The Limits of Pure Exploration in POMDPs: When the Observation Entropy is Enough

Riccardo Zamboni
riccardo.zamboni@polimi.it
Politecnico di Milano

Duilio Cirino
duilio.cirino@mail.polimi.it
Politecnico di Milano

Marcello Restelli
marcello.restelli@polimi.it
Politecnico di Milano

Mirco Mutti
mirco.m@technion.ac.il
Technion

Abstract

The problem of *pure exploration* in Markov decision processes has been cast as maximizing the entropy over the state distribution induced by the agent's policy, an objective that has been extensively studied. However, little attention has been dedicated to state entropy maximization under *partial observability*, despite the latter being ubiquitous in applications, e.g., finance and robotics, in which the agent only receives noisy observations of the true state governing the system's dynamics. How can we address state entropy maximization in those domains? In this paper, we study the simple approach of maximizing the *entropy over observations* in place of true latent states. First, we provide lower and upper bounds to the approximation of the true state entropy that only depend on some properties of the observation function. Then, we show how knowledge of the latter can be exploited to compute a principled regularization of the observation entropy to improve performance. With this work, we provide both a flexible approach to bring advances in state entropy maximization to the POMDP setting and a theoretical characterization of its intrinsic limits.

1 Introduction

A plethora of recent works (Hazan et al., 2019; Lee et al., 2019; Mutti & Restelli, 2020; Tarbouriech et al., 2020; Zhang et al., 2021; Guo et al., 2021; Liu & Abbeel, 2021b;a; Seo et al., 2021; Yarats et al., 2021; Mutti et al., 2021; 2022b;c; Nedergaard & Cook, 2022; Yang & Spaan, 2023; Tiapkin et al., 2023; Jain et al., 2023; Kim et al., 2023; Zisselman et al., 2023; Mutti, 2023) have studied *state entropy maximization* for pure exploration of Markov Decision Processes (MDPs, Puterman, 2014) in the absence of a reward function. In this Maximum State Entropy (MSE) framework, formally introduced by Hazan et al. (2019), the agent aims to maximize the entropy of the state visitation induced by its policy instead of the cumulative sum of rewards, which is a generalization of the Reinforcement Learning (RL, Bertsekas, 2019) problem that often goes as *convex* RL (Zahavy et al., 2021; Geist et al., 2021; Mutti et al., 2022a; 2023) due to the convexity of the entropy function.

Despite the problem being harder than RL, MSE has brought remarkable empirical success as a tool for data collection (Yarats et al., 2022), transition model estimation (Tarbouriech et al., 2020), and policy pre-training (Mutti et al., 2021), as well as an essential building block for improved skills discovery (Liu & Abbeel, 2021a) and generalization across various tasks (Zisselman et al., 2023).

Especially, Mutti et al. (2021); Liu & Abbeel (2021b); Seo et al. (2021); Yarats et al. (2021) have popularized a practical *policy optimization* procedure that allows to address MSE at scale. Their method is based on pairing flexible nearest-neighbors estimators of the entropy (Singh et al., 2003)

with policies implemented through neural networks trained via backpropagation, a recipe for success in complex and high-dimensional domains, e.g., continuous control or learning from images.

Although many facets of the MSE problem have been studied, all of the previous works assume the states, on which the entropy is maximized, to be fully observable to the agent. However, this is not the case for several interesting applications. Let us think of a trading scenario: A trader typically accesses a small portion of the true state, e.g., current stock prices, volumes, and so on, while other parts, e.g., the general sentiment of the market or companies' revenues published quarterly, mostly remain latent, albeit crucial to define the system's dynamics. The same goes for a robotic navigation task, where the state is often accessed through noisy sensory inputs, such as cameras and proximity, rather than true spatial coordinates. An important question arises naturally:

Can we maximize the entropy over states getting partial observations only?

While the problem of addressing a learning objective that is hidden from the agent is fascinating per se, we argue that any improvement in this direction is paramount to bringing MSE closer to practical applications. Unfortunately, the problem we just described is a clear generalization of learning in Partially Observable MDPs (POMDPs, Åström, 1965), which is well-known to be intractable.

In this paper, we study the simple approach of maximizing the entropy over the partial observations in place of the true states. This framework, which we call *Maximum Observation Entropy* (MOE), gives two crucial benefits. On the one hand, we can sidestep the inherent computational hardness of dealing with POMDPs (Papadimitriou & Tsitsiklis, 1987), as the class of policies that are Markovian over observations suffices (Hazan et al., 2019) and we do not need to build complex belief distributions over the true state as the objective is fully observable. Secondly, all of the previous implementations can be directly transferred from MSE to MOE without changes, which gives a head start to the MOE problem instead of implementing new techniques from the ground up.

Surprisingly, we can show that this straightforward approach is not hopeless. We derive formal approximation results on the difference between the entropy induced over observations and the corresponding entropy over the true (latent) states, which can be upper bounded as a function of properties of the observation matrix, namely its maximum singular value or the average entropy of its rows. This is in stark contrast with RL in POMDPs, in which the optimal policy over observations is almost arbitrarily sub-optimal under similar assumptions.

Whereas the approximation bounds characterize settings where optimizing for MOE is enough, they tell little about how to address domains in which the observation matrix is not so well-behaved. In those settings, we show that knowledge of the observation matrix, a reasonable requirement in some applications (e.g., the specifics of sensors and cameras equipped on a robot may be available), can be exploited to improve performance over MOE. First, we derive a principled regularization term that discounts the entropy induced by the observations with the entropy of their emission, intuitively putting more weight on the observations for which the emission process is reliable and less on those that are known to be noisy. Then, we incorporate the latter regularization in an appropriate policy gradient algorithm inspired by previous MSE approaches (Mutti et al., 2021; Liu & Abbeel, 2021b). Finally, we test the algorithm on a set of simple yet illustrative domains to validate our theoretical findings, bringing an algorithmic blueprint for scalable state entropy maximization in POMDPs.

Contributions. Throughout the paper, we make the following contributions:

- In Section 3, we provide the first generalization of the MSE problem to the POMDP setting, introducing MOE as a simple yet flexible and tractable approach;
- In Section 4, we theoretically analyze the gap between MOE and MSE, providing a family of upper and lower bounds that link the approximation error to spectral and information properties of the observation matrix;
- In Section 5, we design a policy gradient algorithm for (general) MOE optimization, and we provide a variation including a principle regularization term to exploit knowledge of the observation matrix (but not the POMDP specification);

- In Section 6, we report an empirical validation that tests the introduced algorithms against an ideal baseline accessing the true states of the POMDP, which both upholds the algorithms' design and our theoretical findings.

Finally, in a concurrent work (Zamboni et al., 2024) we explore state entropy maximization in POMDPs beyond MOE, studying methodologies that exploit observations to build *beliefs* over the true states and then maximize the entropy of the *believed* states. With the combined contributions of this paper and Zamboni et al. (2024), we hope to pave the way for future studies of state entropy maximization in partially observable environments.

2 Preliminaries

In this section, we introduce the most relevant background and the basic notation.

