Towards Unsupervised Multi-Agent Reinforcement Learning via Task-Agnostic Exploration

Riccardo Zamboni
Politecnico di Milano
Milan, Italy
riccardo.zamboni@polimi.it

Mirco Mutti
Technion
Haifa, Israel
mirco.m@technion.ac.il

Marcello Restelli
Politecnico di Milano
Milan, Italy
marcello.restelli@polimi.it

Abstract

In reinforcement learning, we typically refer to *unsupervised* pre-training when we aim to pre-train a policy without a priori access to the task specification, i.e., rewards, to be later employed for efficient learning of downstream tasks. In single-agent settings, the problem has been extensively studied and mostly understood. A popular approach, called *task-agnostic exploration*, casts the unsupervised objective as maximizing the *entropy* of the state distribution induced by the agent's policy, from which principles and methods follow. In contrast, little is known about it in multi-agent settings, which are ubiquitous in the real world. What are the pros and cons of alternative problem formulations in this setting? How hard is the problem in theory, how can we solve it in practice? In this paper, we address these questions by first characterizing those alternative formulations and highlighting how the problem, even when tractable in theory, is non-trivial in practice. Then, we present a scalable, decentralized, trust-region policy search algorithm to address the problem in practical settings. Finally, we provide numerical validations to both corroborate the theoretical findings and pave the way for unsupervised multi-agent reinforcement learning via task-agnostic exploration in challenging domains, showing that optimizing for a specific objective, namely *mixture entropy*, provides an excellent trade-off between tractability and performances.

1 Introduction

Multi-Agent Reinforcement Learning [MARL, 2] recently showed promising results in learning complex behaviors, such as coordination and teamwork [35], strategic planning in the presence of imperfect knowledge [33], and trading [11]. Just like in single-agent RL, however, most of the efforts are focused on tabula rasa learning, that is, without exploiting any prior knowledge gathered from offline data and/or policy pre-training. Despite its generality, learning tabula rasa hinders MARL from addressing real-world situations, where training from scratch is slow, expensive, and arguably unnecessary [1]. In this regard, some progress has been made on techniques specific to the multi-agent setting, ranging from ad hoc teamwork [23] to zero-shot coordination [9], but our understanding of what can be done *instead of* learning tabula rasa is still limited.

In single-agent RL, unsupervised pre-training frameworks [13] have emerged as a viable solution: a policy is pre-trained without a priori access to the task specification, i.e., rewards, to be later employed for efficient learning of downstream tasks. Among others, state-entropy maximization [8, 14] was shown to be a useful tool for policy pre-training [8, 25] and data collection for offline learning [44]. In this setting, the unsupervised objective is casted as maximizing the entropy of the state distribution

First Exploration in AI Today Workshop at ICML (EXAIT at ICML 2025).

induced by the agent's policy. Recently, the potential of entropy objectives in MARL was empirically corroborated by a plethora of works [19, 50, 42, 41] investigating entropic reward-shaping techniques to boost exploration in downstream tasks. Yet, to the best of our knowledge, we are still lacking a principled understanding of how task-agnostic exploration works in multi-agent settings, and how it can be used as an unsupervised pre-training fashion. Let us think of an illustrative example that highlights the central question of this work: multiple autonomous robots deployed in a factory for a production task. The robots' main goal is to perform many tasks over a large set of products, with objectives such as optimizing for costs or energy or throughput might be changing over time depending on the market's condition. Arguably, trying to learn each possible task from scratch is inefficient and unnecessary. On the other hand, one could think to first learn to cover the possible states of the system, to learn its dynamics, and then fine-tune over a specific task. Yet, if everyone is focused on their own exploration, any incentive to collaborate with each other may disappear, especially when coordinating comes at a cost. Similarly, covering the entire space might be unreasonable in most real-world cases. Clearly, a third option is needed.

Research Questions and Paper Structure. Throughout this paper, we will address the following questions: (**Q1**) *Can we formulate in a principled way unsupervised pre-training via task-agnostic exploration in MARL as well?* (Section 3);(**Q2**) *How are different formulations related? Do crucial theoretical differences emerge?* (Section 4); (**Q3**) *Can we explicitly pre-train a policy for task-agnostic exploration in practical multi-agent scenarios?* (Section 5); (**Q4**) *Do crucial differences emerge in practice? Does this have an impact on downstream tasks learning?* (Section 6).

