The Importance of Non-Markovianity in Maximum State Entropy Exploration

Mirco Mutti [*] [1] [2] **Riccardo De Santi** [*] [1] [3] **Marcello Restelli** [1]

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 and how non-Markovian exploration could benefit the sample efficiency of online reinforcement learning 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), especially w.r.t. RL from scratch.

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). Similarly, (Hazan et al., 2019, Lemma 3.3) note that the class of Markovian strategies is *sufficient* for the standard MSE objective. Indeed, a carefully constructed Markovian strategy is able to induce the same state distribution of any history-based (non-Markovian) one by exploiting randomization. Crucially, this result does not hold only for asymptotic state distributions (Puterman, 2014), but also for state distributions that are marginalized over a finite horizon. As a matter of fact, there is little incentive to consider more complicated strategies as they are not providing any benefit on the value of 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: If you know what you have visited already, you can take decisions accordingly. 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

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

Unsupervised Reinforcement Learning Workshop at ICML 2021.

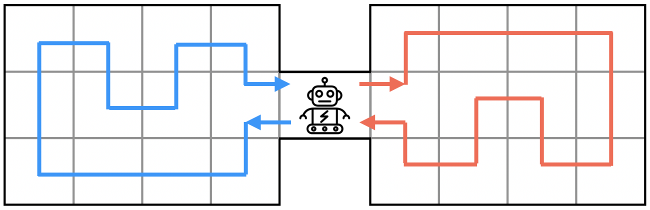


Figure 1: Illustrative two-rooms domain. The agent starts in the middle, colored traces represent optimal strategies to explore the left room and the right room respectively.

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 twice, or the right room twice, or each room once, and all of this outcomes have the same probability. Thus, the agent might explore poorly when considering a single episode, but the exploration is uniform in the average of *infinite trials*. Arguably, this is quite different from how a human being would tackle this problem, i.e., taking intentional decisions in the middle position to visit a room before going to the other. This strategy leads to uniform exploration of the environment in *any trial*, but it is inherently non-Markovian.

Backed by this intuition, we argue that prior work does not recognize the importance of non-Markovianity in MSE exploration due to an hidden infinite-samples assumption in the objective formulation. In this paper, we introduce a new *finite-sample* MSE objective, which targets the expected entropy of the state visitation frequency induced within an episode instead of the entropy of the expected state visitation frequency over infinite samples. In this finite-sample formulation non-Markovian strategies are crucial, and we believe they can benefit a significant range of relevant applications. For example, collecting task-specific samples might be costly in some real-world domains, and a pre-trained non-Markovian strategy is essential to guarantee quality exploration even in a single-trial setting. In another instance, one might aim to pre-train an exploration strategy for a class of multiple environments instead of a single one. A non-Markovian strategy could exploit the history of interactions to swiftly identify the structure of the environment, then employing the environment-specific optimal strategy thereafter. The aim of this paper is to highlight the importance of non-Markovinaity to fulfill the promises of maximum state entropy exploration.

The contributions are organized as follows. First, 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, including the entropy of the induced marginal state distributions in episodic settings. 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, for which we provide lower and upper bounds. 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 in general. Despite the hardness result, we provide a numerical validation of the theory (Section 6), and we comment some potential options to address the problem in a tractable way (Section 7). In Appendix A, we discuss related work, while the missing proofs can be found in Appendix B. Finally, we provide some additional remarks on the role of non-stationarity and state-action entropy objectives (Appendix C).

2. Preliminaries

In this section, we report the notation and the basic background notions we will make use of. We will denote with $\Delta(\mathcal{X})$ a distribution over the space $\mathcal{X}$, and with $[T]$ the set of integers $\{0, \dots, T-1\}$.

Controlled Markov Processes A Controlled Markov Process (CMP) is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, \mu)$, where $\mathcal{S}$ is a finite state space ($|\mathcal{S}| = S$), $\mathcal{A}$ is a finite action space ($|\mathcal{A}| = A$), $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition model, such that $P(s'|a, s)$ denotes the conditional probability of reaching state $s' \in \mathcal{S}$ when selecting action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, and $\mu : \Delta(\mathcal{S})$ is the initial state distribution.

Policies A policy π defines the behavior of an agent interacting with an environment modelled by a CMP. It consists of a sequence of decision rules $\pi := (\pi_1, \pi_2, \dots, \pi_t, \dots)$. Each of them is a map between histories $h := (s_0, a_0, \dots, a_{t-1}, s_t) \in \mathcal{H}_t$ and actions $\pi_t : \mathcal{H}_t \to \Delta(\mathcal{A})$, such that $\pi_t(a|h)$ defines the conditional probability of taking action $a \in \mathcal{A}$ having experienced the history $h \in \mathcal{H}_t$. We denote as $\mathcal{H}$ the space of the histories of arbitrary length. We denote as Π the set of all policies, and as Π^{D} the set of deterministic policies $\pi = (\pi_t)_{t=1}^{\infty}$ such that $\pi_t : \mathcal{H}_t \to \mathcal{A}$. We further define relevant subsets of Π:

- *Non-Markovian* (NM) policies Π_{NM}, where each $\pi \in \Pi_{\mathrm{NM}}$ collapses to a single time-invariant decision rule $\pi = (\pi, \pi, \dots)$ such that $\pi : \mathcal{H} \to \Delta(\mathcal{A})$;
- *Non-Stationary* (NS) policies Π_{NS}, where each $\pi \in \Pi_{\mathrm{NS}}$ is defined through a sequence of Markovian decision rules $\pi = (\pi_1, \pi_2, \dots, \pi_t, \dots)$ such that $\pi_t : \mathcal{S} \to \Delta(\mathcal{A})$;
- *Markovian* (M) policies Π_{M}, where each $\pi \in \Pi_{\mathrm{M}}$ collapses to a single, time-invariant, Markovian decision rule $\pi = (\pi, \pi, \dots)$ such that $\pi : \mathcal{S} \to \Delta(\mathcal{A})$.

