How to Explore with Belief: State Entropy Maximization in POMDPs

Riccardo Zamboni [1] **Duilio Cirino** [1] **Marcello Restelli** [1] **Mirco Mutti** [2]

Abstract

Recent works have studied *state entropy maximization* in reinforcement learning, in which the agent's objective is to learn a policy inducing high entropy over states visitation (Hazan et al., 2019). They typically assume full observability of the state of the system so that the entropy of the observations is maximized. In practice, the agent may only get *partial* observations, e.g., a robot perceiving the state of a physical space through proximity sensors and cameras. A significant mismatch between the entropy over observations and true states of the system can arise in those settings. In this paper, we address the problem of entropy maximization over the *true states* with a decision policy conditioned on partial observations *only*. The latter is a generalization of POMDPs, which is intractable in general. We develop a memory and computationally efficient *policy gradient* method to address a first-order relaxation of the objective defined on *belief* states, providing various formal characterizations of approximation gaps, the optimization landscape, and the *hallucination* problem. This paper aims to generalize state entropy maximization to more realistic domains that meet the challenges of applications.

1. Introduction

The *state entropy maximization* framework, initially proposed in Hazan et al. (2019), is a popular generalization of the Reinforcement Learning (RL, Bertsekas, 2019) problem in which an agent aims to maximize, instead of the cumulative reward, a (self-supervised) objective related to the entropy of the state visitation induced by its policy.

The entropy objective finds motivation as a standalone tool for learning to cover the states of the environment (Hazan et al., 2019), a data collection strategy for offline RL (Yarats et al., 2022), experimental design (Tarbouriech & Lazaric, 2019), or transition model estimation (Tarbouriech et al., 2020; Jin et al., 2020b), and as a surrogate loss for policy pre-training in reward-free settings (Mutti & Restelli, 2020).

While the objective itself is not convex in the policy parameters, it is known to admit a tractable dual formulation (Hazan et al., 2019), and several practical methods, also in combination with neural policies, have been developed (Mutti et al., 2021; Liu & Abbeel, 2021b; Seo et al., 2021; Yarats et al., 2021) as a testament of the promises of the framework for tangible impact on real-world applications.

All of the previous works on state entropy maximization assume the state of the environment is fully observable, such that the agent-environment interaction can be modeled as a Markov Decision Process (MDP, Puterman, 2014). Under this assumption, maximizing the entropy over the observations collected from the environment is well-founded. However, the agent may only receive partial observation from the environment in practice.

Let us think of an autonomous robot for *rescue operations* as an illustrative application: The robot is placed in an unknown terrain with the goal of covering every inch of the ground in order to locate and rescue a wounded human unable to move. The robot cannot access its true position, as well as the human's location-; it can only perceive its surroundings with sensors and cameras. Arguably, maximizing the entropy of the observations, such as changing the camera angle in every direction, is undesirable for the given task. Instead, we would like the robot to maximize the entropy of its position, for which the best policy may entail moving the camera to probe the surroundings and avoid getting stuck, but also to step forward to cover the most ground so that the wounded human can be swiftly located and rescued.

In this paper, we aim to generalize the state entropy maximization framework to scenarios of the kind of the latter, in which the agent only gets partial, potentially noisy, observations over the true state of the environment. Especially, we aim to answer:

How do we maximize the entropy of the true states with a policy conditioned on observations only?

First, we model this setting through a Partially Observable

[1]Politecnico di Milano, Milan, Italy. [2]Technion – Israel Institute of Technology, Haifa, Israel. Correspondence to: Riccardo Zamboni <riccardo.zamboni@polimi.it>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

MDP (POMDP, Åström, 1965) formalism, in which the agent's observations are generated from an observation matrix conditioned on the true state. We consider two distinct learning settings in which the specification of the POMDP is either known or unknown to the agent, respectively.

The former is motivated by domains in which we can train the agent's policy on a simulator of the environment and then deploy the optimal policy in the real world. Still, the problem is non-trivial, as we have to train a policy taking the input available at deployment. Even solving a known POMDP is an intractable problem (Mundhenk et al., 2000) as it requires exponential time in general. Moreover, the problem is optimized by policies conditioned on the history of observations (e.g., Bertsekas, 1976), which are intractable to store. We sidestep the *memory* complexity by defining a clever policy class in which the action distribution is conditioned on a function of the current *belief* on the state of the environment, which is a popular model of uncertainty in POMDPs (Kaelbling et al., 1998). Then, we overcome the computational complexity by considering a first-order relaxation of the state entropy objective, which we optimize via policy gradient (Williams, 1992; Sutton et al., 1999), a methodology that has been previously considered for state entropy maximization in MDPs (e.g., Mutti et al., 2021).

However, a simulator is not available in all the relevant applications. Can we still learn a reasonable policy in those settings? When the POMDP is unknown, we cannot access the entropy over the true states to compute the gradient, as we only have observations. A naive sidestep is to compute the gradient of the entropy over observations, optimizing an objective function that is called Maximum Observation Entropy (MOE). On a recent study of strengths and limitations of MOE, Zamboni et al. (2024) show that the mismatch between the entropy over observations and true states can be significant in relevant domains (e.g., the rescue operation setting we described above). To overcome this limitation, we can instead compute approximate beliefs from observations (Subramanian et al., 2022) and then optimize the entropy of the states sampled from the beliefs as a *proxy* objective that incorporates all of the information available about the entropy on the true states. We can show that the latter is a better approximation than the trivial entropy of observations in general.

Optimizing the proxy objective still involves a crucial issue: The belief is not completely out of the control of the agent, who has an explicit incentive to take actions that maximize the uncertainty of the belief so that the states sampled from the belief will come with higher entropy. We call the latter the *hallucination* problem, as the agent can hack the objective to make herself/himself believe the entropy on the true states is higher than it actually is. To mitigate this effect, we introduce a *regularization* scheme that penalizes the entropy of the belief so that the agent faces a dueling objective that incentives the entropy of the states sampled from the belief on one side and discourages the agent from pursuing beliefs with higher entropy on the other. Finally, we design a policy gradient method for the *regularized* objective, for which we provide extensive theoretical and empirical corroborations, showing that the resulting performance nearly matches the one of the policy that maximizes the true objective when good approximators of the belief are available.

Contributions. We make the following contributions:

- We provide the first generalization of the state entropy maximization problem to POMDPs (Section 3);
- We provide a family of tractable policy gradient methods that address first-order relaxations of the introduced problem with known (Section 4) or unknown (Section 5) POMDP specification, respectively;
- We provide extensive theoretical characterizations of the approximation gap and optimization landscape of the introduced objectives (Section 4 and 5);
- We provide an experimental campaign to uphold the design of the introduced algorithms in a variety of illustrative POMDPs (Section 6).

2. Preliminaries

In this section, we introduce the notation we will use in the paper and the most relevant background on POMDPs (Section 2.1) and state entropy maximization (Section 2.2).

Notation. Let $\mathcal{A}$ a set of size $|\mathcal{A}|$. We denote the T-times Cartesian product of $\mathcal{A}$ as $\mathcal{A}^T := \times_{t=1}^{T} \mathcal{A}$. The simplex on $\mathcal{A}$ is denoted as $\Delta(\mathcal{A}) := \{p \in [0,1]^{|\mathcal{A}|} \mid \sum_{a \in \mathcal{A}} p(a) = 1\}$ and $\mathcal{U}(\mathcal{A})$ denotes a uniform distribution on $\mathcal{A}$. For distributions p_1, p_2 we denote $d^{\text{TV}}(p_1, p_2)$ their total variation distance. We denote as $\mathbb{A} : \mathcal{A} \to \Delta(\mathcal{B})$ a function from elements of $\mathcal{A}$ to distributions over $\mathcal{B}$. For a function $f : \mathbb{R}^n \to \mathbb{R}$, we denote $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ its gradient. For vectors $v = (v_1, \dots, v_T)$ and $u = (u_1, \dots, u_T)$, we use $\oplus$ to denote the concatenation $v \oplus u = (v_1, u_1, \dots, v_T, u_T)$.

2.1. Partially Observable MDPs

A finite-horizon Partially Observable Markov Decision Process (POMDP, Åström, 1965) is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathbb{P}, \mathbb{O}, T, \mu)$ where $\mathcal{S}$ is a set of states of size $S := |\mathcal{A}|$, $\mathcal{A}$ is a set of actions of size $A := |\mathcal{A}|$, $\mathcal{O}$ is a set of observations of size $O := |\mathcal{O}|$, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition model such that $P(s'|s,a)$ is the probability of reaching s' by taking a in s, $\mathbb{O} : \mathcal{S} \to \Delta(\mathcal{O})$ is an observation function such that $\mathbb{O}(o|s)$ is the probability of observing o in s, $T < \infty$ and $\mu \in \Delta(\mathcal{S})$ are the horizon and the initial state distribution of an episode, respectively.

In a POMDP, the interaction process goes as follows. At the start of an episode, an initial state is drawn $s_1 \sim \mu$. For each $t < T$, the agent receives an observation $o_t \sim \mathbb{O}(\cdot|s_t)$ and plays an action a_t, triggering a transition $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$. When the final state s_T is reached, the agent observes $o_T \sim \mathbb{O}(\cdot|s_T)$ and the episode ends. An episode of interaction returns trajectories over states $\tau_{\mathcal{S}} = (s_1, \dots, s_T) \in \mathcal{T}_{\mathcal{S}} \subseteq \mathcal{S}^T$, actions $\tau_{\mathcal{A}} = (a_1, \dots, a_{T-1}) \in \mathcal{T}_{\mathcal{A}} \subseteq \mathcal{A}^{T-1}$, and observations $\tau_{\mathcal{O}} = (o_1, \dots, o_T) \in \mathcal{T}_{\mathcal{O}} \subseteq \mathcal{O}^T$.

Belief. Crucially, the agent cannot access the true state of the POMDP on which the objective function is usually defined.[1] However, it can infer from the observations it receives what is the probability of the process being in a certain state. The latter probability measure, denoted as $\boldsymbol{b} \in \mathcal{B} \subseteq \Delta(\mathcal{S})$ is called a *belief* (Kaelbling et al., 1998). The belief is updated following the Bayes rule. The prior is typically set as $\boldsymbol{b}_1 = \mathcal{U}(\mathcal{S})$. Then, for each $1 < t \leq T$, the posterior of the belief having taken action $a_{t-1} = a$ and received observation $o_t = o$ is computed as

$$\boldsymbol{b}_t^{ao}(s) = \frac{\mathbb{O}(o|s) \sum_{s' \in \mathcal{S}} \mathbb{P}(s|s', a) \boldsymbol{b}_{t-1}(s')}{\sum_{s' \in \mathcal{S}} \mathbb{O}(o|s') \sum_{s'' \in \mathcal{S}} \mathbb{P}(s'|s'', a) \boldsymbol{b}_{t-1}(s'')}.$$

In this sense, the elements of $\mathcal{B}$ can be seen as *belief states* evolving according to the belief-update operator $\mathbb{T}^{ao} : \mathcal{B} \to \mathcal{B}$ such that $\boldsymbol{b}' = \mathbb{T}^{ao}(\boldsymbol{b})$. In the same way as for states, actions, and observations, an episode of interaction generates a trajectory over beliefs $\tau_{\mathcal{B}} = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_T) \in \mathcal{T}_{\mathcal{B}} \subseteq \mathcal{B}^T$.

