} [\pi_{\text{NM}}(h s_t)], \end{aligned}$$

where  $\pi_{\text{NM}} \in \arg \max_{\pi \in \Pi_{\text{NM}}^{\text{D}}} \mathcal{E}(\pi)$ ,  $a_{\text{NM}} = \pi_{\text{NM}}(h_t^*)$ , and

$$\begin{aligned} E_{\max} &= \max_{\pi^* \in \Pi} \mathbb{E}_{h_{T-t}^* \sim p_T^{\pi^*}} [E(d_{h_t^*}(\cdot))], \\ E_{\min,t} &= \min_{h \in \mathcal{H}_{T-t}} E(d_{(h_t^*,h)}(\cdot)), \\ E_{\max,2} &= \max_{h \in \mathcal{H}_{T-t} \setminus \mathcal{H}_{T-t}^*} E(d_{(h_t^*,h)}(\cdot)) \\ \text{s.t. } \mathcal{H}_{T-t}^* &= \arg \max_{h \in \mathcal{H}_{T-t}} E(d_{(h_t^*,h)}(\cdot)). \end{aligned}$$

*Proof.* From the definition of the regret-to-go (Definition 4.1), we have that

$$\mathcal{R}_{T-t}(\pi_{\text{M}}) = \max_{\pi^* \in \Pi} \mathbb{E}_{h_{T-t}^* \sim p_T^{\pi^*}} [\text{Entropy}(d_{h_t^*}(\cdot))] - \mathbb{E}_{h_{T-t} \sim p_T^{\pi_{\text{M}}}} [\text{Entropy}(d_{h_t}(\cdot))],$$

in which we substitute  $E_{\max} = \max_{\pi^* \in \Pi} \mathbb{E}_{h_{T-t}^* \sim p_T^{\pi^*}} [\text{Entropy}(d_{h_t^*}(\cdot))]$ . To derive a lower bound and an upper bound to  $\mathcal{R}_{T-t}(\pi)$  we consider the impact that taking a sub-optimal action  $a \in \mathcal{A} \setminus \{a_{\text{NM}}\}$  in state  $s_t$  would have in a best-case and a worst-case CMP respectively, which is detailed in Lemma A.1 and Lemma A.2. Especially, we can write

$$\begin{aligned} \mathcal{R}_{T-t}(\pi_{\text{M}}) &= E_{\max} - \mathbb{E}_{h_{T-t} \sim p_T^{\pi_{\text{M}}}} [\text{Entropy}(d_{h_t}(\cdot))] \\ &\geq E_{\max} - \pi_{\text{M}}(a_{\text{NM}}|s_t) E_{\max} - (1 - \pi_{\text{M}}(a_{\text{NM}}|s_t)) E_{\max,2} \\ &= (E_{\max} - E_{\max,2})(1 - \pi_{\text{NM}}(a_{\text{NM}}|s_t)) \end{aligned}$$

and

$$\begin{aligned} \mathcal{R}_{T-t}(\pi_{\text{M}}) &= E_{\max} - \mathbb{E}_{h_{T-t} \sim p_T^{\pi_{\text{M}}}} [\text{Entropy}(d_{h_t}(\cdot))] \\ &\leq E_{\max} - \pi_{\text{M}}(a_{\text{NM}}|s_t) E_{\max} - (1 - \pi_{\text{M}}(a_{\text{NM}}|s_t)) E_{\min,t} \\ &= (E_{\max} - E_{\min,t})(1 - \pi_{\text{NM}}(a_{\text{NM}}|s_t)). \end{aligned}$$

Then, we note that the event of taking a sub-optimal action  $a \in \mathcal{A} \setminus \{a_{\text{NM}}\}$  with policy  $\pi_{\text{M}}$  can be modelled by a Bernoulli distribution with parameter  $(1 - \pi_{\text{NM}}(a_{\text{NM}}|s_t))$ . By combining the equation of the variance of a Bernoulli distribution and Lemma 4.3 we obtain

$$\mathbb{V}\text{ar}_{a \sim \pi_{\text{NM}}(s_t)}[a] = \pi_{\text{M}}(a_{\text{NM}}|s_t)(1 - \pi_{\text{M}}(a_{\text{NM}}|s_t)) = \mathbb{V}\text{ar}_{hs_t \sim p_T^{\pi_{\text{NM}}}}[\pi_{\text{NM}}(hs_t)],$$