State Distributions and Visitation Frequency A policy $\pi \in \Pi$ interacting with a CMP induces a t-step state distribution $d_t^\pi(s) := Pr(s_t = s|\pi)$ over $\mathcal{S}$ (Puterman, 2014). This distribution is described by the temporal relation $d_t^\pi(s) = \int_{\mathcal{S}} \int_{\mathcal{A}} d_{t-1}^\pi(s', a') P(s|s', a') \, \mathrm{d}s' \, \mathrm{d}a'$, where $d_t^\pi(\cdot, \cdot) : \Delta(\mathcal{S} \times \mathcal{A})$ is the t-step state-action distribution. We call the asymptotic fixed point of this temporal relation the *stationary state distribution* $d_\infty^\pi(s) := \lim_{t \to \infty} d_t^\pi(s)$, and we denote as $d_\gamma^\pi(s) := (1-\gamma) \sum_{t=0}^\infty \gamma^t d_t^\pi(s)$ its γ-discounted counterpart ($\gamma \in (0,1)$ is the discount factor). A marginalization of the t-step state distribution over a finite horizon T, i.e., $d_T^\pi(s) := \frac{1}{T} \sum_{t \in [T]} d_t^\pi(s)$, is called the *marginal state distribution*. The *state visitation frequency* $d_h(s) = \frac{1}{T} \sum_{t \in [T]} \mathbb{1}(s_t = s|h)$ is a realization of the marginal state distribution, such that $\mathbb{E}_{h \sim p^\pi} \big[d_h(s) \big] = d_T^\pi(s)$, where the distribution over histories $p^\pi : \Delta(\mathcal{H})$ is defined as $p_T^\pi(h) = \mu(s_0) \prod_{t \in [T-1]} \pi(a_t|h_t) P(s_{t+1}|a_t, s_t)$.

Markov Decision Processes A CMP $\mathcal{M}$ paired with a reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 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 when taking action $a \in \mathcal{A}$ in $s \in \mathcal{S}$, and with $R(h) = \sum_{t \in [T]} R(s_t, a_t)$ the return over the horizon T. The performance of a policy π over the MDP $\mathcal{M}^R$ is defined as the *average return* $\mathcal{J}_{\mathcal{M}^R}(\pi) = \mathbb{E}_{h \sim p_T^\pi}[R(h)]$, and $\pi_{\mathcal{J}}^* \in \arg\max_{\pi \in \Pi} \mathcal{J}_{\mathcal{M}^R}(\pi)$ is called an optimal policy. For any MDP $\mathcal{M}^R$, there always exists a deterministic Markovian policy $\pi \in \Pi_{\mathrm{M}}^{\mathrm{D}}$ that is optimal.

Extended MDP The problem of finding an optimal non-Markovian policy with history-length T in an MDP $\mathcal{M}^R$, i.e., $\pi_{\mathrm{NM}}^* \in \arg\max_{\pi \in \Pi_{\mathrm{NM}}} \mathcal{J}_{\mathcal{M}^R}(\pi)$, can be reformulated as the one of finding an optimal Markovian policy $\pi_{\mathrm{M}}^* \in \arg\max_{\pi \in \Pi_{\mathrm{M}}} \mathcal{J}_{\widetilde{\mathcal{M}}_T^R}(\pi)$ in an extended MDP $\widetilde{\mathcal{M}}_T^R$. The extended MDP is defined as $\widetilde{\mathcal{M}}_T^R := (\widetilde{\mathcal{S}}, \widetilde{\mathcal{A}}, \widetilde{P}, \widetilde{R}, \widetilde{\mu})$, in which $\widetilde{\mathcal{S}} \subseteq \mathcal{H}_{[T]} = \mathcal{H}_1 \cup \ldots \cup \mathcal{H}_T$, 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.

Partially Observable MDP A Partially Observable Markov Decision Process (POMDP) (Astrom, 1965; Kaelbling et al., 1998) is described by $\mathcal{M}_\Omega^R := (\mathcal{S}, \mathcal{A}, P, R, \mu, \Omega, O)$, where $\mathcal{S}, \mathcal{A}, P, R, \mu$ are defined as in an MDP, Ω is a finite observation space, and $O : \mathcal{S} \times \mathcal{A} \to \Delta(\Omega)$ is the observation function, such that $O(o|s', a)$ denotes the conditional probability of the observation $o \in \Omega$ when selecting action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. Crucially, while interacting with a POMDP the agent cannot observe the state $s \in \mathcal{S}$, but just the observation $o \in \Omega$. The performance of a policy π is defined as in an MDP.

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:

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

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). While it is well-known (Puterman, 2014) that there exists an optimal deterministic policy $\pi^* \in \Pi_{\mathrm{M}}^{\mathrm{D}}$ for the common average return objective $\mathcal{J}_{\mathcal{M}^R}$, it is not pointless to wonder whether the reward-free objective in (1) would require a more powerful policy class than Π_{M}. Unsurprisingly, Hazan et al. (Hazan et al., 2019, Lemma 3.3) confirm that the set of (randomized) Markovian policies Π_{M} is indeed sufficient for $\mathcal{E}_\infty$ defined over asymptotic (either stationary or discounted) state distributions. In the following theorem and corollary, we build upon common MDP results (Puterman, 2014) to show that Π_{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 $\mathcal{D}_{\mathrm{NM}}^x = \{d_x^\pi(\cdot) : \pi \in \Pi_{\mathrm{NM}}\}$, $\mathcal{D}_{\mathrm{NS}}^x = \{d_x^\pi(\cdot) : \pi \in \Pi_{\mathrm{NS}}\}$, $\mathcal{D}_{\mathrm{M}}^x = \{d_x^\pi(\cdot) : \pi \in \Pi_{\mathrm{M}}\}$ the corresponding sets of distributions. We can prove that:*

(i) The sets of stationary state distributions are equivalent $\mathcal{D}_{\mathrm{NM}}^\infty \equiv \mathcal{D}_{\mathrm{NS}}^\infty \equiv \mathcal{D}_{\mathrm{M}}^\infty$;

(ii) The sets of discounted state distributions are equivalent $\mathcal{D}_{\mathrm{NM}}^\gamma \equiv \mathcal{D}_{\mathrm{NS}}^\gamma \equiv \mathcal{D}_{\mathrm{M}}^\gamma$ for any γ;

(iii) The sets of marginal state distributions are equivalent $\mathcal{D}_{\mathrm{NM}}^T \equiv \mathcal{D}_{\mathrm{NS}}^T \equiv \mathcal{D}_{\mathrm{M}}^T$ for any T.