Policies. We denote the information available to the agent in a given time step as the *information set* $i \in \mathcal{I}$. A policy $\pi : \mathcal{I} \to \Delta(\mathcal{A}) \in \Pi_{\mathcal{I}}$ describes the action selection strategy of the agent, such that $\pi(a|i)$ denotes the probability of taking a given information i and $\Pi_{\mathcal{I}}$ is the policy space with information $\mathcal{I}$. We will later specify the meaning of $\mathcal{I}$, which will be either $\mathcal{O}$, $\mathcal{T}_{\mathcal{O}}$, or $\mathcal{T}_{\mathcal{B}}$ according to the setting.[2]

Distribution over Trajectories. Interacting with a POMDP with a fixed policy induces a specific probability distribution over the generated trajectories. Since we have several trajectories generated simultaneously, we denote the joint trajectory as $\tau = \tau_{\mathcal{S}} \oplus \tau_{\mathcal{A}} \oplus \tau_{\mathcal{O}} \oplus \tau_{\mathcal{B}}$, which probability of being generated under π is given by $p^\pi(\tau) = \mu(s_1) \prod_{t=1}^{T} \mathbb{O}(o_t|s_t) \pi(a_t|i_t) \mathbb{P}(s_{t+1}|s_t, a_t) \mathbb{T}^{o_t a_t}(\boldsymbol{b}_{t+1}|\boldsymbol{b}_t)$.

Interestingly, the belief state formulation allows to extract from τ *believed trajectories* as well, i.e., trajectories $\tau_{\tilde{\mathcal{S}}} = (\tilde{s}_1, \dots, \tilde{s}_T) \in \mathcal{T}_{\tilde{\mathcal{S}}} \subseteq \mathcal{S}^T$ where the states, called *believed* states, are not the true states of the POMDP but samples from the belief $\tilde{s}_t \sim \boldsymbol{b}_t$, which are generated with probability $p(\tau_{\tilde{\mathcal{S}}}|\tau_{\mathcal{B}}) = \prod_{t=1}^{T} \boldsymbol{b}_t(\tilde{s}_t)$.

Distribution over States. A trajectory $\tau_{\mathcal{S}}$ obtained from an interaction episode induces an empirical distribution over true states $d(\tau_{\mathcal{S}}) = (d_{s_1}(\tau_{\mathcal{S}}), \dots, d_{s_S}(\tau_{\mathcal{S}}))$ such that $d_{s_i}(\tau_{\mathcal{S}}) = \frac{1}{T} \sum_{s_t \in \tau_{\mathcal{S}}} \mathbb{1}\{s_t = s_i\}$. This concept can be generalized to any finite set as well, leading to distributions over observations, belief states, and believed states.

2.2. State Entropy Maximization

A POMDP such that $\mathcal{O} = \mathcal{S}$ and $\mathbb{O}(s|s) = 1$ reduces to a finite-horizon Markov Decision Process (MDP, Puterman, 2014) $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbb{P}, T, \mu)$. The true state of the system is fully observable in MDPs, which means the agent can take actions according to a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$.

In the absence of a reward to be maximized, Hazan et al. (2019) proposed a *Maximum State Entropy* (MSE) objective

$$\max_{\pi \in \Pi_{\mathcal{S}}} \left\{ H(d^\pi) := -\sum_{s \in \mathcal{S}} d^\pi(s) \log(d^\pi(s)) \right\} \tag{1}$$

where $d^\pi := \mathbb{E}_{\tau_{\mathcal{S}} \sim p^\pi}[d(\tau_{\mathcal{S}})]$ is the expected state distribution and $H(d^\pi)$ its entropy. The latter objective is known to be *non-concave* w.r.t. the policy yet to admit a dual formulation that is concave w.r.t. the state distribution (Hazan et al., 2019), which allows for efficient computation of an optimal (stochastic in general) Markovian policy.

Mutti et al. (2022a) later formulated a *single-trajectory* version of the state entropy maximization objective,

$$\max_{\pi \in \Pi_{\mathcal{I}}} \left\{ J^{\mathcal{S}}(\pi) := \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^\pi}[H(d(\tau_{\mathcal{S}}))] \right\} \tag{2}$$

in which the agent seeks to maximize the entropy of the empirical state distribution induced in one single trajectory rather than in multiple trajectories (as in Eq. 1). The optimal policy in Eq. 2 is known to be deterministic non-Markovian ($\Pi_{\mathcal{I}} \subseteq \Pi_{\mathcal{T}_{\mathcal{S}}}$) in general, which makes the optimization problem computationally hard (Mutti et al., 2022a).

3. Problem Formulation

State entropy maximization is particularly challenging in POMDPs as the objective function is defined on a space to which the agent has no direct access. It is clear that the ideal goal of maximizing the MSE objective in Eq. 2 as in (fully observable) MDPs is far-fetched under these premises. Addressing MSE in POMDPs includes the following additional and intertwined challenges: **(a)** Defining a proxy objective function compatible with the setting, i.e., on quantities the agent can observe; **(b)** Defining a compact policy class such that policies can be efficiently stored.

In this paper, we will build upon the single-trajectory formulation of the MSE objective, which is closer to the need for

[1] Most of the previous literature in POMDPs define the objective through the maximization of a reward $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Here, we address a different objective that we will formalize in Section 3.

[2] Introducing the information set allows us to work with stationary Markovian policies on the information, which can be non-Markovian policies w.r.t. states or observations.

practical applications (Mutti et al., 2022a). We notice that the common infinite-trajectories relaxation considered in previous works (e.g., Hazan et al., 2019) is still intractable in POMDPs, which leaves minimal benefit over the single-trajectory formulation (see Appendix A.1 for details).

(a) Proxy Objective Functions. Optimizing Eq. 2 is ill-posed in POMDPs without further assumptions because states are not observed. We then seek to design proxy objectives whose maximization leads to policies with good performance on the (ideal) original MSE objective as well. The first and most intuitive choice is to formulate an analogous objective over observations instead of states. The (single-trajectory) *Maximum Observation Entropy* (MOE) is

$$\max_{\pi \in \Pi_{\mathcal{I}}} \left\{ J^{\mathcal{O}}(\pi) := \mathop{\mathbb{E}}_{\tau_{\mathcal{O}} \sim p^{\pi}} [H(d(\tau_{\mathcal{O}}))] \right\} \tag{3}$$

While being rather intuitive, this objective is intrinsically problematic. There can be significant mismatches between observation and state spaces. When the POMDPs are under (respectively over) complete (Liu et al., 2022), i.e., when the number of observations is less (respectively more) than the number of states, it may be hard to link entropy over observations to entropy over states. Moreover, even when $\mathcal{O} = \mathcal{S}$, a random emission function $\mathbb{O}$ could jeopardize any estimate of the state entropy that is based on the entropy of observations. A formal study of the limitations of MOE has been provided in (Zamboni et al., 2024), which characterizes the settings where the entropy of observation is not enough. To address the latter settings, we introduce more reliable proxy objectives in Section 4, 5 along with corresponding assumptions on the information available to the agent.

(b) Deployable Policy Classes. So far, we denoted the policy class as $\Pi_{\mathcal{I}}$ for a generic set $\mathcal{I}$ of the available information. An essential point to be addressed in POMDPs is which policy class to use (Cassandra, 1998). We say a policy class is *deployable* if its policies are conditioned on the information set $\mathcal{I}$ that is available to the agent *at deployment*.[3] We follow a similar definition of deployable policies as for centralized training and decentralized executions in multi-agent settings (Albrecht et al., 2024). It follows that any policy class over true states is not deployable, and this is the case for deterministic non-Markovian policies as well (Mutti et al., 2022a). Yet, other policy classes are deployable, e.g., over observations, trajectories of observations, and trajectories of beliefs. Ideally, we want to employ the richer deployable policy class, which is the space of non-Markovian policies over observations (or, equivalently, over beliefs). Unfortunately, a policy in this class cannot be efficiently stored in general, so we will look for restricted classes with more reasonable memory requirements.

[3]Even in the case a simulator is available to optimize the policy, we still want to deploy the policy in unknown partially observable environments in general.

4. MSE with a POMDP Simulator

First, we consider a simplified setting where:

Assumption 4.1. $\mathbb{P}, \mathbb{O}$ are fully known in training.

This setting encompasses the best-case scenario, in which a (white-box) simulator of the environment is available and the true state of the POMDP can be accessed. Even in this simplified setting, the problem is non-trivial. First, it does not reduce to the MDP problem, as we need to learn a deployable policy. Secondly, the best deployable policy class is problematic in terms of memory complexity. Finally, as the theory demonstrates (Papadimitriou & Tsitsiklis, 1987; Mundhenk et al., 2000), even solving a known POMDP is computationally intractable. These issues drive the algorithmic choices in the following sense:

1. **Memory complexity.** The policy class will be restricted to memory-efficient policies, such that the policy parameters are polynomial in the size of $\mathcal{M}$.
2. **Computational complexity.** A first-order method will be employed, i.e., policy gradient (Williams, 1992; Sutton et al., 1999), to overcome computational hardness.

(1) Unfortunately, the size of $\mathcal{T}_{\mathcal{O}}, \mathcal{T}_{\mathcal{B}}$ is exponential in T, which means that policies over such spaces would require an exponential number of parameters. This leaves the information sets $\mathcal{O}, \mathcal{B}$ as viable options. Similarly, the set $\mathcal{B}$ of belief states reachable in T steps can be extremely large even in simple POMDPs.[4] Policy classes that are efficient to store are $\Pi_{\mathcal{O}}$ and $\Pi_{\tilde{\mathcal{S}}}$, i.e., the set of Markovian policies over observations or believed states. It is known, however, that non-Markovian policies are needed to optimize the single-trajectory MSE in general (Mutti et al., 2022a). An option is to consider the belief, which is a function of the trajectory over states and actions, as a succinct representation of the history, and then to employ a careful parametrization of the policy to get memory efficiency. Formally, we introduce the *Belief-Averaged* (BA) policy class as $\bar{\Pi}_{\mathcal{B}} := \{\pi \in \bar{\Pi}_{\mathcal{B}} : \pi_\theta(\cdot|\boldsymbol{b}) = \langle \theta, \boldsymbol{b} \rangle\} \subseteq \Delta(\mathcal{A})$.

(2) The optimization problem over the latter policy class will be addressed via first-order methods (Williams, 1992), in order to overcome computational hardness. Previous works have considered policy gradient for MSE in MDPs (Mutti et al., 2021; Liu & Abbeel, 2021b). Here, we derive a specialized gradient for the POMDP setting.[5]

Theorem 4.2 (Entropy Policy Gradient in POMDPs)**.** *For a policy $\pi_\theta \in \Pi_{\mathcal{I}}$ parametrized by $\theta \in \Theta \subseteq \mathbb{R}^{IA}$, we have*

$$\nabla_\theta J^i(\pi_\theta) = \mathop{\mathbb{E}}_{\tau \sim p^{\pi}} \left[\nabla_\theta \log \pi_\theta(\tau) H(d(\tau_i)) \right] \tag{4}$$

where $\nabla_\theta \log \pi_\theta(\tau) = \sum_t \nabla_\theta \log \pi_\theta(a_t|i_t)$, $i \in \{\mathcal{S}, \mathcal{O}\}$.