Notation. In the following, we denote $[N] := \{1, 2, \dots, N\}$ for a constant $N < \infty$. We denote a set with a calligraphic letter $\mathcal{A}$ and its size as $|\mathcal{A}|$. We denote $\mathcal{A}^T := \times_{t=1}^{T} \mathcal{A}$ the T-fold Cartesian product of $\mathcal{A}$. The simplex on $\mathcal{A}$ is denoted as $\Delta_{\mathcal{A}} := \{p \in [0,1]^{|\mathcal{A}|} | \sum_{a \in \mathcal{A}} p(a) = 1\}$ and $\Delta_{\mathcal{A}}^{\mathcal{B}}$ denotes the set of conditional distributions $p : \mathcal{A} \to \Delta_{\mathcal{B}}$. Let X a random variable on the set of outcomes $\mathcal{X}$ and corresponding probability measure p_X, we denote as $H_\alpha(X) = \frac{1}{1-\alpha} \log(\sum_{x \in \mathcal{X}} p_X(x)^\alpha)$ the Rènyi entropy of order α, from which we recover the Shannon entropy $H(X) = \lim_{\alpha \to 1} H_\alpha(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log(p_X(x))$ and the min-entropy $H_\infty(X) = \lim_{\alpha \to \infty} H_\alpha(X) = -\log(\max_{x \in \mathcal{X}} p_X(x))$. We denote $\mathbf{x} = (X_1, \dots, X_T)$ a random vector of size T and $\mathbf{x}[t]$ its entry at position $t \in [T]$. For a vector $v \in \mathbb{R}^N$ we denote $\|v\|_\infty := \max_{i \in [N]} v_i$. For a matrix $\mathbb{V} \in \mathbb{R}^{N \times M}$ we denote $\|\mathbb{V}\|_\infty := \max_{ij \in [N] \times [M]} V_{ij}$ its infinity norm, $\mathbb{V}^*$ its conjugate transpose and $\mathbb{V}^{\circ -1}$ its Hadamard inverse, such that $V_{ij}^{\circ -1} = 1/V_{ij}\ \forall i, j$. We further denote $\lambda(\mathbb{V}), \sigma(\mathbb{V})$, the vectors of eigenvalues and singular values of $\mathbb{V}$, respectively. We denote the spectral norm of $\mathbb{V}$ as $\|\mathbb{V}\|_2 := \sqrt{\lambda_{\max}(\mathbb{V}^*\mathbb{V})} = \sigma_{\max}(\mathbb{V})$ where $\lambda_{\max}(\mathbb{V}) := \|\lambda(\mathbb{V})\|_\infty$ and $\sigma_{\max}(\mathbb{V}) := \|\sigma(\mathbb{V})\|_\infty$.

As a base model for interaction, we consider a finite-horizon Partially Observable Markov Decision Process (POMDP, Åström, 1965) without rewards. A POMDP $\mathbb{M} := (\mathcal{S}, \mathcal{X}, \mathcal{A}, \mathbb{O}, \mathbb{P}, \mu, T)$ is composed of a set $\mathcal{S}$ of latent states, a set $\mathcal{X}$ of observations, and a set of actions $\mathcal{A}$, which we let discrete and finite with size $|\mathcal{S}|, |\mathcal{X}|, |\mathcal{A}|$ respectively. At the start of an episode, the initial state s_1 of $\mathbb{M}$ is drawn from an initial state distribution $\mu \in \Delta_{\mathcal{S}}$. An agent interacting with $\mathbb{M}$ never accesses the true state of the system but an observation $x_1 \sim \mathbb{O}(\cdot|s_1)$ where $\mathbb{O} \in \Delta_{\mathcal{X}}^{\mathcal{S}}$ is the *observation function*.[1] Upon observing x_1, the agent takes action $a_1 \in \mathcal{A}$, the system transitions to $s_2 \sim \mathbb{P}(\cdot|s_1, a_1)$ according to the *transition model* $\mathbb{P} \in \Delta_{\mathcal{S}}^{\mathcal{S} \times \mathcal{A}}$, and a new observation $x_2 \sim \mathbb{O}(\cdot|s_2)$ is generated. The process is repeated until s_T is reached and x_T is generated, being $T < \infty$ the horizon of an episode.

The agent selects actions according to a decision *policy* $\pi \in \Delta_{\mathcal{X}}^{\mathcal{A}}$ such that $\pi(a|x)$ denotes the conditional probability of taking action a upon observing x. Deploying a policy π over a POMDP $\mathbb{M}$ induces a specific distribution over states and observations. Let denote as S, X the random variables corresponding to the state and observation respectively, we have that S is distributed as $p_S^\pi \in \Delta_{\mathcal{S}}$, where $p_S^\pi(s) = \frac{1}{T} \sum_{t \in [T]} Pr(s_t = s)$, and X as $p_X^\pi \in \Delta_{\mathcal{X}}$, where $p_X^\pi(x) = \frac{1}{T} \sum_{t \in [T]} Pr(x_t = x)$. Further, it is easy to see that $p_X^\pi(x) = \sum_{s \in \mathcal{S}} p_S^\pi(s) \mathbb{O}(x|s)$. Furthermore, let us denote with $\mathbf{s}, \mathbf{x}, \mathbf{a}$ the random vectors corresponding to sequences of states, observations, and actions of length T, which are supported in $\mathcal{S}^T, \mathcal{X}^T, \mathcal{A}^T$ respectively. We have that $\mathbf{s}$ is distributed as $q_S^\pi \in \Delta_{\mathcal{S}^T}$, where $q_S^\pi(\mathbf{s}) = \prod_{t \in [T]} Pr(s_t = \mathbf{s}[t])$, and $\mathbf{x}, \mathbf{a}$ as $q_{XA}^\pi \in \Delta_{\mathcal{X}^T \times \mathcal{A}^T}$, where $q_{XA}^\pi(\mathbf{x}, \mathbf{a}) = \prod_{t \in [T]} Pr(x_t = \mathbf{x}[t], a_t = \mathbf{a}[t])$. Finally, we denote the empirical distributions induced by $\mathbf{s}, \mathbf{x}$ as $\hat{p}_S(s|\,\mathbf{s}) = \frac{1}{T} \sum_{t \in [T]} \mathbb{1}(\mathbf{s}[t] = s)$ and $\hat{p}_X(x|\mathbf{x}) = \frac{1}{T} \sum_{t \in [T]} \mathbb{1}(\mathbf{x}[t] = x)$, which does not depend on π due to the conditioning on $\mathbf{s}, \mathbf{x}$.

[1] With slight overload of notation, we will equivalently represent the observation function through a stochastic matrix $\mathbb{O} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{X}|}$.

3 Problem Formulation

In the MDP setting, i.e., observations coincide with the true states of the state of the system, Hazan et al. (2019) have formulated the *Maximum State Entropy* (MSE) objective as follows

$$\max_{\pi \in \tilde{\Pi}} \left\{ H(S|\pi) := -\sum\nolimits_{s \in \mathcal{S}} p_S^\pi(s) \log p_S^\pi(s) \right\} \tag{1}$$

where $\tilde{\Pi} \subseteq \Delta_{\mathcal{S}}^{\mathcal{A}}$ is the set of Markovian policies from states to distribution over actions, and $H(S|\pi)$ is the entropy of the state variable S "conditioned" on running the policy π in the MDP. When the MDP is fully known, Hazan et al. (2019) shows that (1) is non-convex, but it admits a convex dual formulation, which is optimized by a stochastic Markovian policy in general.

In principle, we aim to address the same objective (1) in POMDPs as well. However, in the POMDP setting, we cannot access the true states, which are latent, but we have to rely on partial observations generated from those states. Thus, a straightforward adaptation of (1) to POMDPs is to define an analogous objective on observations as a proxy for $H(S|\pi)$, which we cannot access. We define the *Maximum Observation Entropy* (MOE) objective as follows

$$\max_{\pi \in \Pi} \left\{ H(X|\pi) := -\sum\nolimits_{x \in \mathcal{X}} p_X^\pi(x) \log p_X^\pi(x) \right\} \tag{2}$$

where $\Pi \subseteq \Delta_{\mathcal{S}}^{\mathcal{A}}$ is the set of Markovian policies from observations to distribution over actions, and $H(X|\pi)$ is the entropy of the observation X "conditioned" on running the policy π in the POMDP.