Proof Sketch. The results *(i), (ii)* are a consequence of (Hazan et al., 2019, Lemma 3.3), and we refer to Appendix B.1 for a complete proof. Here we focus on the result *(iii)*. From (Puterman, 2014, Theorem 5.5.1) we know that, for any $\pi \in \Pi_{\mathrm{NM}}$, we can build a $\pi' \in \Pi_{\mathrm{NS}}$ having $d_t^{\pi'}(s) = d_t^{\pi}(s)$ for every $t \geq 0$ and $s \in \mathcal{S}$, which implies $\mathcal{D}_{\mathrm{NM}}^T \equiv \mathcal{D}_{\mathrm{NS}}^T$. Thus, it is sufficient to show $\mathcal{D}_{\mathrm{NS}}^T \equiv \mathcal{D}_{\mathrm{M}}^T$. The key point of the proof is that, for any $\pi \in \Pi_{\mathrm{NS}}$, we can build a policy $\pi' \in \Pi_{\mathrm{M}}$ inducing $d_T^{\pi'}(\cdot) = d_T^{\pi}(\cdot)$ by *marginalizing* π over the time-steps $t \in [T-1]$, i.e., $\pi'(a|s) = \sum_{t\in[T-1]} d_t^{\pi}(s)\pi_t(a|s) \big/ \sum_{t\in[T-1]} d_t^{\pi}(s), \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$. This gives $\mathcal{D}_{\mathrm{NS}}^T \subseteq \mathcal{D}_{\mathrm{M}}^T$ and concludes the proof. □

From the equivalence of the sets of induced distributions, it is straightforward to derive the following corollary on the optimality of Markovian policies for objective (1).

Corollary 3.2. *For every CMP $\mathcal{M}$, there exists a Markovian policy $\pi^* \in \Pi_{\mathrm{M}}$ 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), since there is no clear advantage to make up for the additional complexity of the policy. This result might be unsurprising when considering asymptotic distributions, as one can expect a carefully constructed Markovian policy to be able to tie the distribution induced by a non-Markovian (or non-stationary) policy in the limit of the interaction steps. However, it is less evident that a similar property holds for the expectation of final-length interactions alike. Yet, we were able to prove that a Markovian policy that properly exploits randomization can always achieve equivalent state distributions w.r.t. non-Markovian (or non-stationary) counterparts. Note that state distributions are actually *expected* state visitation frequency, and the expectation practically implies an infinite number of realizations. 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, which results in

$$\mathcal{E}(\pi) := \mathop{\mathbb{E}}_{h\sim p_T^\pi} \big[Entropy\big(d_h(\cdot)\big)\big] \tag{2}$$

$$= - \mathop{\mathbb{E}}_{h\sim p_T^\pi} \mathop{\mathbb{E}}_{s\sim d_h} \big[\log d_h(s)\big]. \tag{3}$$

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 (3) 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^π. Indeed, we show in this section that non-Markovianity provides a superior policy class for objective (3). First, we define a performance measure to formally assess this benefit, which we call *regret-to-go*.[1]

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 π as*

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

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 the state 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 π, by how much would we decrease (in expectation) the entropy of the state visits $Entropy(d_{h_T}(\cdot))$ w.r.t. an optimal policy π^*? In particular, we would like to know how limiting the policy π 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.

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

Proof. The result $\mathcal{R}_{T-t}(\pi_{\mathrm{NM}}) = 0$ is straightforward by noting that the set of non-Markovian policies Π_{NM} with arbitrary history-length is as powerful as the general set of policies Π. To show that there exists a deterministic π_{NM}, we consider the extended MDP $\widetilde{\mathcal{M}}_T^R$ obtained from

[1] Note that the entropy function does not enjoy additivity, thus we cannot adopt the usual expected cumulative regret formulation in this setting.

the CMP $\mathcal{M}$ as in Section 2, in which the extended reward function is $\widetilde{R}(\widetilde{s},\widetilde{a}) = Entropy(d_{\widetilde{s}}(\cdot))$ 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 $\widetilde{\pi}_{\mathrm{M}} \in \Pi^{\mathrm{D}}_{\mathrm{M}}$ on $\widetilde{\mathcal{M}}^{R}_{T}$ can be mapped to a non-Markovian policy $\pi_{\mathrm{NM}} \in \Pi^{\mathrm{D}}_{\mathrm{NM}}$ on $\mathcal{M}$, and it is well-known (Puterman, 2014) that for any MDP there exists an optimal deterministic Markovian policy, we have that $\widetilde{\pi}_{\mathrm{M}} \in \arg\max_{\pi\in\Pi_{\mathrm{M}}} \mathcal{J}_{\widetilde{\mathcal{M}}^{R}_{T}}(\pi)$ implies $\pi_{\mathrm{NM}} \in \arg\max_{\pi\in\Pi_{\mathrm{NM}}} \mathcal{E}(\pi)$. □

Whereas Lemma 4.2 ensures that the set of non-Markovian policies Π_{NM} is sufficient for the objective (3), we would like to know if it is also necessary. Especially, we aim to assess whether there exist CMPs in which a Markovian policy $\pi \in \Pi_{\mathrm{M}}$ would suffer non-zero regret-to-go. First, it is worth showing that Markovian policies can rely on randomization to optimize objective (3).

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

$$\mathop{\mathbb{V}\mathrm{ar}}_{a\sim\pi_{\mathrm{M}}(s)} \big[a\big] = \mathop{\mathbb{V}\mathrm{ar}}_{hs\sim p^{\pi_{\mathrm{NM}}}_{T}} \big[\pi_{\mathrm{NM}}(hs)\big], \qquad \forall s \in \mathcal{S},$$

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

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

Especially, Lemma 4.3 shows that, whenever the optimal strategy for objective (3) (i.e., the non-Markovian π_{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., π_{M}) must necessarily be randomized. This is crucial to prove the main result of this section, which establishes a lower bound $\underline{\mathcal{R}}_{T-t}$ and an upper bound $\overline{\mathcal{R}}_{T-t}$ to the expected regret-to-go of any Markovian policy that optimizes objective (3).

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

$$\underline{\mathcal{R}}_{T-t}(\pi_{\mathrm{M}}) = \frac{E_{max} - E_{max,2}}{\pi_{\mathrm{M}}(a_{\mathrm{NM}}|s_t)} \mathop{\mathbb{V}\mathrm{ar}}_{hs_t\sim p^{\pi_{\mathrm{NM}}}_{T}} \big[\pi_{\mathrm{NM}}(hs_t)\big],$$
$$\overline{\mathcal{R}}_{T-t}(\pi_{\mathrm{M}}) = \frac{E_{max} - E_{min,t}}{\pi_{\mathrm{M}}(a_{\mathrm{NM}}|s_t)} \mathop{\mathbb{V}\mathrm{ar}}_{hs_t\sim p^{\pi_{\mathrm{NM}}}_{T}} \big[\pi_{\mathrm{NM}}(hs_t)\big],$$

where $\pi_{\mathrm{NM}} \in \arg\max_{\pi\in\Pi^{\mathrm{D}}_{\mathrm{NM}}} \mathcal{E}(\pi)$, $a_{\mathrm{NM}} = \pi_{\mathrm{NM}}(h^*_t)$ is the unique optimal action in s_t, and E_{max}, $E_{max,2}$, $E_{min,t}$ are given by