which gives

$$\begin{aligned} \mathcal{R}_{T-t}(\pi_{\text{M}}) &\geq \frac{(E_{\text{max}} - E_{\text{max},2})}{\pi_{\text{M}}(a_{\text{NM}}|s_t)} \mathbb{V}\text{ar}_{hs_t \sim p^{\pi_{\text{NM}}}}[\pi_{\text{NM}}(hs_t)] := \underline{\mathcal{R}}_{T-t}(\pi_{\text{M}}) \\ \mathcal{R}_{T-t}(\pi_{\text{M}}) &\leq \frac{(E_{\text{max}} - E_{\text{min},t})}{\pi_{\text{M}}(a_{\text{NM}}|s_t)} \mathbb{V}\text{ar}_{hs_t \sim p^{\pi_{\text{NM}}}}[\pi_{\text{NM}}(hs_t)] := \overline{\mathcal{R}}_{T-t}(\pi_{\text{M}}) \end{aligned}$$

□

**Corollary 4.6.** *Let  $\pi_{\text{M}} \in \Pi_{\text{M}}$  be a Markovian policy such that  $\pi_{\text{M}} \in \arg \max_{\pi \in \Pi_{\text{M}}} \mathcal{E}(\pi)$  on a CMP  $\mathcal{M}$ . Then, for any  $t \in [T]$ , it holds  $\underline{r}_t(\pi_{\text{M}}) \leq r_t(\pi_{\text{M}}) \leq \bar{r}_t(\pi_{\text{M}})$  such that*

$$\begin{aligned} \underline{r}_t(\pi_{\text{M}}) &= \max \left( 0, E_{\text{max}}(\mathcal{V}_t(\pi_{\text{M}}) - \mathcal{V}_{t+1}(\pi_{\text{M}})) \right. \\ &\quad \left. - E_{\text{max},2}\mathcal{V}_t(\pi_{\text{M}}) + E_{\text{min},t+1}\mathcal{V}_{t+1}(\pi_{\text{M}}) \right), \\ \bar{r}_t(\pi_{\text{M}}) &= \max \left( 0, E_{\text{max}}(\mathcal{V}_t(\pi_{\text{M}}) - \mathcal{V}_{t+1}(\pi_{\text{M}})) \right. \\ &\quad \left. - E_{\text{min},t}\mathcal{V}_t(\pi_{\text{M}}) + E_{\text{max},2}\mathcal{V}_{t+1}(\pi_{\text{M}}) \right), \end{aligned}$$

where

$$\mathcal{V}_t(\pi_{\text{M}}) := \frac{1}{\pi_{\text{M}}(a_{\text{NM}}|s_t)} \mathbb{V}\text{ar}_{hs_t \sim p_T^{\pi_{\text{NM}}}}[\pi_{\text{NM}}(hs_t)].$$

*Proof.* From Definition 4.5, we have that  $r_t(\pi_{\text{M}}) = \mathcal{R}_{T-t} - \mathcal{R}_{T-t-1}$ . Recall that

$$\underline{\mathcal{R}}_{T-t}(\pi) = \mathcal{V}_t(\pi)(E_{\text{max}} - E_{\text{max},2}), \quad \overline{\mathcal{R}}_{T-t}(\pi) = \mathcal{V}_t(\pi)(E_{\text{max}} - E_{\text{min},t}),$$

from Theorem 4.4. We can write

$$\begin{aligned} \underline{r}_t(\pi_{\text{M}}) &\geq \underline{\mathcal{R}}_{T-t} - \overline{\mathcal{R}}_{T-t-1} \\ &= E_{\text{max}}(\mathcal{V}_t(\pi_{\text{M}}) - \mathcal{V}_{t+1}(\pi_{\text{M}})) - E_{\text{max},2}\mathcal{V}_t(\pi_{\text{M}}) + E_{\text{min},t+1}\mathcal{V}_{t+1}(\pi_{\text{M}}), \end{aligned}$$

and

$$\begin{aligned} \bar{r}_t(\pi_{\text{M}}) &\leq \overline{\mathcal{R}}_{T-t} - \underline{\mathcal{R}}_{T-t-1} \\ &= E_{\text{max}}(\mathcal{V}_t(\pi_{\text{M}}) - \mathcal{V}_{t+1}(\pi_{\text{M}})) - E_{\text{min},t}\mathcal{V}_t(\pi_{\text{M}}) + E_{\text{max},2}\mathcal{V}_{t+1}(\pi_{\text{M}}). \end{aligned}$$