[4]We can compute $\mathcal{B}$ by means of Algorithm 2 in Appendix A.2.

[5]The full derivation can be found in Appendix A.3.

Algorithm 1 Reg-PG for MaxEnt POMDPs

Input: learning rate α, initial parameters θ_1, number of episodes K, batch size N, information set $\mathcal{I}$, proxy class $j \in \{\mathcal{S}, \mathcal{O}, \tilde{\mathcal{S}}\}$, regularization parameter ρ
for $k = 1$ **to** K **do**
  Sample N trajectories $\{\tau_j^n \sim p^{\pi_{\theta_k}}\}_{n\in[N]}$
  Compute the feedbacks $\{H(d(\tau_j^n))\}_{n\in[N]}$
  Compute $\{\log \pi(\tau_j^n)\}_{n\in[N]}$
  Perform a gradient step $\theta_{k+1} \leftarrow \theta_k + \frac{\alpha}{N}\sum_n^N \log\pi(\tau_j^n)[H(d(\tau_j^n)) - \rho \sum_t H(b_t^n)]$
end for
Output: the last-iterate policy π_θ^K

Algorithmic Architecture. It can be seen that optimizing for different objectives, the policy gradient differs only on the second term of the product, which we refer to as *feedback*. Thus, we propose a general algorithmic framework, which works for any objective, and mimics the structure of REINFORCE (Williams, 1992). The pseudocode is reported in Algorithm 1.[6] The main loop of the algorithm (**2**-**7**) is composed of the main steps: (**3**) sample N trajectories with the current policy, (**4**) extract the feedbacks coherently to the objective being optimized, (**5**) compute the log-policy term and (**6**) perform a gradient ascent step over the parameters space.

Smoothness of the Optimization Landscape. We can prove that the considered objectives are locally smooth, making first-order approaches of the kind described above well-suited for the problem.[7]

Theorem 4.3 (Local Lipschitz Constants)**.** *Let $\pi_1, \pi_2 \in \Pi_{\mathcal{I}}$, let $\mathcal{T}_i(\pi_1, \pi_2) = \{\tau_i \in \mathcal{T}_i : p^{\pi_1}(\tau_i) > 0 \vee p^{\pi_2}(\tau_i) > 0\}$ be the set of realizable trajectories over $i \in \{\mathcal{S}, \mathcal{O}\}$, and let $\tau_i^\star = \arg\max_{\tau\in\mathcal{T}_i(\pi_1,\pi_2)} H(d(\tau))$. It holds*

$$|J^i(\pi_1) - J^i(\pi_2)| \leq T H(d(\tau_i^\star)) d^{TV}(\pi_1, \pi_2).$$

A global (but looser) upper bound of the Lipschitz constant can be derived as $TH_{\max}$, where $H_{\max}$ is the maximum entropy that can be obtained over the support. These results provide an interesting insight into how (a bound on) the smoothness constant behaves, as both the objectives defined over true states or observations have Lipschitz constants that are not directly dependent on the policies themselves.

5. MSE without a POMDP Simulator

The Assumption 4.1 of having access to the POMDP specification is rather restrictive and arguably unreasonable in domains where a (white-box) simulator is not available. To overcome this assumption, we aim to refine the design of our algorithmic solution to work with quantities related to observations only. Luckily, beliefs can still be computed approximately well without access to the POMDP model. Belief approximation techniques have been extensively studied in the literature (e.g., see (Subramanian et al., 2022) for a summary). Here, we do not delve into the technicalities of the latter, which are out of the scope of this work, and we instead assume to have access to an *approximated* oracle to compute beliefs.

Assumption 5.1. Let $a \in \mathcal{A}$ and $o \in \mathcal{O}$. Given an approximate belief $\hat{\boldsymbol{b}}_t \in \Delta(\mathcal{S})$ of the true belief $\boldsymbol{b}_t$, an *oracle belief approximator* gives $\hat{\boldsymbol{b}}_{t+1}$ such that $\|\mathbb{T}^{ao}(\hat{\boldsymbol{b}}_t) - \hat{\boldsymbol{b}}_{t+1}\|_1 \leq \epsilon$.

With the latter, we can follow Algorithm 1 as is, computing approximate beliefs instead of the true beliefs. Yet, we have to change the feedback as we cannot compute the entropy on the true states. Luckily, the trivial MOE feedback (3) is *not* the only option we have. We can use the approximate beliefs to reconstruct *believed trajectories* over states and then compute the feedbacks as their entropy. We call the latter the *Maximum Believed Entropy* (MBE):

$$\max_{\pi\in\Pi_{\mathcal{I}}} \left\{ \tilde{J}(\pi) := \mathop{\mathbb{E}}_{\tau_{\mathcal{B}}\sim p^\pi} \mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}}\sim p(\cdot|\tau_{\mathcal{B}})} [H(d(\tau_{\tilde{\mathcal{S}}}))] \right\} \tag{5}$$

where the update of the belief in p^π is now given by the approximate belief oracle. Notably, we can extend both Theorem 4.2, 4.3 to the MBE objective.[8]

Theorem 5.2. *For a policy $\pi_\theta \in \Pi_{\mathcal{I}}$ parametrized by $\theta \in \Theta \subseteq \mathbb{R}^{SA}$, we have*

$$\nabla_\theta \tilde{J}(\pi) = \mathop{\mathbb{E}}_{\tau_{\mathcal{B}}\sim p^\pi} \mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}}\sim p(\cdot|\tau_{\mathcal{B}})} \left[\nabla_\theta \log \pi_\theta(\tau_{\tilde{\mathcal{S}}}) [H(d(\tau_{\tilde{\mathcal{S}}}))] \right] \tag{6}$$

where $\nabla_\theta \log \pi_\theta(\tau_{\tilde{\mathcal{S}}})$ are defined as in 4.2. Additionally, let $\mathcal{T}_{\mathcal{B}}(\pi_1, \pi_2) = \{\tau_{\mathcal{B}} \in \mathcal{T}_{\mathcal{B}} : p^{\pi_1}(\tau_{\mathcal{B}}) > 0 \vee p^{\pi_2}(\tau_{\mathcal{B}}) > 0\}$, $\tau_{\mathcal{B}}^\star = \arg\max_{\tau\in\mathcal{T}_{\mathcal{B}}(\pi_1,\pi_2)} \mathbb{E}_{\tau\sim\tau_{\mathcal{B}}} H(d(\tau))$, and $\bar{H}(\tau_{\mathcal{B}}^\star) = \mathbb{E}_{\tau_{\tilde{\mathcal{S}}}\sim\tau_{\mathcal{B}}^\star} H(d(\tau_{\tilde{\mathcal{S}}}))$, we have

$$|\tilde{J}(\pi_1) - \tilde{J}(\pi_2)| \leq T \bar{H}(\tau_{\mathcal{B}}^\star) d^{TV}(\pi_1, \pi_2). \tag{7}$$

[6] Note that the meaning and role of the regularization parameters and corresponding regularization term, color-highlighted in the algorithm, will be clarified in the next section.

[7] The full derivation of the result is in Appendix A.4.

[8] A full derivation can be found in Appendix A.3.

Interestingly, compared to the other results in Theorem 4.3, MBE displays an upper bound of the Lipschitz constant that depends on the policies π_1, π_2 directly (through $\tau^\star_{\mathcal{B}}$). Additionally, $\bar{H}(\tau^\star_{\mathcal{B}})$ consists in the best *expected* believed entropy, which is generally smaller than $H(d(\tau^\star_i)), i \in \{\mathcal{S}, \mathcal{O}\}$ of Theorem 4.3.

Objectives Gaps and Hallucinatory Effect. Without Assumption 4.1, we cannot know the value of the MSE objective anymore. Thus, it is hard to keep track of the mismatch between what the agent expects the (latent) performance to be and what it truly is once it is evaluated on the true states of the environment. However, it is possible to show that the true objective lies in a space explicitly encircled by the proxies. First, we provide the following instrumental definitions:

Definition 5.3. We define $\mathcal{T}_{\mathcal{O}}(\tau_{\mathcal{S}}) = \{\tau_{\mathcal{O}} \in \mathcal{T}_{\mathcal{O}} : H(d(\tau_{\mathcal{O}})) \geq H(d(\tau_{\mathcal{S}}))\}, \mathcal{T}(\tau_{\mathcal{S}}) = \{\tau_{\bar{\mathcal{S}}} \in \mathcal{T}_{\bar{\mathcal{S}}} : H(d(\tau_{\bar{\mathcal{S}}})) \geq H(d(\tau_{\mathcal{S}}))\}$ as the set of trajectories over observations and believed states, respectively, for which their entropy is higher than the entropy of a fixed trajectory over true states. We let $\mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}}) = \sum_{\tau_{\mathcal{O}} \in \mathcal{T}_{\mathcal{O}}(\tau_{\mathcal{S}})} p^\pi(\tau_{\mathcal{O}}|\tau_{\mathcal{S}}), \mathbb{P}(\mathcal{T}|\tau_{\mathcal{B}}) = \sum_{\tau_{\bar{\mathcal{S}}} \in \mathcal{T}(\tau_{\mathcal{S}})} \tau_{\mathcal{B}}(\tau_{\mathcal{S}})$ the cumulative probability of drawing a trajectory form the above sets and $\bar{p}_{\mathcal{S}}(\tau_{\mathcal{S}}) = \mathbb{E}_{\tau_{\mathcal{B}} \sim p^\pi(\cdot|\tau_{\mathcal{S}})} \mathbb{P}(\mathcal{T}|\tau_{\mathcal{B}})$ the expected probability of the believed set. Finally, $J^{\mathcal{O}}(\pi|\tau_{\mathcal{S}}) = \mathbb{E}_{\tau_{\mathcal{O}} \sim p^\pi(\cdot|\tau_{\mathcal{S}})}[H(d(\tau_{\mathcal{O}}))], \tilde{J}(\pi|\tau_{\mathcal{S}}) = \mathbb{E}_{\tau_{\mathcal{B}} \sim p^\pi(\cdot|\tau_{\mathcal{S}})} \mathbb{E}_{\tau_{\bar{\mathcal{S}}} \sim \tau_{\mathcal{B}}}[H(d(\tau_{\bar{\mathcal{S}}}))]$ the MOE (MBE) objective for a fixed trajectory on the states.