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

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 B.1, B.2). This gives $\mathcal{R}_{T-t}(\pi_{\mathrm{M}}) \geq E_{max} - \pi_{\mathrm{M}}(a_{\mathrm{NM}}|s_t)E_{max} - \big(1 - \pi_{\mathrm{M}}(a_{\mathrm{NM}}|s_t)\big)E_{max,2}$ and $\mathcal{R}_{T-t}(\pi_{\mathrm{M}}) \leq E_{max} - \pi_{\mathrm{M}}(a_{\mathrm{NM}}|s_t)E_{max} - \big(1 - \pi_{\mathrm{M}}(a_{\mathrm{NM}}|s_t)\big)E_{min,t}$. 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 $\mathbb{V}\mathrm{ar}_{a\sim\pi_{\mathrm{NM}}(s_t)}[a] = \pi_{\mathrm{M}}(a_{\mathrm{NM}}|s_t)\big(1 - \pi_{\mathrm{M}}(a_{\mathrm{NM}}|s_t)\big) = \mathbb{V}\mathrm{ar}_{hs_t\sim p^{\pi_{\mathrm{NM}}}_{T}} \big[\pi_{\mathrm{NM}}(hs_t)\big]$, which concludes the proof. □

Theorem 4.4 basically states that, whenever the Markovian policy (π_{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_{NM} in s_t.[2] Clearly, the value of this regret is related to the probability of taking a sub-optimal action (through the factor $1/\pi_{\mathrm{M}}(a_{\mathrm{NM}}|s_t)$) and the variance of the optimal strategy in s_t over the possible histories (through the factor $\mathbb{V}\mathrm{ar}_{hs_t\sim p^{\pi_{\mathrm{NM}}}_{T}}[\pi_{\mathrm{NM}}(hs_t)]$). To compute the regret-to-go exactly, one should have access to the full structure of the CMP $\mathcal{M}$ and its transition dynamics P. Instead, we provide an upper bound and a lower bound to the regret that only depends 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

[2] Note that this assumption could be easily removed by partitioning the action space in s_t as $\mathcal{A}(s_t) = \mathcal{A}_{opt}(s_t) \cup \mathcal{A}_{sub-opt}(s_t)$, such that $\mathcal{A}_{opt}(s_t)$ are optimal actions and $\mathcal{A}_{sub-opt}(s_t)$ are sub-optimal, and substituting the term $1/\pi_{\mathrm{M}}(a_{\mathrm{NM}}|s_t)$ with $1/\sum_{a\in\mathcal{A}_{opt}(s_t)} \pi_{\mathrm{M}}(a|s_t)$ in the regret bounds.

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

Finally, although the objective (3) 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 to the regret suffered by an optimal Markovian policy.

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 π at step t as $r_t(\pi) := \max\big(0, \mathcal{R}_{T-t}(\pi) - \mathcal{R}_{T-t-1}(\pi)\big)$.*

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

$$\underline{r}_t(\pi_{\mathrm{M}}) = \max\Big(0,\ E_{max}\big(\mathcal{V}_t(\pi_{\mathrm{M}}) - \mathcal{V}_{t+1}(\pi_{\mathrm{M}})\big) - E_{max,2}\mathcal{V}_t(\pi_{\mathrm{M}}) + E_{min,t+1}\mathcal{V}_{t+1}(\pi_{\mathrm{M}})\Big),$$

$$\overline{r}_t(\pi_{\mathrm{M}}) = \max\Big(0,\ E_{max}\big(\mathcal{V}_t(\pi_{\mathrm{M}}) - \mathcal{V}_{t+1}(\pi_{\mathrm{M}})\big) - E_{min,t}\mathcal{V}_t(\pi_{\mathrm{M}}) + E_{max,2}\mathcal{V}_{t+1}(\pi_{\mathrm{M}})\Big),$$

where

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

5. Complexity Analysis

Having established the importance of non-Markovianity in dealing with MSE exploration in a finite-sample regime, it is worth considering how hard it is to optimize the objective 3 within the class of non-Markovian policies. Especially, we aim at characterizing the complexity of the problem:

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

defined over a CMP $\mathcal{M}$. First, we recall that Ψ_0 can be rewritten as the problem of finding a reward-maximizing Markovian policy, i.e., $\widetilde{\pi}_{\mathrm{M}} \in \arg\max_{\pi \in \Pi_{\mathrm{M}}} \mathcal{J}_{\widetilde{\mathcal{M}}_T^R}(\pi)$, over a convenient extended MDP $\widetilde{\mathcal{M}}_T^R$ obtained from CMP $\mathcal{M}$ (see the proof of Lemma 4.2 for further details). We call this problem $\widetilde{\Psi}_0$ and we note that $\widetilde{\Psi}_0 \in \mathrm{P}$, 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 lemma shows that it does not exist a many-to-one reduction from Ψ_0 to $\widetilde{\Psi}_0$.

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

Proof. In the general case, coding any instance of Ψ_0 in the representation required by $\widetilde{\Psi}_0$ holds exponential complexity w.r.t. the input of the initial instance of Ψ_0. □

The latter result informally suggests that $\Psi \notin \mathrm{P}$. Indeed, we can show that $\Psi \notin \mathrm{NP}$.

Lemma 5.2. $\Psi_0 \notin \mathrm{NP}$.

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 Ψ_0 and a candidate solution $\pi \in \Pi_{\mathrm{NM}}$, if there exists an algorithm Λ that can verify the optimality of π in polynomial time, then $\Psi_0 \in NP$. We prove that $\Psi_0 \notin NP$ 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 $\pi \in \Pi_{\mathrm{NM}}$, i.e., a candidate solution of an instance of the problem Ψ_0, a polynomial verifier Λ 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 π is optimal according to Ψ_0, but this operation is exponential w.r.t. the input of Ψ_0. □

As a direct consequence, $\Psi_0 \notin$ NP-complete. We can now prove the main theorem of this section, which shows that Ψ_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 Ψ_0. To derive the reduction we consider two intermediate problems, namely Ψ_1 and Ψ_2. Especially, we aim to show that the following chain of reductions holds