□

### A.3. Proofs of Section 5

**Lemma 5.1.** *A reduction  $\Psi_0 \leq_m \tilde{\Psi}_0$  does not exist.*

*Proof.* In the general case, coding any instance of  $\Psi_0$  in the representation required by  $\tilde{\Psi}_0$  holds exponential complexity w.r.t. the input of the initial instance of  $\Psi_0$ . □

**Lemma 5.2.**  $\Psi_0 \notin \text{NP}$ .

*Proof.* This proof is based on the verifier-based definition of NP. According to this definition, given any instance  $I \in \mathcal{I}_{\Psi_0}$  of problem  $\Psi_0$  and a candidate solution  $\pi \in \Pi_{\text{NM}}$ , if there exists an algorithm  $\Lambda$  that can verify the optimality of  $\pi$  in polynomial time, then  $\Psi_0 \in \text{NP}$ . We prove that  $\Psi_0 \notin \text{NP}$  by showing that a polynomial time verifier does not exist for any instance of the problem. It suffices to note that given a non-Markovian policy  $\pi \in \Pi_{\text{NM}}$ , i.e., a candidate solution of an instance of the problem  $\Psi_0$ , a polynomial verifier  $\Lambda$  does not exist, since any possible verifier has to compute the objective function for every possible well-defined non-Markovian policy in order to determine whether  $\pi$  is optimal according to  $\Psi_0$ , but this operation is exponential w.r.t. the input of  $\Psi_0$ .  $\square$

**Theorem 5.3.**  $\Psi_0 \in \text{NP-hard}$ .

*Proof.* To prove that  $\Psi_0 \in \text{NP-hard}$ , it is sufficient to show that there exists a problem  $\Psi_c \in \text{NP-complete}$  so that  $\Psi_c \leq_p \Psi_0$ . We show this by reducing 3SAT, a well-known NP-complete problem, to  $\Psi_0$ . To derive the reduction we consider two intermediate problems, namely  $\Psi_1$  and  $\Psi_2$ . Especially, we aim to show that the following chain of reductions hold:

$$\Psi_0 \geq_m \Psi_1 \geq_p \Psi_2 \geq_p \text{3SAT}$$

First, we define  $\Psi_1$  and we prove that  $\Psi_0 \geq_m \Psi_1$ . Informally,  $\Psi_1$  is the problem of finding a reward-maximizing Markovian policy  $\pi_M \in \Pi_M$  w.r.t. the entropy objective (2) encoded through a reward function in a convenient POMDP  $\widetilde{\mathcal{M}}_\Omega^R$ . We can build  $\widetilde{\mathcal{M}}_\Omega^R$  from the CMP  $\mathcal{M}$  similarly as the extended MDP  $\widetilde{\mathcal{M}}_T^R$  (see Section 2 and the proof of Lemma 4.2 for details), except that the agent only access the observation space  $\widetilde{\Omega}$  instead of the extended state space  $\widetilde{\mathcal{S}}$ . In particular, we define  $\widetilde{\Omega} = \mathcal{S}$  (note that  $\mathcal{S}$  is the state space of the original CMP  $\mathcal{M}$ ),  $\widetilde{O}(\widetilde{o}|\widetilde{s}) = \widetilde{s}_{-1}$ , and the reward function  $\widetilde{R}$  assigns value 0 to all states  $\widetilde{s} \in \widetilde{\mathcal{S}}$  such that  $|\widetilde{s}| \neq T$ , otherwise (if  $|\widetilde{s}| = T$ ) the reward corresponds to the entropy value of the state visitation frequencies induced by the trajectory codified through  $\widetilde{s}$ .

Then, the reduction  $\Psi_0 \geq_m \Psi_1$  works as follows. We denote as  $\mathcal{I}_{\Psi_i}$  the set of possible instances of problem  $\Psi_i$ . We show that  $\Psi_0$  is harder than  $\Psi_1$  by defining the polynomial-time functions  $\psi$  and  $\phi$  such that any instance of  $\Psi_1$  can be rewritten through  $\psi$  as an instance of  $\Psi_0$ , and a solution  $\pi_{\text{NM}}^* \in \Pi_{\text{NM}}$  for  $\Psi_0$  can be converted through  $\phi$  into a solution  $\pi_M^* \in \Pi_M$  for the original instance of  $\Psi_1$ .