Then, the following theorem holds:

Theorem 5.4 (Proxy Gaps)**.** *For a fixed policy $\pi \in \Pi_{\mathcal{I}}$, the MSE objective $J^{\mathcal{S}}(\pi)$ is bounded by the MOE objective according to*

$$J^{\mathcal{S}}(\pi) \leq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim \bar{p}_{\mathcal{S}}} \left[\frac{1}{\mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}})} J^{\mathcal{O}}(\pi|\tau_{\mathcal{S}}) \right]$$

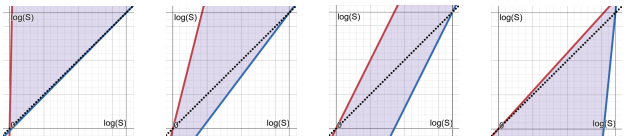
$$J^{\mathcal{S}}(\pi) \geq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim \bar{p}_{\mathcal{S}}} \left[\frac{1}{1 - \mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}})} J^{\mathcal{O}}(\pi|\tau_{\mathcal{S}}) \right] - \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim \bar{p}_{\mathcal{S}}} \left[\frac{\mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}})}{1 - \mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}})} \right] \log O$$

Analogously, $J^{\mathcal{S}}(\pi)$ is bounded by the MBE objective according to

$$J^{\mathcal{S}}(\pi) \leq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim \bar{p}} \left[\frac{1}{\bar{p}_{\mathcal{S}}(\tau_{\mathcal{S}})} \tilde{J}^{\mathcal{S}}(\pi|\tau_s) \right]$$

$$J^{\mathcal{S}}(\pi) \geq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim \bar{p}} \left[\frac{1}{1 - \bar{p}_{\mathcal{S}}(\tau_{\mathcal{S}})} \tilde{J}^{\mathcal{S}}(\pi|\tau_{\mathcal{S}}) \right] - \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim \bar{p}} \left[\frac{\bar{p}_{\mathcal{S}}(\tau_{\mathcal{S}})}{1 - \bar{p}_{\mathcal{S}}(\tau_{\mathcal{S}})} \right] \log S$$

These results show that the true objective (MSE) is upper/lower bounded by the proxies depending on the probability to generate trajectories (over observations or believed states, respectively) with entropy higher than the one of the trajectory that generated them. We refer to this probability as *hallucination probability* and to the resulting phenomenon as **hallucinatory effect**. We show in Figure 1 a visual representation of the MBE gaps in Theorem 5.4. It is evident that for low over-estimation probabilities ($\bar{p}_{\mathcal{S}} = 0.02$), the MBE objective is a good lower bound for the MSE objective,[9] while it is less so as the hallucination probability increases. The full derivation of these results can be found in Appendix A.5.



(a) $\bar{p} = 0.02$ (b) $\bar{p} = 0.25$ (c) $\bar{p} = 0.5$ (d) $\bar{p} = 0.9$

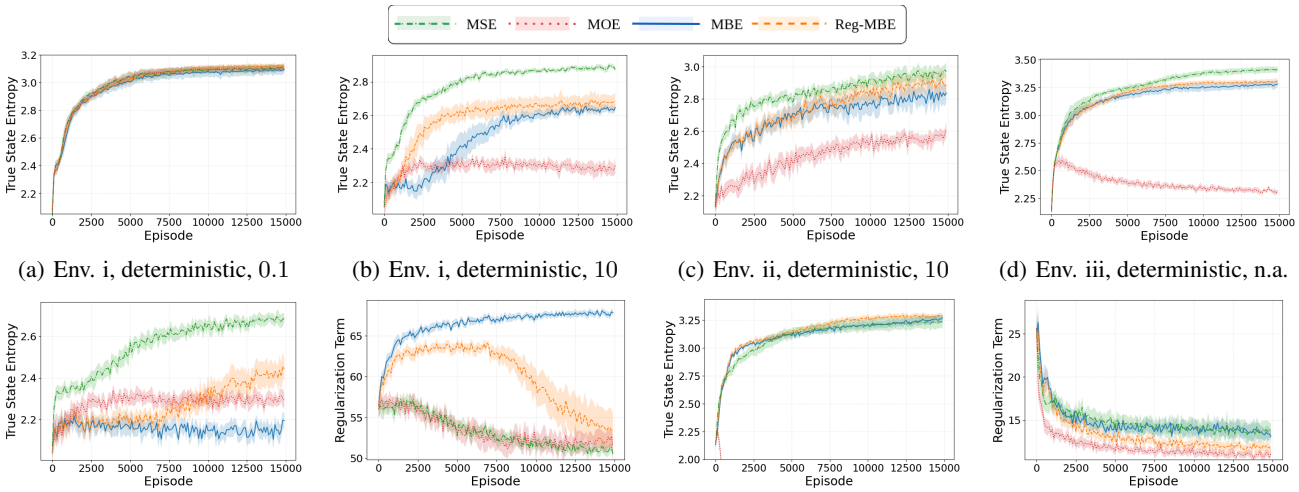
Figure 1. MBE Proxy gaps: for different hallucination probabilities $\bar{p}_{\mathcal{S}}$ and a fixed trajectory τ_s, the y-axis represents the **possible MSE values** contained between the **upper bound** and **lower bound** as $\tilde{\mathcal{J}}^{\mathcal{S}}(\pi|\tau_s)$ varies between 0 and the maximum value $\log(S)$ (the corresponding **MBE values** are plotted over the diagonal to allow a comparisons).

The role of hallucinatory effects is crucial. Indeed, when the effect of hallucinations is negligible, the proxy objectives are reasonable lower bounds to the true MSE objective, and optimizing them guarantees at least a non-degradation of the MSE objective. The hallucinatory effect, i.e., generating over-entropic trajectories due to the randomness of the generating process, on either observations or beliefs, can be controlled by reducing the randomness of the generating process itself. Unfortunately, under Assumption 5.1, we cannot control the observation model as done in Zamboni et al. (2024). However, we have partial control over the trajectory of beliefs that are generated, as they are (approximately) learned and the belief update is conditioned on the taken action. Thus, we can follow the same rationale and derive a regularized objective built upon $\tilde{J}(\pi)$. In particular, we can maintain a valid lower bound to the MBE objective while enforcing the generation of a sequence of low-entropy belief states $\tau_{\mathcal{B}} = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_T)$ with the following:

$$\tilde{J}(\pi) \geq \tilde{J}(\pi) - \rho \mathop{\mathbb{E}}_{\tau_{\mathcal{B}} \sim p^\pi}[H(\tau_{\mathcal{B}})] \geq \tilde{J}(\pi) - \rho \mathop{\mathbb{E}}_{\tau_{\mathcal{B}} \sim p^\pi} \left[\sum\nolimits_t H(\boldsymbol{b}_t) \right] =: \tilde{J}_\rho(\pi)$$

where the second inequality is due to the sub-additivity of the entropy. We call the obtained $\tilde{J}_\rho(\pi)$ *MBE with belief regularization* (Reg-MBE for short). Then, the policy gradient

[9] One may notice that the MOE gap is potentially looser: In many scenarios $\log(O) \gg \log(S)$ while on the other hand $\bar{p}_{\mathcal{S}}(\tau_{\mathcal{S}})$ is the result of an additional expectation with respect to $\mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}})$.



(a) Env. i, deterministic, 0.1 (b) Env. i, deterministic, 10 (c) Env. ii, deterministic, 10 (d) Env. iii, deterministic, n.a.

(e) Env. i, stochastic, 10 (f) Env. i, stochastic, 10 (g) Env. iv, deterministic, n.a. (h) Env. iv, deterministic, n.a.

Figure 2. True state entropy (or *regularization term*) obtained by Algorithm 1 specialized for the feedbacks *MSE, MOE, MBE, MBE with belief regularization* (Reg-MBE). For each plot, we report a tuple (environment, transition noise, observation variance) where the latter is *not available* (n.a.) when observations are deterministic. For each curve, we report the average and 95% c.i. over 16 runs.

$\nabla_\pi \tilde{J}_\rho(\pi)$ for this objective is

$$\nabla_\pi \tilde{J}(\pi) - \rho \mathop{\mathbb{E}}_{\tau_\mathcal{B} \sim p^\pi} \left[\nabla_\theta \log \pi_\theta(\tau_\mathcal{B}) \sum\nolimits_t H(\boldsymbol{b}_t) \right].$$

It is easy to see that whenever $\tilde{J}(\pi)$ is a good proxy (i.e., a tight lower bound) of the true MSE objective, then the regularized objective $\tilde{J}_\rho(\pi)$ will be a reasonable lower bound as well. Most importantly, the regularization term incentives lower-entropy beliefs, which keeps $\tilde{J}(\pi)$ in a region where it approximates MSE well. From these considerations, a *belief-regularized* version of the Algorithm 1 is proposed by simply modifying how the gradient step in (**6**) is computed, as can be seen in the regularized version of Algorithm 1.

6. Numerical Experiments

In this section, we provide an empirical corroboration of the proposed methods and reported claims (results reported in Figure 2 and 3). The section is organized as follows:

- 6.1 We describe the experimental set-up;
- 6.2 We compare the performance driven by the proxy objectives (MOE, MBE, MBE with belief regularization) against the ideal objective (MSE);
- 6.3 We study the impact of belief approximation on MBE-based algorithms (with and without regularization).

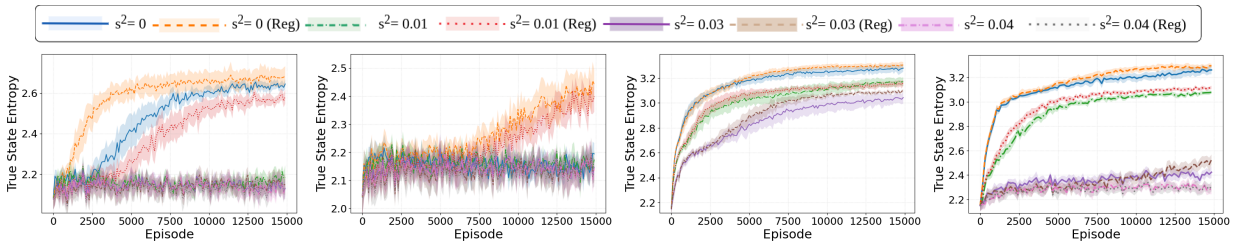
6.1. Experimental Set-Up

We consider the following set of finite domains:

- **(i)** A 5×5-Gridworld with a single room, where $\mathcal{O} = \mathcal{S}$ and the emission matrix $\mathbb{O}$ is such that every row is a (discretized) Gaussian $\mathbb{O}(o|s) = \mathcal{N}(s, \sigma^2)$;
- **(ii)** A 6×6-Gridworld with 4 identical rooms, where $\mathcal{O} = \mathcal{S}$ and the emission matrix $\mathbb{O}$ is such that every row is a (discretized) Gaussian $\mathbb{O}(o|s) = \mathcal{N}(s, \sigma^2)$;
- **(iii)** A 6×6-Gridworld with 4 identical rooms, where $\mathcal{O} = \{1, 2, 3, 4\}$ and the deterministic emission matrix $\mathbb{O}$ such that for every state $\mathbb{O}(s)$ is the id of the room in which the state lies;
- **(iv)** A 6×6-Gridworld with 4 identical rooms, where $\mathcal{O} = \{1, 2\}$ and the deterministic emission matrix $\mathbb{O}$ such that for every state $\mathbb{O}(s)$ is the side of the grid (left rooms or right rooms) the state lies in.

In all the environments described above, the agent has four actions to take, one for moving to the adjacent grid cell in each of the coordinate directions. Moving against a wall undoes the effect of an action. When we say an environment is *deterministic* we mean that the agent actions never fail. In a *stochastic* environment each action has 0.1 failure probability, which has the equivalent effect of taking one of the other three actions at random. Finally, we compare the algorithms designed for the MSE, MOE, MBE objectives presented in previous sections.[10] Irrespective of the optimized objective, their performance is evaluated on the **true state entropy** (Equation 2), which is the ultimate target of state entropy maximization in POMDPs. All of the algorithms optimize a policy within the BA class $\bar{\Pi}_\mathcal{B}$. A visualization of the described environment is provided in Appendix B.3, while the choice of the experimental parameters is discussed in Appendix B.4. Appendix B.5 provides a finer analysis of the choice of the policy class.