Similarly, as in MDPs, we aim to find a policy π that maximizes (2), but we are actually interested in achieving a good performance on (1). It is easy to see how the value of (2) can depart significantly from the true objective (1). Take, for instance, an observation matrix that maps every state to the same observation $\mathbb{O}(\bar{x}|s) = 1 \; \forall s \in \mathcal{S}$. It is clear that every policy is optimal for MOE in this setting, but the entropy on the true states can be arbitrarily bad. While those extreme cases are rather unrealistic, the observation matrix can be truly messed up in practice. We want to understand what are the settings that are worth addressing with MOE and what kind of guarantees we can get. In the following section, we provide answers to these questions by deriving theoretical bounds on the approximation of MSE with MOE that depends on crucial properties of the observation matrix.

4 A Formal Characterization of Maximum Observation Entropy

In this section, we aim to characterize the gap $H(S|\pi) - H(X|\pi)$ induced by a chosen policy π, e.g., the policy that maximizes the MOE objective (2). Due to the POMDP nature, in which only partial information (if any) on the true states is leaked to the agent, we cannot provide any general guarantee on the latter gap, which can be as large as

$$|H(S|\pi) - H(X|\pi)| \leq \max\{\log|\mathcal{S}|, \log|\mathcal{X}|\}. \tag{3}$$

Nonetheless, we can provide *instance-dependent* results that formally characterize the gap according to notable properties of the observation function in the given instance. First, we prove the following.

Theorem 4.1 (Spectral Approximation Bounds)**.** *Let $\mathbb{M}$ a POMDP and let $\pi \in \Pi \subseteq \Delta_{\mathcal{X}}^{\mathcal{A}}$ a policy. Then, it holds*

$$\log\left(\frac{1}{\sigma_{\max}(\mathbb{O}^{\circ -1})}\right) \leq H(S|\pi) - H(X|\pi) \leq \log(\sigma_{\max}(\mathbb{O})).$$

Proof. First, we derive the upper bound. Starting from $H(X|\pi)$, we have

$$\begin{aligned} H(X|\pi) \geq H_2(X|\pi) &= \log\left(\frac{1}{\|p_X^\pi\|_2}\right) = \log\left(\frac{1}{\|\mathbb{O} \cdot p_S^\pi\|_2}\right) \\ &\geq \log\left(\frac{1}{\|\mathbb{O}\|_2 \, \|p_S^\pi\|_2}\right) = \log\left(\frac{1}{\|p_S^\pi\|_2}\right) + \log\left(\frac{1}{\|\mathbb{O}\|_2}\right) = H_2(S|\pi) - \log\left(\sigma_{\max}(\mathbb{O})\right) \end{aligned} \tag{4}$$

where the first inequality comes from $H(V) \geq H_2(V)$ for every variable V and the second inequality from $\|\mathbb{V} \cdot v\|_2 \leq \|\mathbb{V}\|_2 \|v\|_2$ for every matrix $\mathbb{V}$ and vector v. Then, starting from $H(S|\pi)$, we get

$$H(S|\pi) = \|p_S^\pi\|_\infty \log\left(\frac{1}{\|p_S^\pi\|_\infty}\right) + \sum_{s:p_S^\pi(s)<\|p_S^\pi\|_\infty} p_S^\pi(s) \log\left(\frac{1}{p_S^\pi(s)}\right) \tag{5}$$

$$\leq \|p_S^\pi\|_\infty H_\infty(S|\pi) + (1 - \|p_S^\pi\|_\infty) \log\left(\frac{|\mathcal{S}| - 1}{1 - \|p_S^\pi\|_\infty}\right)$$

where the inequality is obtained by letting p_S^π be uniformly distributed outside of the entry $\|p_S^\pi\|_\infty$. By noting $H_\infty(V) \leq H_2(V)$ and plugging (5) back to (4) we get

$$H(X|\pi) \geq \frac{H(S|\pi)}{\|p_S^\pi\|_\infty} + \frac{\|p_S^\pi\|_\infty - 1}{\|p_S^\pi\|_\infty} \log\left(\frac{|\mathcal{S}| - 1}{1 - \|p_S^\pi\|_\infty}\right) + \log\left(\frac{1}{\sigma_{\max}(\mathbb{O})}\right) \tag{6}$$

which gives the result for $\|p_S^\pi\|_\infty \to 1$.[2]

To derive the lower bound, we proceed as follows. We start from the $H(X|\pi)$ definition to write

$$H(X|\pi) = \sum_{x\in\mathcal{X}} p_X^\pi(x) \log\left(\frac{1}{p_X^\pi(x)}\right) = \sum_{x\in\mathcal{X}} p_X^\pi(x) \log\left(\frac{\sum_{s\in\mathcal{S}} p_S^\pi(s)}{\sum_s p_S^\pi(s)\mathbb{O}(x|s)}\right) \sum_{s\in\mathcal{S}} p_S^\pi(s) \tag{7}$$

$$\leq \sum_{x\in\mathcal{X}} p_X^\pi(x) \sum_{s\in\mathcal{S}} p_S^\pi(s) \log\left(\frac{p_S^\pi(s)}{p_S^\pi(s)\mathbb{O}(x|s)}\right) = H(S|\pi) + \sum_{x\in\mathcal{X}} p_X^\pi(x) \sum_{s\in\mathcal{S}} p_S^\pi(s) \log\left(\frac{p_S^\pi(s)}{\mathbb{O}(x|s)}\right) \tag{8}$$

$$\leq H(S|\pi) + \mathop{\mathbb{E}}_{x\sim p_X^\pi} \mathop{\mathbb{E}}_{s\sim p_S^\pi} \left[\log(\mathbb{O}^{\circ -1}(x|s))\right] \leq H(S|\pi) + \log\left(\max_{x\in\mathcal{X}} \max_{s\in\mathcal{S}} \mathbb{O}^{\circ -1}(x|s)\right) \tag{9}$$

$$\leq H(S|\pi) + \log\left(\sigma_{\max}(\mathbb{O}^{\circ -1})\right) \tag{10}$$

where we exploit $p_X^\pi(x) = \sum_{s\in\mathcal{S}} p_S^\pi(s)\mathbb{O}(x|s)$ and $\sum_{s\in\mathcal{S}} p_S^\pi(s) = 1$ to write (7), we first apply the log-sum inequality and we split the logarithm to get (8). Then, in (9), we write the first inequality through the definition of the Hadamard inverse of $\mathbb{O}$ and noting that $p_S^\pi(s) \leq 1 \ \forall s \in \mathcal{S}$, we get the second inequality from $\mathbb{E}[V] \leq \max(V)$ for any random variable V and the monotonicity of the logarithm. Finally, we obtain the result (10) by $\|\mathbb{V}\|_\infty \leq \|\mathbb{V}\|_2 = \sigma_{\max}(\mathbb{V})$ for any matrix $\mathbb{V}$. □

Theorem 4.1 gives bounds on the approximation gap that can be much tighter than the worst-case gap in (3). The bounds relate the gap to the scale of the transformation induced by the observation matrix on the distribution of the latent states, which is captured by the maximum singular value of $\mathbb{O}$ and $\mathbb{O}^{\circ -1}$, respectively. For instance, an observation matrix that maps every state to the same observation $\mathbb{O}(\bar{x}|s) = 1 \ \forall s \in \mathcal{S}$ can lead to a larger gap between MOE and MSE, as visualized in the left-hand side of Figure 1. On the other hand, when the observation matrix maps with high probability each state to a different observation, the gap is necessarily smaller (see the right-hand side of Figure 1). Notably, both sides of the bound collapse to zero when the observation matrix is an identity matrix, i.e., when the states are fully observed.