$$\begin{array}{ccc} \mathcal{I}_{\Psi_1} & \xrightarrow{\psi} & \mathcal{I}_{\Psi_0} \\ & & \downarrow \\ \pi_M^* & \xleftarrow{\phi} & \pi_{\text{NM}}^* \end{array}$$

The function  $\psi$  sets  $\mathcal{S} = \widetilde{\Omega}$  and derives the transition model of  $\mathcal{M}$  from the one of  $\widetilde{\mathcal{M}}_\Omega^R$ , while  $\phi$  converts the optimal solution of  $\Psi_0$  by computing

$$\pi_M^* = \sum_{ho \in \mathcal{H}_o} p_T^{\pi_{\text{NM}}^*}(ho) \pi_{\text{NM}}^*(a|ho) \quad (5)$$

where  $\mathcal{H}_o$  stands for the set of histories  $h \in \mathcal{H}_{[T]}$  ending in the observation  $o \in \Omega$ . Thus, we have that  $\Psi_0 \geq_m \Psi_1$  holds. We now define  $\Psi_2$  as the policy existence problem w.r.t. the problem statement of  $\Psi_1$ . Hence,  $\Psi_2$  is the problem of determining whether the value of a reward-maximizing Markovian policy  $\pi_M^* \in \arg \max_{\pi \in \Pi_M} \mathcal{J}_{\widetilde{\mathcal{M}}_\Omega^R}(\pi)$  is greater than 0. Since computing an optimal policy in POMDPs is in general harder than the relative policy existence problem (Section 3) [Jusena2001], we have that  $\Psi_1 \geq_p \Psi_2$ .

For the last reduction, i.e.,  $\Psi_2 \geq_p \text{3SAT}$ , we extend the proof of Theorem 4.13 in (Mundhenk et al., 2000), which states that the policy existence problem for POMDPs is NP-complete. In particular, we show that this holds within the restricted class of POMDPs defined in  $\Psi_1$ .

The restrictions on the POMDPs class are the following:

1. The reward function  $R(s) \geq 0$  only in the subset of states reachable in T steps, otherwise  $R(s) = 0$
2.  $|\widetilde{\mathcal{S}}| = \widetilde{\mathcal{S}} = |\widetilde{\Omega}|^T$

Both limitations can be overcome in the following ways:

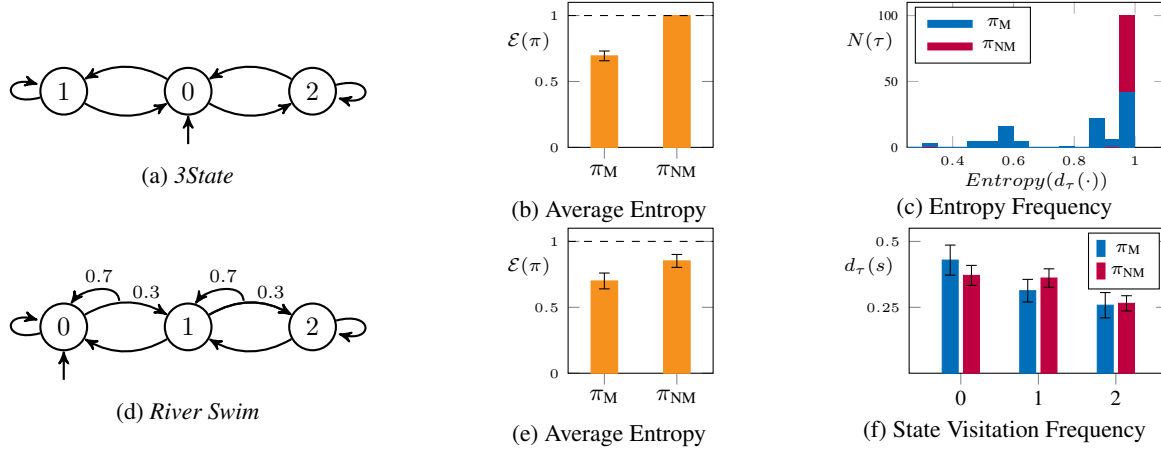


Figure 2: In (a, d), we illustrate the *3State* and *River Swim* CMPs. Then, we report the average entropy induced by an optimal Markovian policy  $\pi_M$  and an optimal non-Markovian policy  $\pi_{NM}$  in the *3State* ( $T = 9$ ) (b) and the *River Swim* ( $T = 10$ ) (e). In (c) we report the entropy frequency in the *3State*, in (f) the state visitation frequency in the *River Swim*. We provide 95% c.i. over 100 runs.