[10] While we only compare algorithms of our design, we note that we could not find any previous algorithm addressing state entropy maximization in POMDPs.



(a) Env. i, deterministic, 10 (b) Env. i, stochastic, 10 (c) Env. iii , deterministic, n.a. (d) Env. iv, deterministic, n.a.

Figure 3. True state entropy obtained by Algorithm 1 with *MBE, MBE with belief regularization* (MBE with Reg) feedbacks under different levels of approximation noise s^2. For each plot, we report a tuple (environment, transition noise, observation variance) where the latter is *not available* (n.a.) when observations are deterministic. For each curve, we report the average and 95% c.i. over 16 runs.

6.2. MSE in POMDP with the Proxy Objectives

In this section, we compare the performance obtained by Algorithm 1 specialized for the different proxy objectives. For the sake of clarity, here we assume the belief updates to be computed exactly, while we study the impact of the belief approximation in the next section.

Figure 2(a) shows that all of the objectives works equally well in easy settings, e.g., deterministic transitions and small observation noise. However, major differences arise when considering harder settings. The MOE objective is sensitive to the quality of the observations, which is evident from the performance degradation in Figures 2(b), 2(c), 2(d). Instead, MBE objectives are remarkably robust to their (diminishing) quality. MBE with belief regularization (Reg-MBE) always performed better than the non-regularized version, showing faster convergence or better final performance.

In Figure 2(e), we see that stochastic transitions are also arduous for MOE and MBE. MBE with belief regularization proved to be better. Interestingly, the true state entropy improvement happens concurrently with the optimization of the regularization term (Figure 2(f)).

Unsurprisingly, optimizing the MSE objective leads to the best performance in most cases, as a testament that whenever the POMDP specification is available in simulation, it is worth training the policy we seek to deploy on the true state entropy. Interestingly, in some limit cases with extreme disentanglement between the observations and the true MSE objective (Figure 2(g)), the belief-regularized MBE proxy performed slightly better.

Finally, Figure 2(f) shows how the MBE is severely hallucinated, an effect that is mitigated with belief regularization.

6.3. The Impact of Belief Approximation

In the previous section, we compared the algorithms in an ideal setting in which the belief is approximated exactly. Here we instead consider the effect of the belief approximation on the same experiments. Especially, to keep full generality of our results, we perturb the exact beliefs with an entry-wise Gaussian noise (with variance $s^2 = \{0, 0.01, 0.03, 0.04\}$ respectively), so that our results do not apply to a single belief approximator but any approximator with a bounded error.[11]

All Figures from 3(a) to 3(d) provide two important evidences. First, when good belief approximators are available, the resulting performance is strikingly similar to the ideal setting with exact beliefs. Secondly, MBE with belief regularization is significantly more robust to perturbations, hinting that mitigating hallucination also alleviates the impact of the approximation error to some extent.

7. Related Work

Below, we summarize the most relevant work on POMDPs, state entropy maximization, and policy optimization.

POMDPs. Learning and planning problems in POMDPs have been extensively studied. In the most general formulation, POMDPs have been shown to be computationally and statistically intractable (Papadimitriou & Tsitsiklis, 1987; Krishnamurthy et al., 2016). Nonetheless, several recent works have analyzed tractable sub-classes of POMDPs under convenient structural assumptions (to name a few Jin et al., 2020a; Golowich et al., 2022; Chen et al., 2022; Liu et al., 2022; Zhan et al., 2023; Zhong et al., 2023). Strides have also been made in modeling beliefs as approximate information states (Subramanian et al., 2022) and in the design of practical algorithms (Hafner et al., 2019).

State Entropy Maximization. State entropy maximization in MDPs has been introduced in Hazan et al. (2019) and then considered in a flurry of subsequent works (Lee et al., 2019; Mutti & Restelli, 2020; Mutti et al., 2021; 2022a;b; Mutti, 2023; Zhang et al., 2021; Guo et al., 2021; Liu & Abbeel, 2021b;a; Seo et al., 2021; Yarats et al., 2021; Nedergaard & Cook, 2022; Yang & Spaan, 2023; Tiapkin et al., 2023; Jain et al., 2023; Kim et al., 2023; Zisselman et al., 2023) addressing the problem from various angles. While Savas

[11] For the sake of clarity, here we report the variance of the perturbation instead of the approximation as in Assumption 5.1.

et al. (2022) study the problem of maximizing the entropy over trajectories induced in a POMDP, we are the first to formulate *state* entropy maximization in the latter setting. Complementary results on when the observation entropy is a sensible target for state entropy maximization in POMDPs are reported in the concurrent work (Zamboni et al., 2024).

Policy Optimization. The use of first-order methods (Sutton et al., 1999; Peters & Schaal, 2008) to address non-concave policy optimization is not new in RL. We considered a *vanilla* policy gradient estimator (Williams, 1992) but several refinements could be made, such as natural gradient (Kakade, 2001), trust-region schemes (Schulman et al., 2015), and importance sampling (Metelli et al., 2018).

8. Conclusion

In this paper, we generalize the state entropy maximization problem in POMDPs. Especially, we aim to learn a policy that maximizes the entropy over the true states of the environment while accessing partial observations only. In the paper, we show that this entails several critical challenges. First, we propose a family of proxy objectives to approximate the ideal (but not accessible) original objective through quantities that are available to the agent. Then, we choose a convenient sub-class of non-Markovian policies that retain compressed information of history without incurring unreasonable memory requirements. Finally, we designe practical first-order algorithms, which are based on policy gradient, to overcome the inherent non-convexity of the considered objective functions.

Future works can extend our results in many directions, which include integrating a belief approximation method into the algorithmic pipeline (e.g., Zintgraf et al., 2019; Subramanian et al., 2022), designing practical implementations for continuous domains (e.g., Liu & Abbeel, 2021b), and investigating more policy classes with succinct representations of the history beyond the one we considered.

We believe that our work can be a crucial first step in the direction of extending state entropy maximization to yet more practical settings, in which the state of the system is often not fully observed.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Albrecht, S. V., Christianos, F., and Schäfer, L. *Multi-agent reinforcement learning: Foundations and modern approaches*. MIT Press, 2024.

Åström, K. J. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

Bertsekas, D. *Reinforcement learning and optimal control*. Athena Scientific, 2019.

Bertsekas, D. P. *Dynamic programming and stochastic control*. Academic Press, Inc., 1976.

Cassandra, A. R. *Exact and approximate algorithms for partially observable Markov decision processes*. PhD Thesis, Brown University, 1998.

Chen, F., Bai, Y., and Mei, S. Partially observable rl with b-stability: Unified structural condition and sharp sample-efficient algorithms. *arXiv preprint arXiv:2209.14990*, 2022.

Golowich, N., Moitra, A., and Rohatgi, D. Planning in observable pomdps in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 2022.

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.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2019.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2019.

Jain, A. K., Lehnert, L., Rish, I., and Berseth, G. Maximum state entropy exploration using predecessor and successor representations. In *Advances in Neural Information Processing Systems*, 2023.

Jin, C., Kakade, S., Krishnamurthy, A., and Liu, Q. Sample-efficient reinforcement learning of undercomplete pomdps. In *Advances in Neural Information Processing Systems*, 2020a.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, 2020b.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Kakade, S. M. A natural policy gradient. In *Advances in Neural Information Processing Systems*, 2001.

Kim, D., Shin, J., Abbeel, P., and Seo, Y. Accelerating reinforcement learning with value-conditional state entropy exploration. In *Advances in Neural Information Processing Systems*, 2023.

Krishnamurthy, A., Agarwal, A., and Langford, J. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, 2016.

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 *International Conference on Machine Learning*, 2021a.

Liu, H. and Abbeel, P. Behavior from the void: Unsupervised active pre-training. In *Advances in Neural Information Processing Systems*, 2021b.

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, 2022.

Metelli, A. M., Papini, M., Faccio, F., and Restelli, M. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems*, 2018.

Mundhenk, M., Goldsmith, J., Lusena, C., and Allender, E. Complexity of finite-horizon Markov decision process problems. *Journal of the ACM*, 47(4):681–720, 2000.

Mutti, M. *Unsupervised reinforcement learning via state entropy maximization*. PhD Thesis, Università di Bologna, 2023.

Mutti, M. and Restelli, M. An intrinsically-motivated approach for learning highly exploring and fast mixing policies. In *AAAI Conference on Artificial Intelligence*, 2020.

Mutti, M., Pratissoli, L., and Restelli, M. Task-agnostic exploration via policy gradient of a non-parametric state entropy estimate. In *AAAI Conference on Artificial Intelligence*, 2021.

Mutti, M., De Santi, R., and Restelli, M. The importance of non-Markovianity in maximum state entropy exploration. In *International Conference on Machine Learning*, 2022a.

Mutti, M., Mancassola, M., and Restelli, M. Unsupervised reinforcement learning in multiple environments. In *AAAI Conference on Artificial Intelligence*, 2022b.

Nedergaard, A. and Cook, M. k-means maximum entropy exploration. *arXiv preprint arXiv:2205.15623*, 2022.

Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 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.

Savas, Y., Hibbard, M., Wu, B., Tanaka, T., and Topcu, U. Entropy maximization for partially observable Markov decision processes. *IEEE Transactions on Automatic Control*, 67(12):6948–6955, 2022.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

Seo, Y., Chen, L., Shin, J., Lee, H., Abbeel, P., and Lee, K. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*, 2021.

Subramanian, J., Sinha, A., Seraj, R., and Mahajan, A. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *The Journal of Machine Learning Research*, 23(1):483–565, 2022.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 1999.

Tarbouriech, J. and Lazaric, A. Active exploration in Markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Tarbouriech, J., Shekhar, S., Pirotta, M., Ghavamzadeh, M., and Lazaric, A. Active model estimation in Markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Tiapkin, D., Belomestny, D., Calandriello, D., Moulines, E., Munos, R., Naumov, A., Perrault, P., Tang, Y., Valko, M., and Menard, P. Fast rates for maximum entropy exploration. In *International Conference on Machine Learning*, 2023.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Yang, Q. and Spaan, M. T. CEM: Constrained entropy maximization for task-agnostic safe exploration. In *AAAI Conference on Artificial Intelligence*, 2023.

Yarats, D., Fergus, R., Lazaric, A., and Pinto, L. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, 2021.

Yarats, D., Brandfonbrener, D., Liu, H., Laskin, M., Abbeel, P., Lazaric, A., and Pinto, L. Don't change the algorithm, change the data: Exploratory data for offline reinforcement learning. *arXiv preprint arXiv:2201.13425*, 2022.

Zamboni, R., Cirino, D., Restelli, M., and Mutti, M. The limits of pure exploration in POMDPs: When the observation entropy is enough. *Reinforcement Learning Journal*, 2024.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. Pac reinforcement learning for predictive state representations. In *International Conference on Learning Representations*, 2023.