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

First, we define Ψ_1 and we prove that $\Psi_0 \geq_m \Psi_1$. Informally, Ψ_1 is the problem of finding a reward-maximizing Markovian policy $\pi_{\mathrm{M}} \in \Pi_{\mathrm{M}}$ w.r.t. the entropy objective (3) 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 $\mathcal{M}$), and $\widetilde{O}(\widetilde{o}|\widetilde{s}) = \widetilde{s}_{-1}$. 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 Ψ_i. We show that Ψ_0 is harder than Ψ_1 by defining the polynomial-time functions ψ and ϕ such that any instance of Ψ_1 can be rewritten through ψ as an instance of Ψ_0, and a solution $\pi^*_{\mathrm{NM}} \in \Pi_{\mathrm{NM}}$ for Ψ_0 can be converted through ϕ into a solution $\pi^*_{\mathrm{M}} \in \Pi_{\mathrm{M}}$ for the original instance of Ψ_1. The function ψ sets $\mathcal{S} = \widetilde{\Omega}$ and derives the transition model of $\mathcal{M}$ from the one of $\widetilde{\mathcal{M}}^R_{\Omega}$, while ϕ converts the optimal solution of Ψ_0 by computing $\pi^*_{\mathrm{M}} = \sum_{ho \in \mathcal{H}_o} p^{\pi^*_{\mathrm{NM}}}(ho) \pi^*_{\mathrm{NM}}(a|ho)$, 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 Ψ_2 as the policy existence problem w.r.t. the problem statement of Ψ_1. Hence, Ψ_2 is the problem of determining whether the value of a reward-maximizing Markovian policy $\pi^*_{\mathrm{M}} \in \arg\max_{\pi \in \Pi_{\mathrm{M}}} \mathcal{J}_{\widetilde{\mathcal{M}}^R_{\Omega}}(\pi)$ 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 Ψ_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. □

6. 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 Section 7 why we believe that the development of scalable methods is not hopeless even in this challenging setting.

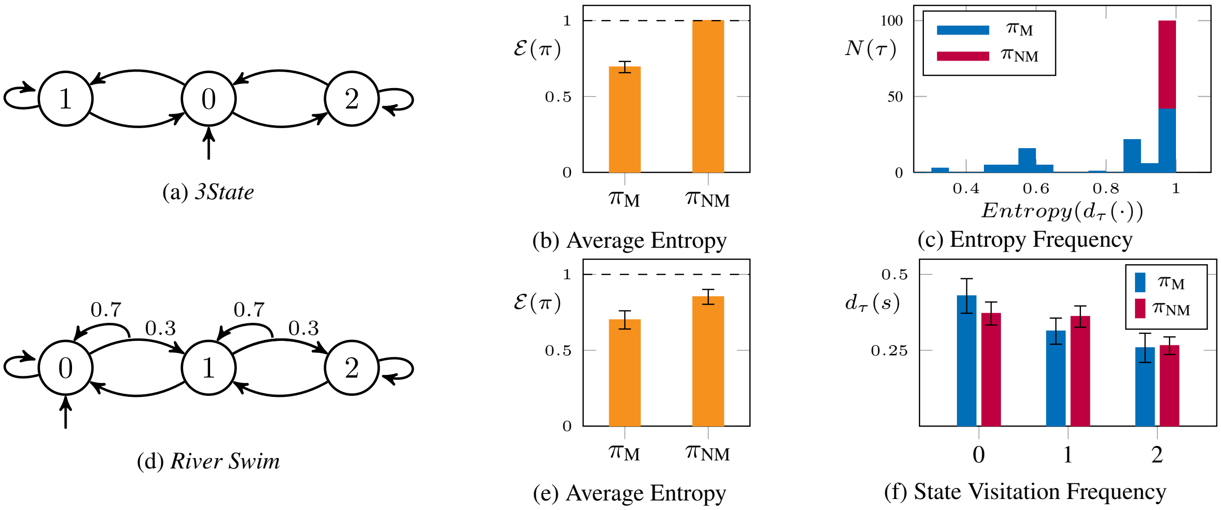
In this section, we consider a *3State* ($S = 3, A = 2, T = 9$), which is a simple abstraction of the two-rooms in Figure 1, and a *River Swim* (Strehl & Littman, 2008) ($S = 3, A = 2, T = 10$) that are depicted in Figure 2a, 2d respectively. Especially, we compare the expected entropy (3) achieved by an optimal non-Markovian policy $\pi_{\mathrm{NM}} \in \arg\max_{\pi \in \Pi_{\mathrm{NM}}} \mathcal{E}(\pi)$, which is obtained by solving the extended MDP as described in the proof of Lemma 4.2, against an optimal Markovian policy $\pi_{\mathrm{M}} \in \arg\max_{\pi \in \Pi_{\mathrm{M}}} \mathcal{E}(\pi)$, which is obtained from π_{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, π_{M} cannot match the performance of π_{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 π_{NM} is able to do that, and it always realizes the optimal trajectory (Figure 2c). Instead, π_{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 π_{M} and π_{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.

7. Discussion and Conclusion

In the previous sections, we detailed 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., $\pi : \mathcal{H} \to \Delta(\mathcal{A})$. 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 $\pi : \mathcal{H}_H \to \Delta(\mathcal{A})$ 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 $\pi : f(\mathcal{H}) \to \Delta(\mathcal{A})$, 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



(a) *3State*

(b) Average Entropy

(c) Entropy Frequency

(d) *River Swim*

(e) Average Entropy

(f) State Visitation Frequency

Figure 2: In **(a, d)**, we illustrates the *3State* and *River Swim* CMPs. Then, we report the average entropy induced by an optimal Markovian policy π_{M} and an optimal non-Markovian policy π_{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.

length S, which is updated as $\boldsymbol{z}_{t+1} \leftarrow \lambda \boldsymbol{z}_t + \mathbf{1}_{s_t}$, 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 $\boldsymbol{z}_0 = 0$. The discount factor λ 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 $\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{z}, \lambda)$ and a parametric policy over the discount $\pi_{\boldsymbol{\nu}}(\lambda)$. To avoid a direct dependence on S, one can define the vector $\boldsymbol{z}$ 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 (Williams & Zipser, 1989; 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 a thorough 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.

References

Astrom, K. J. Optimal control of markov decision processes with incomplete state estimation. *Journal Mathematical Analysis and Applications*, 1965.

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.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 1998.

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.

Williams, R. J. and Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1989.