1. It suffices to add states with deterministic transitions so that  $T = m \cdot n$  can be defined a priori, where  $T$  is the number of steps needed to reach the state with positive reward through every possible path. Here  $m$  is the number of clauses, and  $n$  is the number of variables in the 3SAT instance, as defined in (Mundhenk et al., 2000).
2. The POMDPs class defined by  $\Psi_1$  is such that  $\tilde{S} = |\tilde{\Omega}|^T$ . Noticing that the set of observations corresponds with the set of variables and that from the previous point  $T = m \cdot n$ , we have that  $|\tilde{\Omega}|^T = n^{m \cdot n}$ , while the POMDPs class used by the proof hereinabove has  $\tilde{S} = m \cdot n^2$ . Notice that  $n \geq 2$  and  $m \geq 1$  implies that  $n^{m \cdot n} \geq m \cdot n^2$ . Moreover, notice that every instance of 3SAT has  $m \geq 1$  and  $n \geq 3$ . Hence, to extend the proof to the POMDPs class defined by  $\Psi_1$  it suffices to add a set of states  $\tilde{S}_p$  s.t.  $R(s) = 0 \forall s \in \tilde{S}_p$ .

Since the chain  $\Psi_0 \geq_m \Psi_1 \geq_p \Psi_2 \geq_p$  3SAT holds, we have that  $\Psi_0 \geq_p$  3SAT. Moreover, since 3SAT  $\in$  NP-complete and  $\Psi_0 \notin$  NP (thanks to Lemma 5.2), we conclude that  $\Psi_0 \in$  NP-hard.  $\square$

## B. Numerical Validation

Despite the hardness result of Theorem 5.3, we provide a brief numerical validation around the potential of non-Markovianity in MSE exploration. Crucially, the reported analysis is limited to simple domains and short time horizons, and it has to be intended as an illustration of the theoretical claims reported in previous sections. Whereas a comprehensive evaluation of the practical benefits of non-Markovianity in MSE exploration is left as future work, we discuss in Appendix C why we believe that the development of scalable methods is not hopeless even in this challenging setting.

In this section, we consider a *3State* ( $S = 3, A = 2, T = 9$ ), which is a simple abstraction of the two-rooms in Figure 1, and a *River Swim* (Strehl & Littman, 2008) ( $S = 3, A = 2, T = 10$ ) that are depicted in Figure 2a, 2d respectively. Especially, we compare the expected entropy (2) achieved by an optimal non-Markovian policy  $\pi_{NM} \in \arg \max_{\pi \in \Pi_{NM}} \mathcal{E}(\pi)$ , which is obtained by solving the extended MDP as described in the proof of Lemma 4.2, against an optimal Markovian policy  $\pi_M \in \arg \max_{\pi \in \Pi_M} \mathcal{E}(\pi)$ , which is obtained from  $\pi_{NM}$  through a marginalization over histories (as mentioned in the proof sketch of Theorem 3.1). In confirmation of the result in Theorem 4.4,  $\pi_M$  cannot match the performance of  $\pi_{NM}$  (see Figure 2b, 2e). In *3State*, an optimal strategy requires going left when arriving in state 0 from state 2 and vice versa. The policy  $\pi_{NM}$  is able to do that, and it always realizes the optimal trajectory (Figure 2c). Instead,  $\pi_M$  is uniform in 0 and it often runs into sub-optimal trajectories. In the *River Swim*, the main hurdle is to reach state 2 from the initial one. Whereas  $\pi_M$  and  $\pi_{NM}$  are equivalently good in doing so, as reported in Figure 2f, only the non-Markovian strategy is able to balance the visitations in the previous states when it eventually reaches 2. The difference is already noticeable with a short horizon and it would further increase with a longer  $T$ .

## C. Discussion

Throughout the paper, we advocated for the importance of non-Markovianity when optimizing a finite-sample MSE objective, but we also proved that the corresponding optimization problem is NP-hard in its general formulation. Despite the hardness result, we believe that it is not hopeless to learn exploration policies with some form of non-Markovianity, while still preserving an edge over Markovian strategies. In the following paragraphs, we discuss potential avenues to derive practical methods for relevant relaxations to the general class of non-Markovian policies.