Zhang, C., Cai, Y., Huang, L., and Li, J. Exploration by maximizing Rényi entropy for reward-free rl framework. In *AAAI Conference on Artificial Intelligence*, 2021.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*, 2023.

Zintgraf, L., Shiarlis, K., Igl, M., Schulze, S., Gal, Y., Hofmann, K., and Whiteson, S. Varibad: A very good method for Bayes-adaptive deep rl via meta-learning. In *International Conference on Learning Representations*, 2019.

Zisselman, E., Lavie, I., Soudry, D., and Tamar, A. Explore to generalize in zero-shot rl. In *Advances in Neural Information Processing Systems*, 2023.

A. Proofs and Additional Material

A.1. Infinite-trajectories Formulations

As for MDP theory, the Maximum State Entropy proxies can be formulated in an infinite trajectories form, namely as for Equation (1) it is possible to write the infinite trajectories formulation of the MOE (MBE) as

$$\max_{\pi\in\Pi_{\mathcal{I}}} \left\{ J^{\mathcal{O}}_{\infty} = H(d^{\pi}_{\mathcal{O}}) := -\sum_{o\in\mathcal{O}} d^{\pi}_{\mathcal{O}}(o)\log(d^{\pi}_{\mathcal{O}}(o)) \right\}$$
$$\max_{\pi\in\Pi_{\mathcal{I}}} \left\{ \tilde{J}_{\infty} = H(d^{\pi}_{\mathcal{S}}) := -\sum_{s\in\mathcal{S}} d^{\pi}_{\mathcal{S}}(s)\log(d^{\pi}_{\mathcal{S}}(s)) \right\}$$

where $d^{\pi}_{\mathcal{O}} := \mathbb{E}_{\tau_{\mathcal{O}}\sim p^{\pi}}[d(\tau_{\mathcal{O}})]$ and $d^{\pi}_{\mathcal{S}} := \mathbb{E}_{\tau_{\mathcal{B}}\sim p^{\pi}}\mathbb{E}_{\tau_{\tilde{\mathcal{S}}}\sim p(\cdot|\tau_{\mathcal{B}})}[d(\tau_{\tilde{\mathcal{S}}})]$ is the expected observation (believed state) distribution. These objectives are linked to the single trajectory ones through Jensen's Inequality due to the concavity of the entropy function, namely

$$J^{\mathcal{O}} = \mathop{\mathbb{E}}_{\tau_{\mathcal{O}}\sim p^{\pi}}[H(d(\tau_{\mathcal{O}})] \leq H(\mathop{\mathbb{E}}_{\tau_{\mathcal{O}}\sim p^{\pi}} d(\tau_{\mathcal{O}})) = H(d^{\pi}_{\mathcal{O}}) = J^{\mathcal{O}}_{\infty}$$
$$\tilde{J} = \mathop{\mathbb{E}}_{\tau_{\mathcal{B}}\sim p^{\pi}}\mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}}\sim p(\cdot|\tau_{\mathcal{B}})}[H(d(\tau_{\tilde{\mathcal{S}}})] \leq H(\mathop{\mathbb{E}}_{\tau_{\mathcal{B}}\sim p^{\pi}}\mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}}\sim p(\cdot|\tau_{\mathcal{B}})}[d(\tau_{\tilde{\mathcal{S}}})]) = H(d^{\pi}_{\mathcal{S}}) = \tilde{J}_{\infty}$$

Interestingly, the MBE objective has a clean and neat equivalent formulation in belief-state POMDPs that can be turned into a dual problem as for MDPs, yet the resulting problem is still intractable. More specifically, having defined belief states, we can encode the POMDP $\mathcal{M}$ into a corresponding *belief* MDP $\mathcal{M}_{\mathcal{B}} := (\mathcal{B}, \mathcal{A}, \widetilde{\mathbb{P}}, \mathbb{B}, \boldsymbol{b}_0, T)$ where

- $\mathcal{B}$ is a finite set of states such that each $\boldsymbol{b}\in\mathcal{B}$ corresponds to a belief state, and $\mathcal{B}$ is obtained by running Algorithm 2 in $\mathcal{M}$;
- $\mathcal{A}$ is the set of actions in $\mathcal{M}$;
- $\widetilde{\mathbb{P}} : \mathcal{B}\times\mathcal{A}\to\Delta_{\mathcal{B}}$ is the transition model of the belief MDP defined in a few lines;
- $\boldsymbol{b}_0\in\mathcal{B}$ is the initial state;
- T is the horizon length.

To fully characterize $\mathcal{M}_{\mathcal{B}}$, we can extract the transition model $\widetilde{\mathbb{P}}$ from $\mathcal{M}$ as

$$\begin{aligned}\widetilde{\mathbb{P}}(\boldsymbol{b}'|\boldsymbol{b},a) &= \sum_{\{o\in\mathcal{O}|\boldsymbol{b}'=\mathbb{T}^{ao}(\boldsymbol{b})\}} P(o|\boldsymbol{b},a) = \sum_{\{o\in\mathcal{O}|\boldsymbol{b}'=\mathbb{T}^{ao}(\boldsymbol{b})\}}\sum_{s\in\mathcal{S}} P(o|s)P(s|\boldsymbol{b},a)\\ &= \sum_{\{o\in\mathcal{O}|\boldsymbol{b}'=\mathbb{T}^{ao}(\boldsymbol{b})\}}\sum_{s\in\mathcal{S}}\mathbb{O}(o|s)\sum_{s'\in\mathcal{S}}\boldsymbol{b}(s')\mathbb{P}(s|s',a).\end{aligned}$$

Let us denote as $d^{\pi}\in\Delta_{\mathcal{S}}$ the expected finite-horizon state distribution induced by a policy $\pi\in\Pi_{\mathcal{I}}$ on the true (unobserved) states. Then, we can define the objective function of our problem as

$$\max_{\pi\in\Pi_{\mathcal{I}}} H(d^{\pi}) = \min_{\pi\in\Pi_{\mathcal{I}}}\mathop{\mathbb{E}}_{s\sim d^{\pi}}[\log d^{\pi}(s)] \tag{8}$$

Following standard techniques for MDPs (Puterman, 2014), we can obtain the optimal planning policy for (8) by solving the dual convex program

$$\begin{aligned}&\underset{\substack{\boldsymbol{d}\in\Delta_{\mathcal{S}}\\ \{\boldsymbol{\omega_t}\in\Delta_{\mathcal{B}\times\mathcal{A}}\}_{t\in[1:T]}}}{\text{maximize}} \quad H(d)\\ &\text{subject to} \sum_{a'\in\mathcal{A}}\omega_{t+1}(\boldsymbol{b}',a') = \sum_{\boldsymbol{b}\in\mathcal{B},a\in\mathcal{A}}\omega_t(\boldsymbol{b},a)\widetilde{\mathbb{P}}(\boldsymbol{b}'|\boldsymbol{b},a) \qquad \forall\boldsymbol{b}'\in\mathcal{B},\ \forall t\in 1\dots T\\ &\qquad d(s) = \frac{1}{T}\sum_{t}^{T}\sum_{\boldsymbol{b}\in\mathcal{B},a\in\mathcal{A}}\omega_t(\boldsymbol{b},a)\boldsymbol{b}(s) \qquad \forall(s,a)\in\mathcal{S}\times\mathcal{A}\end{aligned}$$

and then obtaining the resulting (non-stationary) policy from the solution $\boldsymbol{\omega}^*$ as $\pi_t(a|\boldsymbol{b}) = \omega_t^*(\boldsymbol{b},a)/\sum_{a'\in\mathcal{A}}\omega_t^*(\boldsymbol{b},a'), \forall(\boldsymbol{b},a) \in \mathcal{B}\times\mathcal{A}$. As one may notice, while this problem has a neat and concise formulation, the dimensionality of the optimization problem does not scale with the dimension of $\mathcal{M}$.

A.2. Belief Set Computation

The belief states set reachable in a T step interaction with a POMDP can be computed via the following Algorithm

Algorithm 2 $Belief\_set(\boldsymbol{b}, \mathcal{B}, t, T)$

Input: belief $\boldsymbol{b}$, set $\mathcal{B}$, step t, horizon T
if $t < T$ **then**
  for $(o,a) \in \mathcal{O}\times\mathcal{A}$ **do**
    $\boldsymbol{b}' = T^{ao}(b)$
    if $\boldsymbol{b}' \notin \mathcal{B}$ **then**
      $\mathcal{B} = Belief\_set(\boldsymbol{b}', \mathcal{B}\cup\{\boldsymbol{b}'\}, t+1, T)$
    end if
  end for
end if
return $\mathcal{B}$

A.3. Proofs of Theorem 4.2 & Theorem 5.2: Policy Gradients Computation

Let us denote $\tau = \tau_{\mathcal{S}} \oplus \tau_{\mathcal{O}} \oplus \tau_{\mathcal{B}} \oplus \tau_{\tilde{\mathcal{S}}}$ and for a generic $i \in \{\mathcal{S}, \mathcal{O}, \mathcal{B}, \tilde{\mathcal{S}}\}$ we denote $\tau|_i$ as the trajectory τ_i extracted from τ. This is done to be able to use any kind of policy class considered in the main paper as well. For a generic single trajectory objective defined with $J \in \{J^{\mathcal{S}}, J^{\mathcal{O}}, \tilde{J}\}$ it is possible to write:

$$\begin{aligned}\nabla_\theta J(\pi) &= \nabla_\theta \mathop{\mathbb{E}}_{\tau\sim p^\pi}[H(d(\tau|_I))] \\ &= \nabla_\theta \sum_\tau p^\pi(\tau) H(d(\tau|_i)) \\ &= \sum_\tau \Big(\nabla_\theta p^\pi(\tau)\Big) H(d(\tau|_i))\end{aligned}$$

Thanks to the usual log-trick

$$\begin{aligned}&= \sum_\tau p^\pi(\tau)\Big(\nabla_\theta \log p^\pi(\tau)\Big) H(d(\tau|_i)) \\ &= \mathop{\mathbb{E}}_{\tau\sim p^\pi}\left[\nabla_\theta \log p^\pi(\tau) H(d(\tau|_i))\right]\end{aligned}$$

The computation of the gradient is then reconducted to the calculation of the log-policy term $\nabla_\theta \log p^\pi(\tau)$ for the generic class $\pi \in \Pi_{\mathcal{I}}$. It follows that

$$\begin{aligned}\nabla_\theta \log p^\pi(\tau) &= \nabla_\theta \log\Big(\mu(s_1)\prod_{t=1}^{T}\mathcal{O}(o_t|s_t)\pi(a_t|i_t)\mathbb{P}(s_{t+1}|s_t,a_t)\mathbb{T}^{o_t a_t}(b_{t+1}|b_t)\Big) \\ &= \nabla_\theta\Big(\log(\mu(s_1)) + \sum_{t=1}^{T}\log(\mathcal{O}(o_t|s_t)) + \log(\pi(a_t|i_t)) + \log(\mathbb{P}(s_{t+1}|s_t,a_t)) + \log(\mathbb{T}^{o_t a_t}(b_{t+1}|b_t))\Big)\end{aligned}$$

Where the only terms depending on θ are indeed the $\mathcal{I}$-specific log-policy terms, leading to

$$\nabla_\theta \log p^\pi(\tau) = \sum_t^T \nabla_\theta \log \pi_\theta(a_t|i_t)$$

which leads to the standard REINFORCE-like formulation of policy gradients.