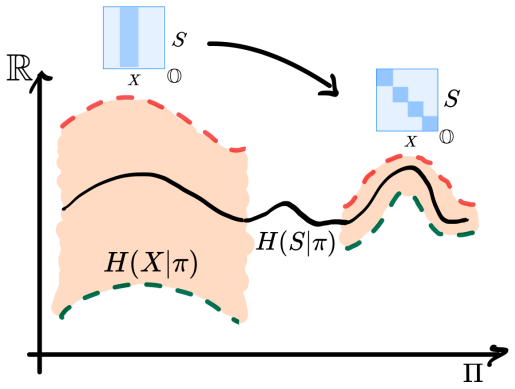


Figure 1: Spectral Bound behavior for two different observation matrices $\mathbb{O}$. MOE values compatible with MSE values are in orange.

The bounds in Theorem 4.1 only focus on spectral properties of the observation matrix $\mathbb{O}$. In a similar vein, we can provide an analogous characterization based on information properties of $\mathbb{O}$.

[2]Note that (6) is a tighter version of the upper bound than the one provided in the theorem statement, although it directly depends on the state distribution p_S^π beyond spectral properties of $\mathbb{O}$.

Theorem 4.2 (Information Approximation Bound)**.** *Let $\mathbb{M}$ a POMDP, let $\pi \in \Pi \subseteq \Delta_{\mathcal{X}}^{\mathcal{A}}$ a policy, and let $H(X|S,\pi) = \mathbb{E}_{s\sim p_S^\pi}[H(\mathbb{O}(\cdot|s))]$. Then, it holds*

$$H(S|\pi) \geq H(X|\pi) - H(X|S,\pi).$$

Proof. Starting from $H(X|\pi)$ definition, we can write

$$\begin{aligned} H(X|\pi) &= \sum_{x\in\mathcal{X}} p_X^\pi(x) \log \frac{1}{p_X^\pi(x)} = \sum_{x\in\mathcal{X}}\sum_{s\in\mathcal{S}} \mathbb{O}(x|s)p_S^\pi(s) \log \frac{1}{\sum_{s'\in\mathcal{S}} \mathbb{O}(x|s')p_S^\pi(s')} \\ &\leq \sum_{x\in\mathcal{X}}\sum_{s\in\mathcal{S}} \mathbb{O}(x|s)p_S^\pi(s) \log \frac{1}{\mathbb{O}(x|s)p_S^\pi(s)} \qquad (11) \\ &= \sum_{x\in\mathcal{X}}\sum_{s\in\mathcal{S}} \mathbb{O}(x|s)p_S^\pi(s) \log \frac{1}{p_S^\pi(s)} + \sum_{x\in\mathcal{X}}\sum_{s\in\mathcal{S}} \mathbb{O}(x|s)p_S^\pi(s) \log \frac{1}{\mathbb{O}(x|s)} \qquad (12) \\ &= H(S|\pi) + \sum_{s\in\mathcal{S}} p_S^\pi(s) H(\mathbb{O}(\cdot|s)) = H(S|\pi) + H(X|S,\pi) \qquad (13) \end{aligned}$$

where we get (11) by noting $\sum_{s'\in\mathcal{S}} \mathbb{O}(x|s')p_S^\pi(s') \geq \mathbb{O}(x|s)p_S^\pi(s)$, we split the logarithm to write (12), we let $\sum_{x\in\mathcal{X}} \mathbb{O}(x|s) = 1$ and $\sum_{s\in\mathcal{S}} p_S^\pi(s) H(\mathbb{O}(\cdot|s)) = H(X|S,\pi)$ to obtain the result in (13). □

Theorem 4.2 essentially states that the gap between the entropy on observations and true states is small as long as the policy π induces visits to states where the observation function has low entropy, which is captured by the term $H(X|S,\pi) = \mathbb{E}_{s\sim p_S^\pi}[H(\mathbb{O}(\cdot|s))]$. When a policy visits states emitting observations with high entropy, the bound on the gap will be loose, as visualized in the left-hand side of Figure 2. Instead, when the most visited states emit almost deterministic observations, then the bound on the gap is tighter (see the right-hand side in Figure 2). Just as Theorem 4.1, also the latter bound is tight when the true states are fully observed, collapsing the gap to zero.

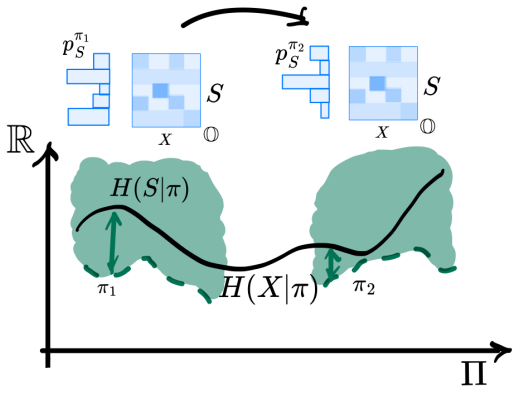


Figure 2: Information Approximation Bound behavior for two different p_S^π. MSE values compatible with MOE values are in green.

The combination of Theorems 4.1, 4.2 yield a nice description of the instances that is reasonable to address with a MOE approach, i.e., those for which the gap between the resulting policy and the optimal MSE policy is small thanks to the properties of the observation matrix. Unfortunately, policies in POMDPs have control over neither the spectral properties of the observation function nor whether the visited states have low-entropy observation distributions. In other words, while being descriptive, these results do not provide any further tool to actively address MSE in POMDPs. In the next section, we reformulate the bound in Theorem 4.2 around quantities that can be actively controlled by a policy conditioned on observations and we provide a family of policy gradient algorithms to learn a MOE policy in those relevant instances.

Before diving into algorithmic solutions, it is interesting to confront the properties making a state entropy maximization problem on POMDPs easy and analogous requirements for RL in POMDPs. In the latter setting, we generally ask for an observation function that leaks significant information on the latent state. For instance, this is captured by a lower bound on the minimum singular value of $\mathbb{O}$ in the *revealing* POMDP assumption (e.g., Liu et al., 2022). Instead, in state entropy maximization, we care less about identifying the latent state, and we can just focus on observations as long as $\mathbb{O}$ does not dramatically jeopardize the underlying state distribution.

Algorithm 1 PG for MOE (**Reg-MOE**)

Input: learning rate α, number of iterations K, batch size N
Initialize the policy parameters θ_1
for $k = 1, \dots, K$ **do**
  Sample N trajectories $\{(\mathbf{x}_i, \mathbf{a}_i)\}_{i\in[N]}$ with the policy π_{θ_k}
  Compute $\{H(X|\mathbf{x}_i)\}_{i\in[N]}$ and $\{\nabla_\theta \log \pi_\theta(\mathbf{x}_i, \mathbf{a}_i) = \sum_{t\in[T]} \nabla_\theta \log \pi_\theta(\mathbf{a}_i[t]|\mathbf{x}_i[t])\}_{i\in[N]}$
  Update the policy parameters in the gradient direction
    $\theta_{k+1} \leftarrow \theta_k + \alpha \frac{1}{N} \sum_i^N \nabla_\theta \log \pi_\theta(\mathbf{x}_i, \mathbf{a}_i)\big(H(X|\mathbf{x}_i) - \beta \sum_{x\in\mathcal{X}} p_X(x|\mathbf{x}_i) H(\mathbb{O}(x|\cdot)\big)$
end for
Output: the final policy π_{θ_K}

5 Towards Principled Policy Gradients for MOE

In the previous section, we analyzed the theoretical guarantees we get on the state entropy maximization problem by optimizing the MOE objective (2), but we did not yet describe how the latter optimization can be performed. Here we propose a family of Policy Gradient algorithms (Peters & Schaal, 2008) to learn a MOE policy from sampled interactions with the POMDP.