**Finite-Length Histories** Throughout the paper, we considered non-Markovian policies that condition their decisions on histories of arbitrary length, i.e.,  $\pi : \mathcal{H} \rightarrow \Delta(\mathcal{A})$ . However, the complexity of optimizing such policies grows exponentially with the length of the history. To avoid this exponential blowup, one can define a class of non-Markovian policies  $\pi : \mathcal{H}_H \rightarrow \Delta(\mathcal{A})$  in which the decisions are conditioned on histories of a finite length  $H > 1$  that are obtained from a sliding window on the full history. The optimal policy within this class would still retain better regret guarantees than an optimal Markovian policy, but it would not achieve zero regret in general. With the length parameter  $H$  one can trade-off the learning complexity with the regret according to the structure of the domain. For instance,  $H = 2$  would be sufficient to achieve zero regret in the *3State* domain, whereas in the *River Swim* domain any  $H < T$  would cause some positive regret.

**Compact Representations of the History** Instead of setting a finite length  $H$ , one can choose to perform function approximation on the full history to obtain a class of policies  $\pi : f(\mathcal{H}) \rightarrow \Delta(\mathcal{A})$ , where  $f$  is a function that maps an history  $h$  to some compact representation. An interesting option is to use the notion of *eligibility traces* (Sutton & Barto, 2018) to encode the information of  $h$  in a vector of length  $S$ , which is updated as  $\mathbf{z}_{t+1} \leftarrow \lambda \mathbf{z}_t + \mathbf{1}_{s_t}$ , where  $\lambda \in (0, 1)$  is a discount factor,  $\mathbf{1}_{s_t}$  is a vector with a unit entry at the index  $s_t$ , and  $\mathbf{z}_0 = 0$ . The discount factor  $\lambda$  acts as a smoothed version of the length parameter  $H$ , and it can be dynamically adapted while learning. Indeed, this eligibility traces representation is particularly convenient for policy optimization (Deisenroth et al., 2013), in which we could optimize in turn a parametric policy over actions  $\pi_\theta(\cdot | \mathbf{z}, \lambda)$  and a parametric policy over the discount  $\pi_\nu(\lambda)$ . To avoid a direct dependence on  $S$ , one can define the vector  $\mathbf{z}$  over a discretization of the state space.

**Deep Recurrent Policies** Another noteworthy way to do function approximation on the history is to employ recurrent neural networks (Hochreiter & Schmidhuber, 1997) to represent the non-Markovian policy. This kind of recurrent architecture is already popular in RL. In this paper we are providing the theoretical ground to motivate the use of deep recurrent policies to address maximum state entropy exploration.

**Non-Markovian Control with Tree Search** In principle, one can get a realization of actions from the optimal non-Markovian policy without ever computing it, e.g., by employing a Monte-Carlo Tree Search (MCTS) (Kocsis & Szepesvári, 2006) approach to select the next action to take. Given the current state  $s_t$  as a root, we can build the tree of trajectories from the root through repeated simulations of potential action sequences. With a sufficient number of simulations and a sufficiently deep tree, we are guaranteed to select the optimal action at the root. If the horizon is too long, we can still cut the tree at any depth and approximately evaluate a leaf node with the entropy induced by the path from the root to the leaf. The drawback of this MCTS procedure is that we would require to have access to a simulator with reset (or a reliable estimate of the transition model) to actually build the tree.

Having reported interesting directions to learn non-Markovian exploration policies in practice, we would like to mention some relevant online RL settings that might benefit from such exploration policies. We leave as future work a more formal definition of the settings and an empirical study.

**Single-Trial RL** In many relevant real-world scenarios, where data collection might be costly or non-episodic in nature, we cannot afford multiple trials to achieve the desired exploration of the environment. Non-Markovian exploration policies guarantee a good coverage of the environment in a single trial and they are particularly suitable for online learning processes.

**Learning in Latent MDPs** In a latent MDP scenario (Hallak et al., 2015; Kwon et al., 2021) an agent interacts with an (unknown) environment drawn from a class of MDPs to solve an online RL task. A non-Markovian exploration policy pre-trained on the whole class could exploit the memory to perform a fast identification of the specific context that has been drawn, quickly adapting to the optimal environment-specific policy.

To conclude, we believe that this work sheds some light on the, previously neglected, importance of non-Markovianity to address maximum state entropy exploration, and we believe it can provide inspiration for future empirical and theoretical contributions on the matter.