A.4. Proofs of Theorem 4.3 & Theorem 5.2: Lipschitz Constants Computation

MSE/MOE (Theorem 4.3): Let us define the set of reachable trajectories in T steps by following a generic policy π_i as $T_i = \{\tau \in T_i : p^{\pi_i}(\tau) > 0\}$, it follows that for both MSE and MOE objective, by defining τ as $\tau_{\mathcal{S}}$ or $\tau_{\mathcal{O}}$ respectively, we can see that

$$\begin{aligned}|J(\pi_1) - J(\pi_2)| &= \Big| \mathop{\mathbb{E}}_{\tau\sim p^{\pi_1}}[H(d(\tau))] - \mathop{\mathbb{E}}_{\tau\sim p^{\pi_2}}[H(d(\tau))]\Big| \\ &\leq \sum_{\tau\in T_1\cup T_2} H(d(\tau))\Big|p^{\pi_1}(\tau) - p^{\pi_2}(\tau)\Big|\end{aligned}$$

By defining $\tau^\star \in \arg\max_{\tau\in T_1\cup T_2} H[d(\tau)]$

$$\leq H[d(\tau^\star)] \sum_{\tau\in T_1\cup T_2} \Big|p^{\pi_1}(\tau) - p^{\pi_2}(\tau)\Big|$$

We notice that $p^{\pi_i} = \prod_t^\pi p_t^{\pi_i}$ and that the total variation between two product distributions can be upper-bounder by the summation over the per-step total variations, namely $d^{\text{TV}}(\prod_t^\pi p_t^{\pi_i}, \prod_t^\pi p_t^{\pi_j}) \leq \sum_t^T d^{\text{TV}}(p_t^{\pi_i}, p_t^{\pi_j})$, leading to

$$\begin{aligned}&= H[d(\tau^\star)] d^{\text{TV}}(p^{\pi_1}, p^{\pi_2}) \\ &\leq H[d(\tau^\star)] \sum_t^T d^{\text{TV}}(p_t^{\pi_1}, p_t^{\pi_2})\end{aligned}$$

The only difference between the two distributions (for a fixed step) consists of the policies

$$= TH[d(\tau^\star)] d^{\text{TV}}(\pi^1, \pi^2) = \mathcal{L}(\pi_1, \pi_2) d^{\text{TV}}(\pi^1, \pi^2)$$

It follows a (bound on a) Lipschitz constant dependent on the two policies to be compared that is directly proportional to the best single trajectory (in terms of entropy) reachable by the policies themselves. Any policy able to generate a maximum entropic trajectory will have the highest possible Lipschitz constant. The constant then gets steeper as the quality of the policies improves.

MBE (Theorem 5.2): Similarly to the previous steps,

$$\begin{aligned}|\tilde{J}(\pi_1) - \tilde{J}(\pi_2)| &= \Big| \mathop{\mathbb{E}}_{\tau_{\mathcal{B}}\sim p^{\pi_1}} \mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}}\sim\tau_{\mathcal{B}}(\cdot)}[H(d(\tau_{\tilde{\mathcal{S}}}))] - \mathop{\mathbb{E}}_{\tau_{\mathcal{B}}\sim p^{\pi_2}} \mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}}\sim\tau_{\mathcal{B}}(\cdot)}[H(d(\tau_{\tilde{\mathcal{S}}}))]\Big| \\ &\leq \sum_{\tau_{\mathcal{B}}\in T_1\cup T_2} \sum_{\tau_{\tilde{\mathcal{S}}}} \tau_{\mathcal{B}}(\tau_{\tilde{\mathcal{S}}}) H(d(\tau_{\tilde{\mathcal{S}}}))\Big|p^{\pi_1}(\tau_{\mathcal{B}}) - p^{\pi_2}(\tau_{\mathcal{B}})\Big| \\ &= \sum_{\tau_{\mathcal{B}}\in T_1\cup T_2} \mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}}\sim\tau_{\mathcal{B}}} H(d(\tau_{\tilde{\mathcal{S}}}))\Big|p^{\pi_1}(\tau_{\mathcal{B}}) - p^{\pi_2}(\tau_{\mathcal{B}})\Big|\end{aligned}$$

Again let us define $\tau_{\mathcal{B}}^\star \in \arg\max_{\tau_{\mathcal{B}}\in T_1\cup T_2} \mathbb{E}_{\tau_{\tilde{\mathcal{S}}}\sim\tau_{\mathcal{B}}} H(d(\tau_{\tilde{\mathcal{S}}}))$

$$\begin{aligned}&\leq \mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}}\sim\tau_{\mathcal{B}}^\star} H(d(\tau_{\tilde{\mathcal{S}}})) d^{\text{TV}}(p^{\pi_1}, p^{\pi_2}) \\ &\leq \mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}}\sim\tau_{\mathcal{B}}^\star} H(d(\tau_{\tilde{\mathcal{S}}})) \sum_t^T d^{\text{TV}}(p_t^{\pi_1}, p_t^{\pi_2}) \\ &= T \mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}}\sim\tau_{\mathcal{B}}^\star} H(d(\tau_{\tilde{\mathcal{S}}})) d^{\text{TV}}(\pi^1, \pi^2) = \tilde{\mathcal{L}}(\pi_1, \pi_2) d^{\text{TV}}(\pi^1, \pi^2)\end{aligned}$$

Again, the (local) Lipschitz constant $\tilde{\mathcal{L}}(\pi_1, \pi_2)$ is dependent on the maximum (expected) entropy that can be induced by one of the policies. One may notice that $\mathcal{L}(\pi_1, \pi_2)$ will be usually higher than $\tilde{\mathcal{L}}(\pi_1, \pi_2)$.

A.5. Proofs of Theorem 5.4: Proxy Gaps

MOE: Let us define the set of observation-trajectories that have an entropy higher than the entropy of a fixed trajectory over true states, namely $\mathcal{T}_{\mathcal{O}}(\tau_{\mathcal{S}}) = \{\tau_{\mathcal{O}} \in \mathcal{T}_{\mathcal{O}} : H(d(\tau_{\mathcal{O}})) \geq H(d(\tau_{\mathcal{S}}))\}$. It follows that by employing the conditional trajectory probability $p^{\pi}(\tau_{\mathcal{O}}|\tau_{\mathcal{S}})$ one can define the probability $\mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}}) = \sum_{\tau_{\mathcal{O}} \in \mathcal{T}_{\mathcal{O}}(\tau_{\mathcal{S}})} p^{\pi}(\tau_{\mathcal{O}}|\tau_{\mathcal{S}})$. It follows that

$$\begin{aligned} J^{\mathcal{S}} - J^{\mathcal{O}} &= \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - J^{\mathcal{O}}(\pi|\tau_{\mathcal{S}}) \Big] \\ &= \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - \mathop{\mathbb{E}}_{\tau_o \sim p^{\pi}(\cdot|\tau_{\mathcal{S}})} H(d(\tau_o)) \Big] \\ &= \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - \sum_{\tau_o} p^{\pi}(\tau_o|\tau_{\mathcal{S}}) H(d(\tau_o)) \Big] \end{aligned}$$

By definition $H(d(\tau_{\mathcal{O}} \in \mathcal{T}_{\mathcal{O}}(\tau_{\mathcal{S}}))) \geq H(d(\tau_{\mathcal{S}}))$, and by positivity of the entropy function $H(d(\tau_{\mathcal{O}} \notin \mathcal{T}_{\mathcal{O}}(\tau_{\mathcal{S}}))) \geq 0$

$$\begin{aligned} &\leq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - \mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}}) H(d(\tau_{\mathcal{S}})) \Big] \\ &\leq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[(1 - \mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}})) H(d(\tau_{\mathcal{S}})) \Big] \end{aligned}$$

It follows that

$$J^{\mathcal{S}}(\pi) \leq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[\frac{1}{\mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}})} J^{\mathcal{O}}(\pi|\tau_{\mathcal{S}}) \Big]$$

In the same way, focusing on the terms inside the outer expectation for simplicity, one obtains:

$$\begin{aligned} J^{\mathcal{S}} - J^{\mathcal{O}} &= \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - J^{\mathcal{O}}(\pi|\tau_{\mathcal{S}}) \Big] \\ &= \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - \mathop{\mathbb{E}}_{\tau_o \sim p^{\pi}(\cdot|\tau_{\mathcal{S}})} H(d(\tau_o)) \Big] \\ &= \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - \sum_{\tau_o} p^{\pi}(\tau_o|\tau_{\mathcal{S}}) H(d(\tau_o)) \Big] \end{aligned}$$

Again, one notices that $H(d(\tau_{\mathcal{O}} \in \mathcal{T}_{\mathcal{O}}(\tau_{\mathcal{S}})))) \leq H(d(\tau_{\mathcal{S}}))$ and $H(d(\tau_{\mathcal{O}} \in \mathcal{T}_{\mathcal{O}}(\tau_{\mathcal{S}})))) \leq \log(O)$, from which the inner expectation turns out to be bounded by the use of the complementary probability $\mathbb{P}(\mathcal{T}_{\mathcal{O}}^{C}|\tau_{\mathcal{S}}) = \sum_{\tau_{\mathcal{O}} \notin \mathcal{T}_{\mathcal{O}}(\tau_{\mathcal{S}})} p^{\pi}(\tau_{\mathcal{O}}|\tau_{\mathcal{S}})$

$$\begin{aligned} &\geq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[(1 - \mathbb{P}(\mathcal{T}_{\mathcal{O}}^{C}|\tau_{\mathcal{S}})) H(d(\tau_{\mathcal{S}})) - \mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}}) \log(O) \Big] \\ &= \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[\mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}}) H(d(\tau_{\mathcal{S}})) - \mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}}) \log(O) \Big] \end{aligned}$$

Leading to

$$J^{\mathcal{S}}(\pi) \geq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[\frac{J^{\mathcal{O}}(\pi|\tau_{\mathcal{S}}) - \mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}}) \log(O)}{1 - \mathbb{P}(\mathcal{T}_{\mathcal{O}}|\tau_{\mathcal{S}})} \Big]$$

MBE: Let us define the similar set for hallucinated trajectories $\mathcal{T}(\tau_{\mathcal{S}}) = \{\tau_{\tilde{\mathcal{S}}} \in \mathcal{T}_{\tilde{\mathcal{S}}} : H(d(\tau_{\tilde{\mathcal{S}}})) \geq H(d(\tau_{\mathcal{S}}))\}, \mathbb{P}(\mathcal{T}|\tau_{\mathcal{B}}) = \sum_{\tau_{\mathcal{S}} \in \mathcal{T}(\tau_{\mathcal{S}})} \tau_{\mathcal{B}}(\tau_{\mathcal{S}})$.