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. Related Work

Hazan et al. (Hazan et al., 2019) were the first to consider an entropic measure over the state distribution as a sensible learning objective for an agent interacting with a reward-free environment (Jin et al., 2020). Especially, they propose an algorithm, called MaxEnt, that learns a mixture of policies that collectively maximize the Shannon entropy, i.e., (1), of the discounted state distribution. The final mixture is learned through a conditional gradient method, in which the algorithm iteratively estimates the state distribution of the current mixture to define an intrinsic reward function, and then identifies the next policy to be added by solving a specific RL sub-problem with this reward. A similar methodology has been obtained by Lee et al. (Lee et al., 2019) from a game-theoretic perspective on the MSE exploration problem. Their algorithms, called SMM, targets the Shannon entropy of the marginal state distribution instead of the discounted distribution of MaxEnt. Another approach based on the conditional gradient method is FW-AME (Tarbouriech & Lazaric, 2019), which learns a mixture of policies to maximize the entropy of the stationary state-action distribution. As noted in (Tarbouriech & Lazaric, 2019), the mixture of policies might suffer a slow mixing to the asymptotic distribution for which the entropy is maximized. In (Mutti & Restelli, 2020), the authors present a method (IDE^{3}AL) to learn a single exploration policy that simultaneously accounts for the entropy of the stationary state-action distribution and the mixing time.

Even if they are sometimes evaluated on continuous domains (especially (Hazan et al., 2019; Lee et al., 2019)), the methods we mentioned require an accurate estimate of either the state distribution (Hazan et al., 2019; Lee et al., 2019) or the transition model (Tarbouriech & Lazaric, 2019; Mutti & Restelli, 2020), which hardly scales to high-dimensional domains. A subsequent work (Mutti et al., 2021b) proposes an approach to estimate the entropy of the state distribution through a non-parametric method, and then to directly optimize the estimated entropy via policy optimization. Their algorithm, called MEPOL, is able to learn a single exploration policy that maximizes the entropy of the marginal state distribution in challenging continuous control domains. Liu and Abbeel (Liu & Abbeel, 2021b) combine non-parametric entropy estimation with learned state representations into an algorithm, called APT, that successfully addresses MSE exploration problems in visual-inputs domains. Seo et al. (Seo et al., 2021) shows that even random state representations are sufficient to learn MSE exploration policies from visual inputs.

Whereas all of the previous approaches accounts for the Shannon entropy in their objectives, recent works (Zhang et al., 2020a; Guo et al., 2021) consider alternative formulations. Especially, Zhang et al. (Zhang et al., 2020a) argues that the Rényi entropy provides a superior incentive to cover all of the corresponding space than the Shannon entropy, and they propose a method to optimize the Rényi of the state-action distribution via gradient ascent (MaxRényi). On an orthogonal direction, the authors of (Guo et al., 2021) consider a reformulation of the entropy function that accounts for the underlying geometry of the space. They present a method, called GEM, to learn an optimal policy for the geometry-aware entropy objective.

Table 1: Overview of the methods addressing MSE exploration in a controlled Markov process. For each method, we report the nature of the corresponding MSE objective, i.e., the entropy function (Entropy), whether it considers stationary, discounted, or marginal distributions (Distribution), and if it accounts for the state space $\mathcal{S}$ or the state-action space $\mathcal{SA}$ (Space). We also specify if the method learns a single policy rather than a mixture of policies (Mixture), and if it supports non-parametric entropy estimation (Non-Parametric).

Algorithm	Entropy	Distribution	Space	Mixture	Non-Parametric
MaxEnt (Hazan et al., 2019)	Shannon	discounted	state	✓	✗
FW-AME (Tarbouriech & Lazaric, 2019)	Shannon	stationary	state-action	✓	✗
SMM (Lee et al., 2019)	Shannon	marginal	state	✓	✗
IDE^{3}AL (Mutti & Restelli, 2020)	Shannon	stationary	state-action	✗	✗
MEPOL (Mutti et al., 2021b)	Shannon	marginal	state	✗	✓
MaxRényi (Zhang et al., 2020a)	Rényi	discounted	state-action	✗	✗
GEM (Guo et al., 2021)	geometric-aware	marginal	state	✗	✗
APT (Liu & Abbeel, 2021b)	Shannon	marginal	state	✗	✓
RE3 (Seo et al., 2021)	Shannon	marginal	state	✗	✓

B. Missing Proofs

B.1. Proofs of Section 3

Theorem 3.1. *Let $x \in \{\infty, \gamma, T\}$, and let $\mathcal{D}^x_{\mathrm{NM}} = \{d^\pi_x(\cdot) : \pi \in \Pi_{\mathrm{NM}}\}$, $\mathcal{D}^x_{\mathrm{NS}} = \{d^\pi_x(\cdot) : \pi \in \Pi_{\mathrm{NS}}\}$, $\mathcal{D}^x_{\mathrm{M}} = \{d^\pi_x(\cdot) : \pi \in \Pi_{\mathrm{M}}\}$ the corresponding sets of distributions. We can prove that:*

(i) The sets of stationary state distributions are equivalent $\mathcal{D}^\infty_{\mathrm{NM}} \equiv \mathcal{D}^\infty_{\mathrm{NS}} \equiv \mathcal{D}^\infty_{\mathrm{M}}$;

(ii) The sets of discounted state distributions are equivalent $\mathcal{D}^\gamma_{\mathrm{NM}} \equiv \mathcal{D}^\gamma_{\mathrm{NS}} \equiv \mathcal{D}^\gamma_{\mathrm{M}}$ for any γ;

(iii) The sets of marginal state distributions are equivalent $\mathcal{D}^T_{\mathrm{NM}} \equiv \mathcal{D}^T_{\mathrm{NS}} \equiv \mathcal{D}^T_{\mathrm{M}}$ for any T.

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

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

For $t = 0$, we have that $d^\pi_0(\cdot) = d^{\pi'}_0(\cdot) = \mu(\cdot)$, which is the initial state distribution. We proceed by induction to show that if $d^\pi_{t-1}(\cdot) = d^{\pi'}_{t-1}(\cdot)$, then we have

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

Since $d^\pi_t(s) = d^{\pi'}_t(s)$ holds for any $t \geq 0$ and $\forall s \in \mathcal{S}$, we have $d^\pi_\infty(\cdot) = d^{\pi'}_\infty(\cdot)$, $d^\pi_\gamma(\cdot) = d^{\pi'}_\gamma(\cdot)$, $d^\pi_T(\cdot) = d^{\pi'}_T(\cdot)$, and thus $\mathcal{D}^x_{\mathrm{NS}} \supseteq \mathcal{D}^x_{\mathrm{NM}}$. Then, $\mathcal{D}^x_{\mathrm{NM}} \equiv \mathcal{D}^x_{\mathrm{NS}}$ follows.