First, we define a space of *parametric* policies $\pi_\theta \in \Pi_\Theta \subseteq \Pi$ where $\theta \in \Theta \subseteq \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ are differentiable policy parameters.[3] The expression of the MOE objective does not allow for an easy computation of policy gradients. However, if we let $H(X\,|\,\mathbf{x}) = -\sum_{x\in\mathcal{X}} \hat{p}_X(x|\mathbf{x}) \log \hat{p}_X(x|\mathbf{x})$ the observation entropy induced by a sequence of observations $\mathbf{x}$, we can write a convenient trajectory-based counterpart of (2), namely

$$\max_{\pi_\theta\in\Pi_\Theta} \Big\{ \mathbf{H}(X|\pi_\theta) := \sum\nolimits_{(\mathbf{x},\mathbf{a})\in\mathcal{X}^T\times\mathcal{A}^T} q^{\pi_\theta}_{XA}(\mathbf{x},\mathbf{a}) H(X\,|\,\mathbf{x}) \Big\}. \tag{14}$$

Notably, the trajectory-based objective (14) is a lower bound to the MOE objective (2), due to the concavity of the entropy function and the Jensen's inequality (see Mutti et al., 2022a). Thus, optimizing for (14) guarantees a non-degradation of our initial objective function (2), while it allows for an easy derivation of the gradient ∇_θ w.r.t. the policy parameters.[4]

Proposition 5.1 (Policy Gradient for MOE)**.** *Let $\pi_\theta \in \Pi_\Theta$ a parametric policy and let the* policy scores $\nabla_\theta \log \pi_\theta(\mathbf{x},\mathbf{a}) = \sum_{t\in[T]} \nabla_\theta \log \pi_\theta(\mathbf{a}[t]\,|\,\mathbf{x}[t])$*. We can compute the* policy gradient *of π_θ as*

$$\nabla_\theta \mathbf{H}(X|\pi_\theta) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{a})\sim q^\pi_{XA}} \Big[\nabla_\theta \log \pi_\theta(\mathbf{x},\mathbf{a}) H(X\,|\,\mathbf{x}) \Big].$$

With the latter result, we can design a policy gradient algorithm based on REINFORCE (Williams, 1992). The procedure, described in Algorithm 1, initializes the policy parameters and then performs several iterations of gradient ascent updates. As we shall see in the next section, Algorithm 1 can be a simple yet effective solution to MOE optimization in various settings. However, the resulting policy can be underwhelming in domains where the observation matrix is particularly challenging. While we cannot overcome the barriers established in Theorems 4.1, 4.2, we can still exploit additional information on the observation function to further improve the performance.

Known Observation Matrix. With the knowledge of $\mathbb{O}$, we are tempted to directly optimize the lower bound to $H(S|\pi)$ provided in Theorem 4.2 by trading-off high entropy on observations ($H(X|\pi)$) with the entropy of their emission ($H(X|S,\pi)$). Unfortunately, we do not have access to the state distribution d^π_S to compute the expectation $H(X|S,\pi) = \mathbb{E}_{s\sim p^\pi_S}[H(\mathbb{O}(\cdot|s))]$. Nonetheless, we can rework the lower bound into an alternative form where all of the terms are known and can be controlled by a policy conditioned on observations only, as it demonstrates the following corollary to Theorem 4.2.

[3] See Deisenroth et al. (2013, Section 1.3) for common choices of parametric policy spaces.
[4] The proof can be found in Appendix A.

Corollary 5.2 (Actionable Lower Bound)**.** *Let* $\mathbb{M}$ *a POMDP, let* $\pi \in \Pi \subseteq \Delta_{\mathcal{X}}^{\mathcal{A}}$ *a policy, and let* $H(S|X,\pi) = \mathbb{E}_{x\sim p_X^\pi}[H(\mathbb{O}(x|\cdot))]$*. Then, it holds*

$$H(S|\pi) \geq H(X|\pi) - H(S|X,\pi) + \log(\sigma_{\max}(\mathbb{O})).$$

Proof. The result follows straightforwardly through further manipulation of Theorem 4.2. We have,

$$H(S|\pi) \geq H(X|\pi) - H(X|S,\pi) = H(X|\pi) - H(S|X,\pi) + H(S|\pi) - H(X|\pi) \tag{15}$$
$$\geq H(X|\pi) - \sum_{x\in\mathcal{X}} p_X^\pi(x) H(\mathbb{O}(x|\cdot)) + \log(\sigma_{\max}(\mathbb{O})) \tag{16}$$

where (15) is the result of the application of the Bayes rule to the conditional entropy $H(X|S,\pi)$ and (16) follows from the fact that $H(X|\pi) - H(S|\pi) \geq -\log(\sigma_{\max}(\mathbb{O}))$ due to Theorem 4.1. □

From the latter result, we get a lower bound to $H(S|\pi)$ that can be controlled, as we flipped the conditioning from $H(X|S,\pi)$ to $H(S|X,\pi)$, which we can compute by taking an expectation with the observation distribution. Visually, when a policy visits observations that can be emitted by many states, the bound on the gap will be looser (Figure 3, left-hand side). When the visited observations are emitted by specific states with high probability, then the bound on the gap is tighter (Figure 3, right-hand side).

Inspired by the rational provided by this bound, it is then possible to explicitly account for the effect of dealing with observation only: for every $\beta \in (0,1)$, we can write a regularized version of (14) as

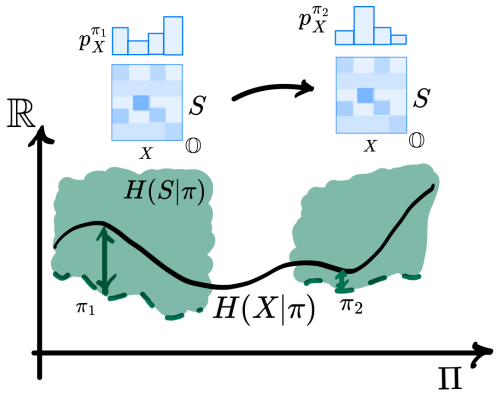


Figure 3: Actionable Lower Bound behavior for two different p_X^π. MSE values compatible with MOE values are in green.

$$\mathbf{H}_\beta(X|\pi) := \sum\nolimits_{(\mathbf{x},\mathbf{a})\in\mathcal{X}^T\times\mathcal{A}^T} q_{XA}^\pi(\mathbf{x},\mathbf{a})\Big(H(X|\,\mathbf{x}) - \beta \sum\nolimits_{x\in\mathcal{X}} \hat{p}_X(x|\,\mathbf{x}) H(\mathbb{O}(x|\cdot)\Big),$$

which we call *Regularized MOE* (Reg-MOE), and a slight variation of the Algorithm 1 (highlighted in the pseudocode) to optimize the regularized objective. In the next section, we provide an empirical validation of the proposed PG algorithms to describe their respective strengths and weaknesses. Note that the presented algorithms can be further enhanced with the same technical solutions of advanced policy optimization algorithms for the MSE objective (e.g., Mutti et al., 2021; Liu & Abbeel, 2021b; Seo et al., 2021; Yarats et al., 2021) to address continuous and high-dimensional domains.

6 Numerical Validation

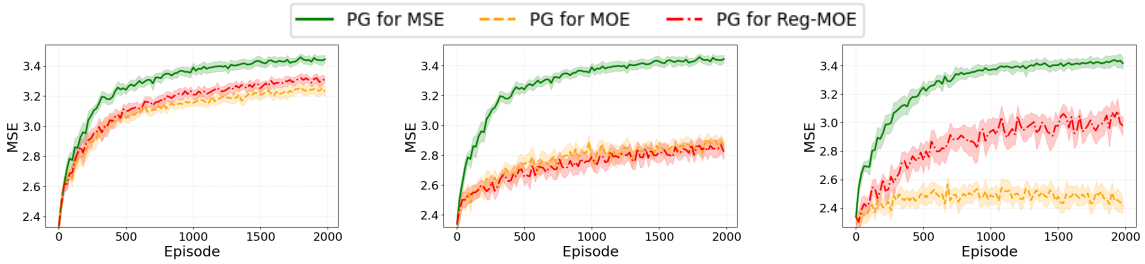
Here we provide a brief numerical validation of the theoretical results provided in Section 4 and the algorithmic solutions proposed in Section 5. Especially, we aim to show that

(a) Optimizing MOE is particularly effective when the observation matrix is "well-behaved";

(b) Optimizing MOE is bound to fail when the observation matrix is not "well-behaved";

(c) Additional knowledge of the observation structure can be sometimes exploited to improve the performance in the latter challenging cases by optimizing the regularized MOE.