$$\begin{aligned} J^{\mathcal{S}}(\pi) - \tilde{J}(\pi) &= \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - \tilde{J}^{\mathcal{S}}(\pi|\tau_{\mathcal{S}}) \Big] \\ &= \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - \mathop{\mathbb{E}}_{\tau_{\mathcal{O}} \tau_{\mathcal{A}}, \tau_{\mathcal{B}} \sim p^{\pi}(\cdot|\tau_{\mathcal{S}})} \mathop{\mathbb{E}}_{\tau_{\tilde{\mathcal{S}}} \sim \tau_{\mathcal{B}}} H(d(\tau_{\tilde{\mathcal{S}}})) \Big] \\ &= \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - \mathop{\mathbb{E}}_{\tau_{\mathcal{O}} \tau_{\mathcal{A}}, \tau_{\mathcal{B}} \sim p(\cdot|\tau_{\mathcal{S}})} \sum_{\tau_{\tilde{\mathcal{S}}}} \tau_{\mathcal{B}}(\tau_{\tilde{\mathcal{S}}}) H(d(\tau_{\tilde{\mathcal{S}}})) \Big] \end{aligned}$$

Again $H(d(\tau \in \mathcal{T}_{\mathcal{S}}(\tau_{\mathcal{S}}))) \geq H(d(\tau_{\mathcal{S}}))$ and $H(d(\tau \notin \mathcal{T}_{\mathcal{S}}(\tau_{\mathcal{S}}))) \geq 0$

$$\leq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[H(d(\tau_{\mathcal{S}})) - \mathop{\mathbb{E}}_{\tau_{\mathcal{O}} \tau_{\mathcal{A}}, \tau_{\mathcal{B}} \sim p(\cdot|\tau_{\mathcal{S}})} [\mathbb{P}(\mathcal{T}|\tau_{\mathcal{B}})] H(d(\tau_{\mathcal{S}})) \Big]$$
$$\leq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[(1 - \mathop{\mathbb{E}}_{\tau_{\mathcal{O}} \tau_{\mathcal{A}}, \tau_{\mathcal{B}} \sim p(\cdot|\tau_{\mathcal{S}})} \mathbb{P}(\mathcal{T}|\tau_{\mathcal{B}})) H(d(\tau_{\mathcal{S}})) \Big]$$

We call $\bar{p}_{\mathcal{S}}(\tau_{\mathcal{S}}) = \mathbb{E}_{\tau_{\mathcal{B}} \sim p^{\pi}(\cdot|\tau_{\mathcal{S}})} \mathbb{P}(\mathcal{T}|\tau_{\mathcal{B}})$, it follows that

$$J^{\mathcal{S}}(\pi) \leq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[\frac{1}{\bar{p}(\tau_{\mathcal{S}})} \tilde{J}^{\mathcal{S}}(\pi|\tau_{\mathcal{S}}) \Big]$$

In the same way as before, by simply changing the definitions accordingly, one obtains that:

$$J^{\mathcal{S}}(\pi) \geq \mathop{\mathbb{E}}_{\tau_{\mathcal{S}} \sim p^{\pi}(\cdot)} \Big[\frac{\tilde{J}^{\mathcal{S}}(\pi|\tau_{\mathcal{S}}) - \bar{p}(\tau_{\mathcal{S}}) \log S}{1 - \bar{p}(\tau_{\mathcal{S}})} \Big]$$

B. Experimental Details

B.1. Code Repository and Reproducibility

The code is available at this link.

B.2. Wall-Clock Time for the Main Experiments

The computational cost of Algorithm 1 is mostly due to the sampling of trajectories in line 3, whereas line 4 and 5 can be computed concurrently to the sampling process and the parameters updates at line 6 are done in parallel. It follows that the computational complexity of Algorithm 1 is

- $\mathcal{O}(KT)$ when we can access to parallel simulators for the POMDP;
- $\mathcal{O}(NKT)$ when the N trajectories are sampled sequentially (e.g., interacting with a real-world system).

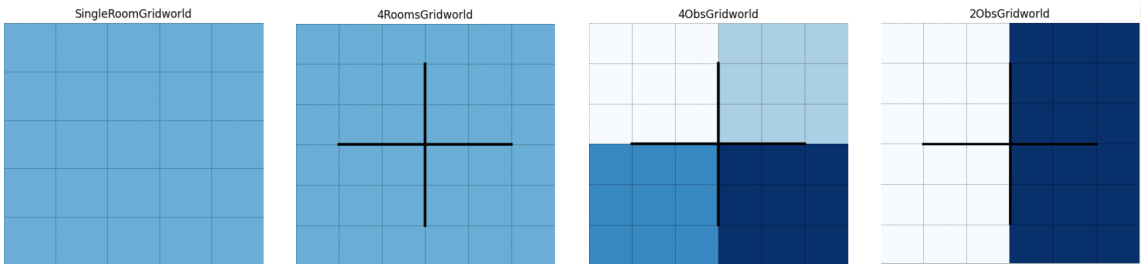
Additionally, here we report a table with the wall-clock time for running the main experiments in the paper. Note that all of the experiments take less than an hour of training on general-purpose CPUs.

Experiment	MSE	MOE	MBE	MBE-Reg
Figure 2.a	1802	1797	2497	2342
Figure 2.b	1800	1791	2537	2465
Figure 2.c	2594	2603	3483	3455
Figure 2.d	2535	2510	3535	3305
Figure 2.e	1780	1803	2423	2582
Figure 2.g	2810	2746	3452	3515

Table 1. Wall-clock time [sec] of the main experiments on general-purpose CPUs.

B.3. POMDP Domains Visualization

In Figure 4 we report a visualization of the four types of domain taken into account.



(a) Single Room (b) Four Rooms (c) Four Rooms with 4 Observations (d) Four Rooms with 2 Observations

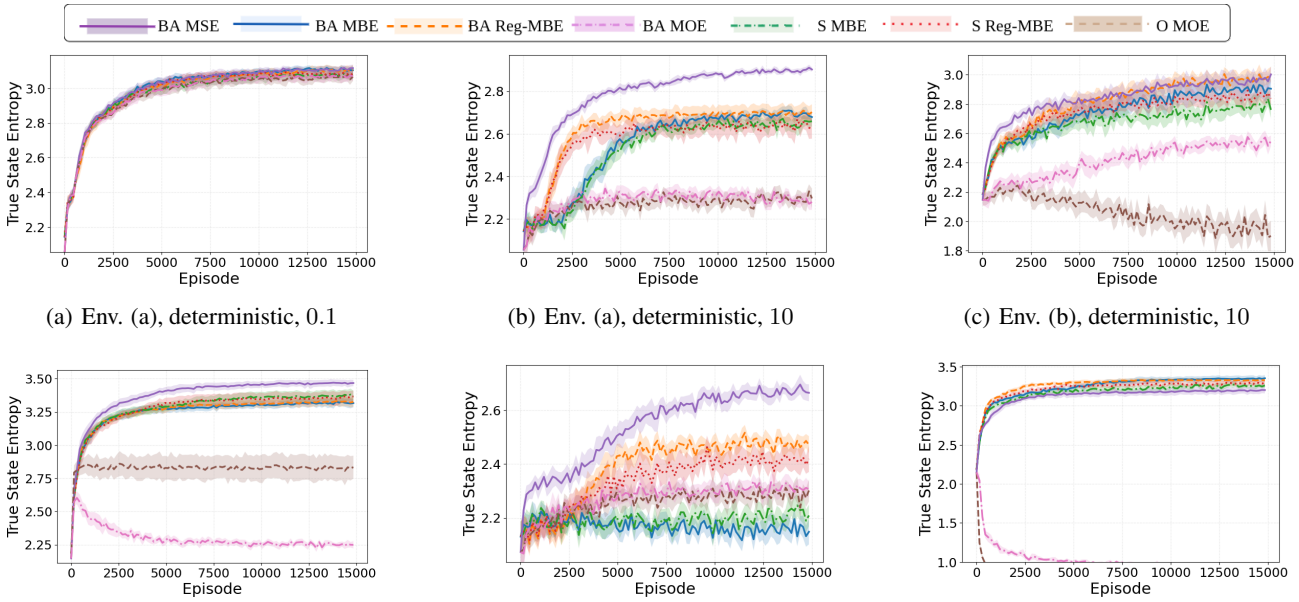
Figure 4. Environments Visualization.

B.4. Choice of Hyperparameters

The **learning rate** was selected as $\alpha = 0.3$. The **batch size** was selected to be $N = 10$ after tuning. As for the **time horizon**, $T = S$ in all the experiments. This makes the exploration task more challenging as every state can be visited at most once. The best regularization term ρ was found to be approximately equal to 0.02.

B.5. Policy Class Investigation

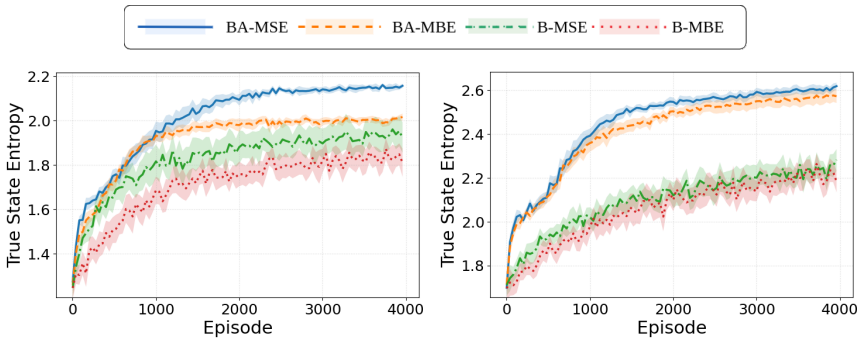
As already described, a plethora of deployable policy classes are possible for addressing MSE in POMDPs. In the main paper, we focused on belief-averaged policies. In Figure 5, we show how this policy class is superior (or non-worse) to other possible options, being implicitly non-Markovian over observations while being memory efficient. In Figure 6, we show that belief-averaged policies perform better than (direct-parametrization) Markovian policies over belief states, even in the case when the belief states set is manageable in size.



(a) Env. (a), deterministic, 0.1 (b) Env. (a), deterministic, 10 (c) Env. (b), deterministic, 10

(d) Env. (c), deterministic, n.a. (e) Env. (a), stochastic, 10 (f) Env. (d), deterministic, n.a.

Figure 5. True state entropy obtained by Algorithm 1 specialized for the feedbacks *MSE, MOE, MBE, MBE with belief regularization* (Reg-MBE) over different policy classes with direct parametrization: Markovian over observation (O), Belief Averaged (BA), Markovian over hallucinated states (S). For each plot, we report a tuple (environment, transition noise, observation variance) where the latter is *not available* (n.a.) when observations are deterministic. For each curve, we report the average and 95% c.i. over 16 runs. BA confirms to be the policy class with generally higher performance in all the considered instances.



(a) Env. (a) ($|\mathcal{S}| = 9$), deterministic, 0.2 (b) Env. (a) ($|\mathcal{S}| = 16$), deterministic, 0.2

Figure 6. True state entropy obtained by Algorithm 1 with MSE and MBE employing belief averaged policies (BA) and Markovian policies over belief states (B). For each plot, we report a tuple (environment, transition noise, observation variance) where the latter is *not available* (n.a.) when observations are deterministic. For each curve, we report the average and 95% c.i. over 16 runs. Limited size instances were reported since $|\mathcal{B}| = 10^4$ in 6(a) and $|\mathcal{B}| = 10^5$ in 6(b) leading to memory issues in the policies storage. Even in these cases, BA shows higher performances.