2 Preliminaries

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

Notation. 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 $|\mathcal{A}|$. For a (finite) set $\mathcal{A} = \{1, 2, \dots, i, \dots\}$, we denote $-i = \mathcal{A}/\{i\}$ the set of all its elements but i. $\mathcal{A}^T := \times_{t=1}^T \mathcal{A}$ is the T-fold Cartesian product of $\mathcal{A}$. The simplex on $\mathcal{A}$ is $\Delta_\mathcal{A} := \{p \in [0,1]^{|\mathcal{A}|} | \sum_{a\in\mathcal{A}} p(a) = 1\}$ and $\Delta_\mathcal{A}^\mathcal{B}$ denotes the set of conditional distributions $p : \mathcal{A} \to \Delta_\mathcal{B}$. Let X, X' random variables on the set of outcomes $\mathcal{X}$ and corresponding probability measures $p_X, p_{X'}$, we denote the Shannon entropy of X as $H(X) = -\sum_{x\in\mathcal{X}} p_X(x) \log(p_X(x))$ and the Kullback-Leibler (KL) divergence as $D_{\mathrm{KL}}(p_X \| p_{X'}) = \sum_{x\in\mathcal{X}} p_X(x) \log(p_X(x)/p_{X'}(x))$. We denote $\mathbf{x} = (X_1, \dots, X_T)$ a random vector of size T and $\mathbf{x}[t]$ its entry at position $t \in [T]$.

Interaction Protocol. As a model for interaction, we consider finite-horizon Markov Games [MGs, 16] *without rewards*. A MG $\mathcal{M} := (\mathcal{N}, \mathcal{S}, \mathcal{A}, \mathbb{P}, \mu, T)$ is composed of a set of agents $\mathcal{N}$, a set $\mathcal{S} = \times_{i\in[\mathcal{N}]}\mathcal{S}_i$ of states, and a set of (joint) actions $\mathcal{A} = \times_{i\in[\mathcal{N}]}\mathcal{A}_i$, which we let discrete and finite with size $|\mathcal{S}|, |\mathcal{A}|$ respectively. At the start of an episode, the initial state s_1 of $\mathcal{M}$ is drawn from an initial state distribution $\mu \in \Delta_\mathcal{S}$. Upon observing s_1, each agent takes action $a_1^i \in \mathcal{A}_i$, the system transitions to $s_2 \sim \mathbb{P}(\cdot|s_1, a_1)$ according to the *transition model* $\mathbb{P} \in \Delta^\mathcal{S}_{\mathcal{S}\times\mathcal{A}}$. The process is repeated until s_T is reached and s_T is generated, being $T < \infty$ the horizon of an episode. Each agent acts according to a *policy*, that can be either Markovian, i.e. $\pi^i \in \Delta_\mathcal{S}^{\mathcal{A}^i}$, or Non-Markovian over, i.e. $\pi^i \in \Delta^{\mathcal{A}^i}_{\mathcal{S}^t\times\mathcal{A}^t}$.[1] Also, we will denote as *decentralized-information* policies the ones conditioned on either $\mathcal{S}_i$ or $\mathcal{S}_i^t \times \mathcal{A}_i^t$ for agent i, and *centralized-information* ones the ones conditioned over the full state or state-actions sequences. It follows that the joint action is taken according to the *joint* policy $\Delta_\mathcal{S}^\mathcal{A} \ni \pi = (\pi^i)_{i\in[\mathcal{N}]}$.

Induced Distributions. Now, let us denote as S and S_i the random variables corresponding to the joint state and i-th agent state respectively. Then the former is distributed as $d^\pi \in \Delta_\mathcal{S}$, where $d^\pi(s) = \frac{1}{T}\sum_{t\in[T]} Pr(s_t = s|\pi, \mu)$, the latter is distributed as $d_i^\pi \in \Delta_{\mathcal{S}_i}$, where $d_i^\pi(s_i) = \frac{1}{T}\sum_{t\in[T]} Pr(s_{t,i} = s_i|\pi, \mu)$. Furthermore, let us denote with $\mathbf{s}, \mathbf{a}$ the random vectors corresponding to sequences of (joint) states, and actions of length T, which are supported in $\mathcal{S}^T, \mathcal{A}^T$ respectively. We define $p^\pi \in \Delta_{\mathcal{S}^T\times\mathcal{A}^T}$, where $p^\pi(\mathbf{s}, \mathbf{a}) = \prod_{t\in[T]} Pr(s_t = \mathbf{s}[t], a_t = \mathbf{a}[t])$. Finally, we denote the empirical state distribution induced by $K \in \mathbb{N}^+$ trajectories $\{\mathbf{s}_k\}_{k\in[K]}$ as $d_K(s) = \frac{1}{KT}\sum_{k\in[K]}\sum_{t\in[T]} \mathbb{1}(\mathbf{s}_k[t] = s)$.

Convex MDPs and Task-Agnostic Exploration. In the MDP setting ($|\mathcal{N}| = 1$), the problem of task-agnostic exploration, together with others as imitation-learning, diverse skill discovery and

[1] In general, we will denote the set of valid per-agent policies with Π^i and the set of joint policies with Π.