Intuitively, an observation matrix is "well-behaved" when it does not induce a significant transformation of the state distribution, keeping the approximation gap between MOE and MSE small. Thanks to Theorems 4.1, 4.2 we can provide a formal characterization of this property. In the



(a) Well-behaved observations $\mathbb{E}[H(\mathbb{O})] \approx 1$

(b) Challenging observations $\mathbb{E}[H(\mathbb{O})] \approx 2.2$

(c) Challenging observations with structure $\mathbb{E}[H(\mathbb{O})] \approx 1.85$

Figure 4: Entropy on latent states (MSE) achieved by *PG for MSE*, *PG for MOE*, and *PG for Reg-MOE* in gridworlds with various $\mathbb{O}$. We report the average and 95% c.i. over 16 runs.

experiments below, we measure the latter through the average entropy of the observation function $\mathbb{E}[H(\mathbb{O})] = \sum_{s \in \mathcal{S}} H(\mathbb{O}(\cdot|s))/|\mathcal{S}|$ on the lines of the information bound in Theorem 4.2.

In Figure 4a we test (a) by showing that the performance of the algorithms accessing observations only, i.e., *PG for MOE* and *PG for Reg-MOE*, is remarkably close to the ideal baseline having access to the true states, i.e., *PG for MSE*. This is due to the low average entropy of the observation function: Although the agent cannot know its exact position, maximizing the entropy of observations still leads to a large entropy over the latent states.

This is not the case in the experiment in Figure 4b, where the gridworld configuration is the same, but the observation function is now more challenging, i.e., more entropic on average. The significant gap between the algorithms optimizing MOE and the ideal baseline is a testament of (b) and a corroboration of the theoretical limits of the MOE approach, which are formally provided in Theorems 4.1, 4.2. *PG for MOE* and *PG for Reg-MOE* can still successfully maximize the entropy over observations, but cannot avoid a significant mismatch with the resulting entropy over latent states.

However, not all the domains with challenging (i.e., entropic) observations are hopeless for the MOE approach, especially when we can exploit knowledge on how the observations are themselves generated. In Figure 4c, we report a further experiment in which the observation matrix has a block with very high entropy (in which observations are almost random) and a block with nearly deterministic observations. *PG for MOE* does not exploit the structure of $\mathbb{O}$ and cannot distinguish between observations that are *reliable* from those that are not. Instead, the regularization term in *PG for Reg-MOE* leads to more visitations of reliable observations (i.e., generated with lower entropy) effectively reducing the gap with the ideal baseline (*PG for MSE*), which corroborates both (c) and the result in Corollary 5.2.

As a bottom line, this numerical validation shows that the MOE approach, while not being a solution to every POMDP instance, can still provide a remarkable performance on domains where the observation matrix is not too challenging or when its knowledge can be exploited.

7 Related Work

This work rests in the intersection between POMDPs, state entropy maximization, and policy optimization. Here we report a list of the most relevant contributions in those areas.

POMDPs. Learning and planning problems in POMDPs have been extensively studied. In the most general formulation, POMDPs are known to be computationally and statistically intractable (Papadimitriou & Tsitsiklis, 1987; Krishnamurthy et al., 2016; Jin et al., 2020). Nonetheless, several recent works have analyzed tractable sub-classes of POMDPs under convenient structural assumptions, such as (Jin et al., 2020; Golowich et al., 2022; Chen et al., 2022; Liu et al., 2022; Zhan et al., 2023; Zhong et al., 2023).

State Entropy Maximization. State entropy maximization in MDPs has been introduced in Hazan et al. (2019), from which followed a variety of subsequent works focusing on the problem from various perspectives (Lee et al., 2019; Mutti & Restelli, 2020; Mutti et al., 2021; 2022b;c; 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). Among them, Savas et al. (2022) indeed study the problem of maximizing the entropy over trajectories induced in a POMDP, yet we are the first to formulate *state* entropy maximization in POMDPs in this paper and, concurrently, Zamboni et al. (2024).

Policy Optimization. First-order methods have been extensively employed to address non-concave policy optimization (Sutton et al., 1999; Peters & Schaal, 2008). In this work, we proposed a *vanilla* policy gradient estimator (Williams, 1992) as a first step, yet several further 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 Conclusions

In this paper, we made a step forward into generalizing state entropy maximization in POMDPs. Specifically, we addressed the problem of learning a policy conditioned only by observations that target the entropy over the latent states. We proposed the simple approach of optimizing the entropy over observations in place of latent states and we formally characterized the instances where it is effective by deriving approximation bounds of the latent objective that depend on the structure of the observation matrix. Finally, we design a family of policy gradient algorithms to optimize the observation entropy in practice and to exploit knowledge of the observation structure when available.

Before concluding, it is worth mentioning that state entropy maximization can find further motivation in POMDPs beyond its common use in MDP settings. While how those methods can benefit offline data collection and transition model estimation is less obvious under partial observability, it is worth noting that the reward in a POMDP is usually defined over the true states, such that pre-training a policy to explore over them is still relevant (Eysenbach et al., 2021). Moreover, the policy we aim to learn is commonly a mapping from beliefs to actions, for which ensuring good exploration over beliefs is important. Interestingly, in POMDPs one may choose to pre-train belief representations alone, to be transferred to various downstream tasks. For this problem, accessing data with coverage over the space of beliefs and, consequently, true states is essential. Exploring other potential uses of state entropy maximization in POMDPs is a nice future direction.

Future works may extend our results in many other directions, such as enhancing our algorithms by incorporating recent advancements in policy optimization for MSE (e.g., Liu & Abbeel, 2021b) and designing alternative objectives and algorithms to target domains where the entropy of observations is not enough (Zamboni et al., 2024). To conclude, we believe that this work sets a crucial first step in the direction of extending state entropy maximization to yet more practical settings.

References

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

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

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

Marc Peter Deisenroth, Gerhard Neumann, Jan Peters, et al. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2):1–142, 2013.

Benjamin Eysenbach, R. Salakhutdinov, and S. Levine. The information geometry of unsupervised reinforcement learning. *International Conference on Learning Representations*, 2021.

Matthieu Geist, Julien Pérolat, Mathieu Lauriere, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*, 2021.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning in observable pomdps in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 2022.

Zhaohan Daniel Guo, Mohammad Gheshlagi Azar, Alaa Saade, Shantanu Thakoor, Bilal Piot, Bernardo Avila Pires, Michal Valko, Thomas Mesnard, Tor Lattimore, and Rémi Munos. Geometric entropic exploration. *arXiv preprint arXiv:2101.02055*, 2021.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2019.

Arnav Kumar Jain, Lucas Lehnert, Irina Rish, and Glen Berseth. Maximum state entropy exploration using predecessor and successor representations. In *Advances in Neural Information Processing Systems*, 2023.

Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. In *Advances in Neural Information Processing Systems*, 2020.

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

Dongyoung Kim, Jinwoo Shin, Pieter Abbeel, and Younggyo Seo. Accelerating reinforcement learning with value-conditional state entropy exploration. In *Advances in Neural Information Processing Systems*, 2023.

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

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

Hao Liu and Pieter Abbeel. Aps: Active pretraining with successor features. In *International Conference on Machine Learning*, 2021a.