(i) $\mathcal{D}^\infty_{\mathrm{NM}} \equiv \mathcal{D}^\infty_{\mathrm{NS}} \equiv \mathcal{D}^\infty_{\mathrm{M}}$

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

(ii) $\mathcal{D}^\gamma_{\mathrm{NM}} \equiv \mathcal{D}^\gamma_{\mathrm{NS}} \equiv \mathcal{D}^\gamma_{\mathrm{M}}$

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

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

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

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

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

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

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

We can prove the statement by showing that $\mathcal{D}^{T}_{\text{M}} \supseteq \mathcal{D}^{T}_{\text{NS}}$. 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 $\pi' \in \Pi_{\text{M}}$ which marginalizes π over the time steps, such that

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

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

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

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

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

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

B.2. Proofs of Section 4

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

$$\underset{a\sim\pi_{\mathrm{M}}(s)}{\mathbb{V}\mathrm{ar}}\big[a\big] = \underset{hs\sim p_T^{\pi_{\mathrm{NM}}}}{\mathbb{V}\mathrm{ar}}\big[\pi_{\mathrm{NM}}(hs)\big], \qquad \forall s \in \mathcal{S},$$

where hs is any history $hs \in \mathcal{H}_{[T]}$ such that the final state is 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

$$\begin{aligned}
\underset{\pi}{\mathbb{V}\mathrm{ar}}\big[a|s\big] &= \underset{\pi}{\mathbb{E}}\big[a^2|s\big] - \underset{\pi}{\mathbb{E}}\big[a|s\big]^2 \\
&= \underset{h}{\mathbb{E}}\Big[\underset{\pi}{\mathbb{E}}\big[a^2|s,h\big]\Big] - \underset{h}{\mathbb{E}}\Big[\underset{\pi}{\mathbb{E}}\big[a|s,h\big]\Big]^2 \\
&= \underset{h}{\mathbb{E}}\Big[\underset{\pi}{\mathbb{V}\mathrm{ar}}\big[a|s,h\big] + \underset{\pi}{\mathbb{E}}\big[a|s,h\big]^2\Big] - \underset{h}{\mathbb{E}}\Big[\underset{\pi}{\mathbb{E}}\big[a|s,h\big]\Big]^2 \\
&= \underset{h}{\mathbb{E}}\Big[\underset{\pi}{\mathbb{V}\mathrm{ar}}\big[a|s,h\big]\Big] + \underset{h}{\mathbb{E}}\Big[\underset{\pi}{\mathbb{E}}\big[a|s,h\big]^2\Big] - \underset{h}{\mathbb{E}}\Big[\underset{\pi}{\mathbb{E}}\big[a|s,h\big]\Big]^2 \\
&= \underset{h}{\mathbb{E}}\Big[\underset{\pi}{\mathbb{V}\mathrm{ar}}\big[a|s,h\big]\Big] + \underset{h}{\mathbb{V}\mathrm{ar}}\Big[\underset{\pi}{\mathbb{E}}\big[a|s,h\big]\Big].
\end{aligned}$$

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

$$\underset{\pi}{\mathbb{V}\mathrm{ar}}\big[a|s\big] = \underset{hs\sim p_T^{\pi}}{\mathbb{E}}\Big[\underset{\pi}{\mathbb{V}\mathrm{ar}}\big[a|hs\big]\Big] + \underset{hs\sim p_T^{\pi}}{\mathbb{V}\mathrm{ar}}\Big[\underset{\pi}{\mathbb{E}}\big[a|hs\big]\Big]. \tag{4}$$

If we take the policy π that optimizes the objective (3), we have that the action a on the left-hand side of equation (4) is distributed according to $\pi_{\mathrm{M}} \arg\max_{\pi\in\Pi_{\mathrm{M}}} \mathcal{E}(\pi)$, while on the right-hand side is distributed according to $\pi_{\mathrm{NM}} \arg\max_{\pi\in\Pi_{\mathrm{NM}}} \mathcal{E}(\pi)$, i.e.,

$$\underset{a\sim\pi_{\mathrm{M}}(s)}{\mathbb{V}\mathrm{ar}}\big[a\big] = \underset{hs\sim p_T^{\pi_{\mathrm{NM}}}}{\mathbb{E}}\Big[\underset{a\sim\pi_{\mathrm{NM}}(hs)}{\mathbb{V}\mathrm{ar}}\big[a\big]\Big] + \underset{hs\sim p_T^{\pi_{\mathrm{NM}}}}{\mathbb{V}\mathrm{ar}}\Big[\underset{a\sim\pi_{\mathrm{NM}}(hs)}{\mathbb{E}}\big[a\big]\Big]. \tag{5}$$

From Lemma 4.2, we know that there exists a deterministic optimal non-Markovian policy $\pi_{\mathrm{NM}} \in \Pi^{\mathrm{D}}_{\mathrm{NM}}$, which gives $\mathbb{E}_{hs\sim p_T^{\pi_{\mathrm{NM}}}}\big[\mathbb{V}\mathrm{ar}_{a\sim\pi_{\mathrm{NM}}(hs)}[a]\big] = 0$ in (5) and concludes the proof. □

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

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

where the maximum is attained by

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

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

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

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

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

where the minimum is attained by

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

Proof. The worst-case CMP $\underline{\mathcal{M}}$ is designed such that the agent cannot recover from a sub-optimal action $a_t \in \mathcal{A} \setminus \pi_{\mathrm{NM}}(h_t^*)$ 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. □

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

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

where $\pi_{\mathrm{NM}} \in \arg\max_{\pi\in\Pi_{\mathrm{NM}}^{\mathrm{D}}} \mathcal{E}(\pi)$, $a_{\mathrm{NM}} = \pi_{\mathrm{NM}}(h_t^*)$ is the unique optimal action in s_t, and E_{max}, $E_{max,2}$, $E_{min,t}$ are given by

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

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

$$\mathcal{R}_{T-t}(\pi_{\mathrm{M}}) = \max_{\pi^*\in\Pi} \mathop{\mathbb{E}}_{h_{T-t}^*\sim p^{\pi^*}} \big[Entropy\big(d_{h_T^*}(\cdot)\big)\big] - \mathop{\mathbb{E}}_{h_{T-t}\sim p^{\pi_{\mathrm{M}}}} \big[Entropy\big(d_{h_T}(\cdot)\big)\big],$$

in which we substitute $E_{max} = \max_{\pi^*\in\Pi} \mathbb{E}_{h_{T-t}^*\sim p^{\pi^*}} \big[Entropy\big(d_{h_T^*}(\cdot)\big)\big]$. 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 $a \in \mathcal{A} \setminus \{a_{\mathrm{NM}}\}$ in state s_t would have in a best-case and a worst-case CMP respectively, which is detailed in Lemma B.1 and Lemma B.2. Especially, we can write