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

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, 2022.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems*, 2018.

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

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

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

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022a.

Mirco Mutti, Riccardo De Santi, and Marcello Restelli. The importance of non-Markovianity in maximum state entropy exploration. In *International Conference on Machine Learning*, 2022b.

Mirco Mutti, Mattia Mancassola, and Marcello Restelli. Unsupervised reinforcement learning in multiple environments. In *AAAI Conference on Artificial Intelligence*, 2022c.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Convex reinforcement learning in finite trials. *Journal of Machine Learning Research*, 24(250):1–42, 2023.

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

Christos H Papadimitriou and John N Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 2008.

Martin L Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Yagiz Savas, Michael Hibbard, Bo Wu, Takashi Tanaka, and Ufuk Topcu. Entropy maximization for partially observable Markov decision processes. *IEEE Transactions on Automatic Control*, 67 (12):6948–6955, 2022.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, 2015.

Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*, 2021.

Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. Nearest neighbor estimates of entropy. *American Journal of Mathematical and Management Sciences*, 23 (3-4):301–321, 2003.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, 1999.

Jean Tarbouriech, Shubhanshu Shekhar, Matteo Pirotta, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active model estimation in Markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Pierre Perrault, Yunhao Tang, Michal Valko, and Pierre Menard. Fast rates for maximum entropy exploration. In *International Conference on Machine Learning*, 2023.

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

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

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, 2021.

Denis Yarats, David Brandfonbrener, Hao Liu, Michael Laskin, Pieter Abbeel, Alessandro Lazaric, and Lerrel Pinto. Don't change the algorithm, change the data: Exploratory data for offline reinforcement learning. *arXiv preprint arXiv:2201.13425*, 2022.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. In *Advances in Neural Information Processing Systems*, 2021.

Riccardo Zamboni, Duilio Cirino, Marcello Restelli, and Mirco Mutti. How to explore with belief: State entropy maximization in pomdps. In *International Conference on Machine Learning*, 2024.

Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Pac reinforcement learning for predictive state representations. In *International Conference on Learning Representations*, 2023.

Chuheng Zhang, Yuanying Cai, Longbo Huang, and Jian Li. Exploration by maximizing Rényi entropy for reward-free rl framework. In *AAAI Conference on Artificial Intelligence*, 2021.

Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*, 2023.

Ev Zisselman, Itai Lavie, Daniel Soudry, and Aviv Tamar. Explore to generalize in zero-shot rl. In *Advances in Neural Information Processing Systems*, 2023.

A Missing Proofs

Here we report the derivations of the policy gradient reported in Proposition 5.1. Especially, we write

$$\begin{aligned}\nabla_\theta \mathbf{H}(X|\pi_\theta) &= \nabla_\theta \sum\nolimits_{(\mathbf{x},\mathbf{a})\in\mathcal{X}^T\times\mathcal{A}^T} q^{\pi_\theta}_{XA}(\mathbf{x},\mathbf{a}) H(X|\,\mathbf{x}) \\ &= \sum\nolimits_{(\mathbf{x},\mathbf{a})\in\mathcal{X}^T\times\mathcal{A}^T} \Big(\nabla_\theta q^{\pi_\theta}_{XA}(\mathbf{x},\mathbf{a})\Big) H(X|\,\mathbf{x}) \\ &= \sum\nolimits_{(\mathbf{x},\mathbf{a})\in\mathcal{X}^T\times\mathcal{A}^T} q^{\pi_\theta}_{XA}(\mathbf{x},\mathbf{a}) \nabla_\theta \log q^{\pi_\theta}_{XA}(\mathbf{x},\mathbf{a}) H(X|\,\mathbf{x}) \\ &= \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{a})\sim q^{\pi_\theta}_{XA}} \left[\nabla_\theta \log q^{\pi_\theta}_{XA}(\mathbf{x},\mathbf{a}) H(X|\,\mathbf{x})\right]\end{aligned}$$

by exploiting the linearity of the expectation to go from the first to the second equality, then applying the common log-trick (Peters & Schaal, 2008) and finally recognising the sum as an expectation again.

To derive the gradient we then have to provide the calculation of the *policy scores* $\nabla_\theta \log q^{\pi_\theta}_{XA}(\mathbf{x},\mathbf{a})$. For every $\pi \in \Pi_\Theta$, we notice that $q^{\pi_\theta}_{XA}(\mathbf{x},\mathbf{a}) = \prod_{t\in[T]} Pr(x_t = \mathbf{x}[t])\pi_\theta(a_t = \mathbf{a}[t]|x_t = \mathbf{x}[t])$ and that the only term depending on θ is the policy itself. By exploiting the properties of the logarithm we have

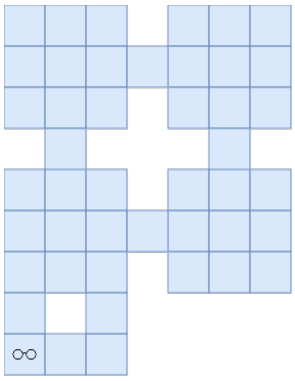
$$\nabla_\theta \log q^{\pi_\theta}_{XA}(\mathbf{x},\mathbf{a}) = \sum_{t\in[T]} \nabla_\theta \log \pi_\theta(a_t = \mathbf{a}[t]|x_t = \mathbf{x}[t])$$

which leads to the standard REINFORCE formulation (Williams, 1992).

B Additional Details on the Experiments

In the following, we report additional details on the experiments of Section 6. Specifically, we describe the employed domains and their properties in Appendix B.1, we comment on the choice of hyper-parameters in Appendix B.2, and on the effect of the regularization on the results of *PG for Reg-MOE* in Appendix B.3.

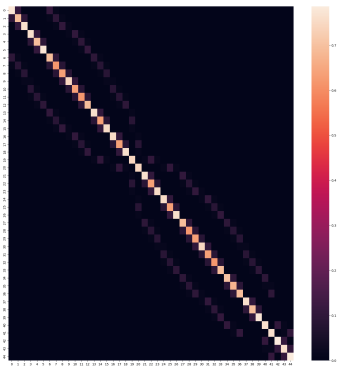
B.1 Domains



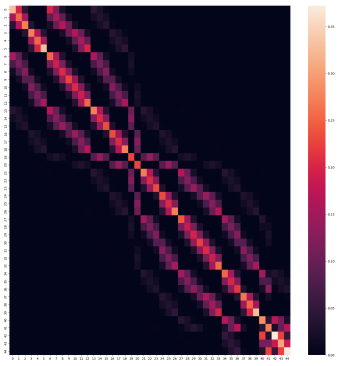
Most of the reported experiments refer to the grid-world reported on the left, which is composed of a set of rooms connected by narrow corridors. The grid is composed of 44 cells, which define both the set of states ($|\mathcal{S}| = 44$) and observations ($|\mathcal{X}| = 44$). The set of actions $\mathcal{A}$ include an action to move to the adjacent cell in every direction ($|\mathcal{A}| = 4$). To every action is associated a probability of failure $\bar{p} = 0.1$ that leads the agent to an adjacent cell (at random) different from the one intended by the taken action. The episode horizon is $T = 55$ and the initial state distribution μ was set to be a deterministic over the top-left cell.

The glasses icon in the bottom left cell of the grid represents a state that "flips" the behavior of the observations. This is only relevant in the experiment in Figure 4c and is better explained below. All the experiments of Section 6 were performed with a regularization factor $\beta = 0.8$ (for *PG for Reg-MOE*) and a learning rate of $\alpha = 0.9$. Finally, the batch size was $N = 6$ and the number of independent runs was set to 16.

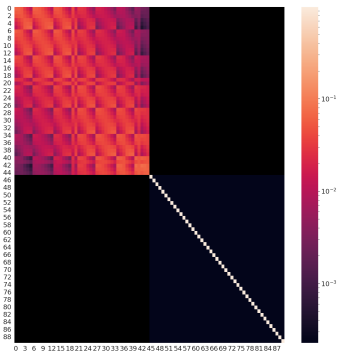
Observations. The observations were set to be Gaussian distributions $\mathcal{G}(0, \sigma^2)$ over the Manhattan distance centered in the true state and without caring about any obstacles, with 0 mean and different values of variance σ^2. The resulting observation matrices are reported in Figure 5. Finally, the effect of "wearing" the glasses (i.e., reaching the bottom-left cell of the grid) is to make the observation function fully deterministic. Note that the information on whether the agent wears the glasses is encoded in the state themselves, doubling the size of the set of states to $|\mathcal{S}| = 88$.



(a) Observation matrix of the experiment in Figure 4a



(b) Observation matrix of the experiment in Figure 4b



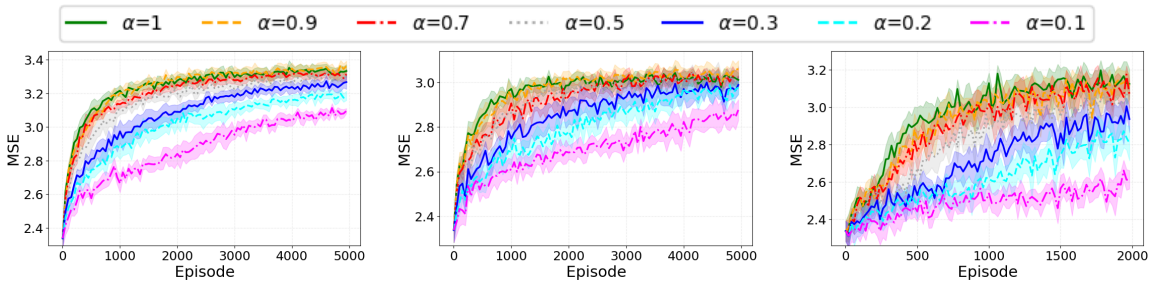
(c) Observation matrix of the experiment in Figure 4c

Figure 5: Heatmaps representing the observation matrix $\mathbb{O}$ employed in the experiments of Section 6. Note that in Figure 5c the colormap has logarithmic scale.

B.2 Hyper-Parameters Selection

In this section, we briefly discuss the choice behind the selection of specific hyper-parameters employed in the experiments.

Learning Rate. As for the learning rate α, a value of $\alpha = 0.9$ was selected across the experiments. As one can see from Figure 6, the best performance were reached with a learning rate between $\alpha = 1$ and $\alpha = 0.7$, so $\alpha = 0.9$ can be seen as a robust choice across the boards.



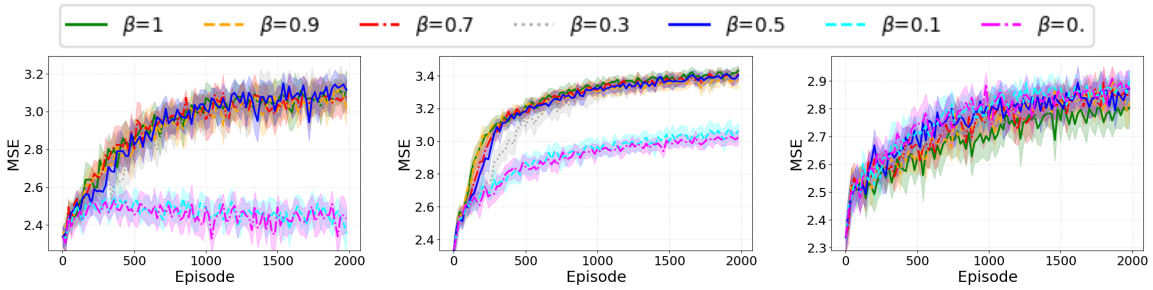
(a) PG for MOE (observation matrix as in Figure 5a)

(b) PG for MOE (observation matrix as in Figure 5b)

(c) PG for Reg-MOE (observation matrix as in Figure 5c)

Figure 6: Comparison of the performance with different values of the learning rate for various algorithms and domains.

Regularization. As for the regularization term β, the best performance for the various instances was generally reached with $\beta \in (0.3, 1)$, as shown in Figure 7 (the learning rate is fixed to $\alpha = 0.9$). For lower values of β, the effect of the regularization is almost negligible, while for higher values of β the agent tended to over-optimize the regularization term in place of the entropy over observations, reducing performance. As one would expect, the best value for the regularization depend on the specific POMDP instance.



(a) With glasses (variance of observations $\sigma^2 = 10$)

(b) With glasses (variance of observations $\sigma^2 = 1$)

(c) Without glasses (variance of observations $\sigma^2 = 0.25$)

Figure 7: A comparison of different values of regularization for varying emission matrices' quality and settings with and without glasses. For the low value of regularization, the performances of Reg-MOE are equivalent to the MOE performances.

B.3 Further Insights on the Effect of the Regularization

In this section, we further investigate the effect of the regularization term. For this specific test, we consider a different gridworld configuration than previous experiments, which is reported on the right. The observation matrix is designed as a Gaussian $\mathcal{G}(0, \sigma^2)$ over the Manhattan distance in the blue rooms, while it is deterministic (and thus fully revealing) in the red room. For this experiment, we set the variance to $\sigma^2 = 1$, the regularization term to $\beta = 0.3$, and the horizon $T = 40$. As for the remaining parameters, they are kept as in the previous experiments.

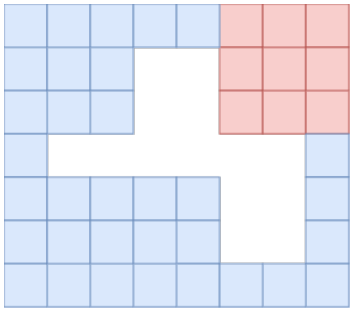
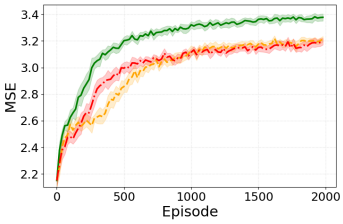
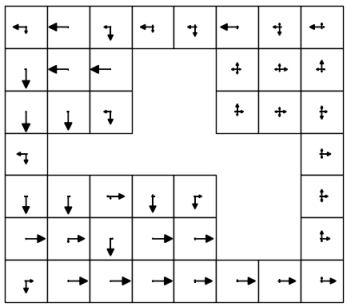


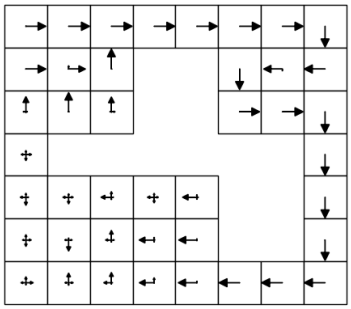
Figure 8 shows that, in this experiment, the two learned policies have similar performances. Yet, as can be seen, while the policy trained with *PH for MOE* tries to explore the environment uniformly, the one trained with *PG for Reg-MOE* successfully explored the portion of the grid with lower entropy in the observations, to later address a deeper exploration of the remaining rooms. This behaviour exactly aligns with the role of the regularization term, which should indeed make the agent prefer observations that are emitted with lower entropy by the observation function.



(a) MSE performance of PG for MSE (green), PG for MOE (orange), PG for Reg-MOE (red)



(b) PG for MOE



(c) PG for Reg-MOE

Figure 8: Comparison of the policies learned by PG for MOE and PG for Reg-MOE over 2000 episodes. The magnitude of each arrow is proportional to the probability of the policy to choose that action, after marginalizing over all the possible observations emitted in that state.