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

and

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

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

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

which gives

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

□

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

$$\begin{aligned}\underline{r}_t(\pi_{\mathrm{M}}) = \max\Big(0,\ & E_{max}\big(\mathcal{V}_t(\pi_{\mathrm{M}}) - \mathcal{V}_{t+1}(\pi_{\mathrm{M}})\big) \\ & - E_{max,2}\mathcal{V}_t(\pi_{\mathrm{M}}) + E_{min,t+1}\mathcal{V}_{t+1}(\pi_{\mathrm{M}})\Big), \\ \overline{r}_t(\pi_{\mathrm{M}}) = \max\Big(0,\ & E_{max}\big(\mathcal{V}_t(\pi_{\mathrm{M}}) - \mathcal{V}_{t+1}(\pi_{\mathrm{M}})\big) \\ & - E_{min,t}\mathcal{V}_t(\pi_{\mathrm{M}}) + E_{max,2}\mathcal{V}_{t+1}(\pi_{\mathrm{M}})\Big),\end{aligned}$$

where

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

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

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

from Theorem 4.4. We can write

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

and

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

□

B.3. Proofs of Section 5

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

Proof. 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, a well-known NP-complete problem, to Ψ_0. To derive the reduction we consider two intermediate problems, namely Ψ_1 and Ψ_2. Especially, we aim to show that the following chain of reductions hold:

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

First, we define Ψ_1 and we prove that $\Psi_0 \geq_m \Psi_1$. Informally, Ψ_1 is the problem of finding a reward-maximizing Markovian policy $\pi_{\mathrm{M}} \in \Pi_{\mathrm{M}}$ w.r.t. the entropy objective (3) encoded through a reward function in a convenient POMDP $\widetilde{\mathcal{M}}^R_\Omega$. We can build $\widetilde{\mathcal{M}}^R_\Omega$ from the CMP $\mathcal{M}$ similarly as the extended MDP $\widetilde{\mathcal{M}}^R_T$ (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 $\widetilde{s} \in \widetilde{S}$ such that $|\widetilde{s}| \neq T$, otherwise (if $|\widetilde{s}| = T$) the reward corresponds to the entropy value of the state visitation frequences 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 Ψ_i. We show that Ψ_0 is harder than Ψ_1 by defining the polynomial-time functions ψ and ϕ such that any instance of Ψ_1 can be rewritten through ψ as an instance of Ψ_0, and a solution $\pi^*_{\mathrm{NM}} \in \Pi_{\mathrm{NM}}$ for Ψ_0 can be converted through ϕ into a solution $\pi^*_{\mathrm{M}} \in \Pi_{\mathrm{M}}$ for the original instance of Ψ_1.

$$\begin{tikzcd}
\mathcal{I}_{\Psi_1} \arrow[r, "\psi"] & \mathcal{I}_{\Psi_0} \arrow[d] \\
\pi^*_{\mathrm{M}} & \pi^*_{\mathrm{NM}} \arrow[l, "\phi"']
\end{tikzcd}$$

The function ψ sets $\mathcal{S} = \widetilde{\Omega}$ and derives the transition model of $\mathcal{M}$ from the one of $\widetilde{\mathcal{M}}^R_\Omega$, while ϕ converts the optimal solution of Ψ_0 by computing

$$\pi^*_{\mathrm{M}} = \sum_{ho \in \mathcal{H}_o} p_T^{\pi^*_{\mathrm{NM}}}(ho) \pi^*_{\mathrm{NM}}(a|ho) \tag{6}$$

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 Ψ_2 as the policy existence problem w.r.t. the problem statement of Ψ_1. Hence, Ψ_2 is the problem of determining whether the value of a reward-maximizing Markovian policy $\pi^*_{\mathrm{M}} \in \arg\max_{\pi \in \Pi_{\mathrm{M}}} \mathcal{J}_{\widetilde{\mathcal{M}}^R_\Omega}(\pi)$ 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 Ψ_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. $|\widetilde{\mathcal{S}}| = \widetilde{S} = |\widetilde{\Omega}|^T$

Both limitations can be overcome in the following ways:

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 Ψ_1 is such that $\widetilde{S} = |\widetilde{\Omega}|^T$. 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 $|\widetilde{\Omega}|^T = n^{m \cdot n}$, while the POMDPs class used by the proof hereinabove has $\widetilde{S} = m \cdot n^2$. 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 Ψ_1 it suffices to add a set of states $\widetilde{S}_p$ s.t. $R(s) = 0 \; \forall s \in \widetilde{S}_p$.

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. □

C. Non-Stationary Policies and State-Action Distributions

Throughout the paper, we mostly considered the class of non-Markovian policies against the class of Markovian policies. However, one could wonder how non-stationary policies fare in dealing with the finite-sample MSE objective (3). Here we would like to report some informal results on this point. On the one hand, we can prove that there exists a deterministic non-stationary policy $\pi_{\mathrm{NS}} \in \Pi_{\mathrm{NS}}^{\mathrm{D}}$ such that $\pi_{\mathrm{NS}} \in \arg\max_{\pi \in \Pi_{\mathrm{NS}}} \mathcal{E}(\pi)$ through a standard backward-induction argument. Nonetheless, differently from the result in Lemma 4.2 for non-Markovian policies, π_{NS} is not guaranteed to suffer zero regret. Especially, it might happen that $\pi_{\mathrm{NS}}(s_t) \neq a_{\mathrm{NM}}$ as the non-stationary policy is uncertain about the past. One should account for the probability of this event to extend the regret bounds of Theorem 4.4 to non-stationary policies. Whereas π_{NS} is not zero-regret in general, it can suffer a lower regret than an optimal Markovian policy. Formally establishing this regret gap might be an interesting direction for future works.

To assess the importance of non-Markovianity in MSE exploration, we adopted the most common state distribution formulation. As we mentioned in Appendix A, other works in the MSE literature considered the entropy of the state-action distribution in the objective function. Analogously, we can recast the finite-sample objective as $\mathbb{E}_{h \sim p_T^\pi}[Entropy(d_h(\cdot, \cdot))]$ and replicate most of the results that we reported in the paper. In this alternative formulation, the class of non-Markovian policies would still preserve an edge over the classes of non-stationary and Markovian policies. Accounting for the entropy of the state-action distribution might be crucial in the settings where the exploration over the action space cannot be overlooked (e.g., a single-trial setting).