inverse RL, is a special case of *convex RL* [8, 48, 45]. In such framework, the general task is defined via an F-bounded concave[2] utility function $\mathcal{F} : \Delta_{\mathcal{S}} \to (-\infty, F]$, with $F < \infty$, that is a function of the state distribution d^π. This allows for a generalization of the standard RL objective, which is a linear product between a reward vector and the state(-action) distribution [34]. Usually, some regularity assumptions are enforced on the function $\mathcal{F}$. In the following, we align with the most common Lipschitz assumption, explicitly described in Appendix B. More recently, [26] noticed that in many practical scenarios only a finite number of $K \in \mathbb{N}^+$ episodes/trials can be drawn while interacting with the environment, and in such cases one should focus on d_K rather than d^π. As a result, they contrast the *infinite-trials* objective defined as $\zeta_\infty(\pi) := \mathcal{F}(d^\pi)$ with a *finite-trials* one, namely $\zeta_K(\pi) := \mathbb{E}_{d_K \sim p_K^\pi} \mathcal{F}(d_k)$, noticing that convex MDPs (cMDPs) are characterized by the fact that $\zeta_K(\pi) \leqslant \zeta_\infty(\pi)$, differently from standard (linear) MDPs for which equality holds. In single-agent convex RL, task-agnostic exploration is defined as solving a cMDP equipped with an entropy functional [8], namely $\mathcal{F}(d^\pi) := H(d^\pi)$.

3 Problem Formulation

This section addresses the first of the research questions:

(**Q1**) *Can we formulate in a principled way unsupervised pre-training via task-agnostic exploration in MARL as well?*

In fact, when a reward function is not available, the core of the problem resides in finding a well-behaved problem formulation coherent with the task. [7] recently introduced a convex generalization of MGs called **convex Markov Games** (cMGs), namely a tuple $\mathcal{M}_{\mathcal{F}} := (\mathcal{N}, \mathcal{S}, \mathcal{A}, \mathbb{P}, \mathcal{F}, \mu, T)$, that consists in a MG equipped with (non-linear) functions of the *stationary joint state* distribution $\mathcal{F}(d^\pi)$. We expand over this definition, by noticing that task-agnostic exploration can be casted as solving a cMG equipped with an entropy functional, namely $\mathcal{F}(\cdot) := H(\cdot)$. Yet, important new questions arise: Over which distributions should agents compute the entropy? How much information should they have access to? Can we define objectives accounting for a finite number of trials? Different answers depict different objectives.

Joint Objectives. The first and most straightforward way to formulate the problem is to define it as in the MDP setting, with the joint state distribution simply taking the place of the single-agent state distribution. In this case, we define *infinite-trials* and *finite-trials Joint* objectives, respectively

$$\max_{\pi=(\pi^i \in \Pi^i)_{i\in[\mathcal{N}]}} \Big\{ \zeta_\infty(\pi) := \mathcal{F}(d^\pi) \Big\} \qquad \max_{\pi=(\pi^i \in \Pi^i)_{i\in[\mathcal{N}]}} \Big\{ \zeta_K(\pi) := \mathop{\mathbb{E}}_{d_K \sim p_K^\pi} \mathcal{F}(d_K) \Big\} \tag{1}$$

In task-agnostic exploration tasks, an optimal (joint) policy will try to cover the joint state space uniformly, either in expectation or over a finite number of trials respectively. In this, the joint formulation is rather intuitive as it describes the most general case of multi-agent exploration. Moreover, as each agent sees a difference in performance explicitly linked to others, this objective should be able to foster coordinated exploration. As we shall see, this comes at a price.

Disjoint Objectives. One might look for formulations that fully embrace the multi-agent setting, such as defining a set of functions supported on per-agent state distributions rather than joint distributions. This intuition leads to *infinite-trials* and *finite-trials Disjoint* objectives:

$$\Big\{ \max_{\pi^i \in \Pi^i} \zeta^i_\infty(\pi^i, \cdot) := \mathcal{F}(d_i^{\pi^i, \cdot}) \Big\}_{i\in[\mathcal{N}]} \qquad \Big\{ \max_{\pi^i \in \Pi^i} \zeta^i_K(\pi^i, \cdot) := \mathop{\mathbb{E}}_{d_K \sim p_K^{\pi^i, \cdot}} \mathcal{F}(d_{K,i}) \Big\}_{i\in[\mathcal{N}]} \tag{2}$$

According to these objectives, each agent will try to maximize her own marginal state entropy separately, neglecting the effect of her actions over others performances. In other words, we expect this objective to hinder the potential coordinated exploration, where one has to take as step down as so allow a better performance overall.

Mixture Objectives. At last, we introduce a problem formulation that will later prove capable of uniquely taking advantage of the structure of the problem. First, we introduce the following:

Assumption 3.1 (Uniformity). *The agents have the same state space $\mathcal{S}_i = \mathcal{S}_j = \tilde{\mathcal{S}}, \forall (i, j) \in \mathcal{N} \times \mathcal{N}$.*[3]

[2] In practice, the function can be either convex, concave, or even non-convex. The term is used to distinguish the objective from the standard (linear) RL objective. We will assume $\mathcal{F}$ is concave if not mentioned otherwise.

[3] One should notice that even in cMGs where this is not (even partially) the case, the assumption can be enforced by padding together the per-agent states.

Under this assumption, we will drop the agent subscript when referring to the per-agent states and use $\tilde{\mathcal{S}}$ instead. Interestingly, this assumption allows us to define a particular distribution: $\tilde{d}^\pi(\tilde{s}) := \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} d_i^\pi(\tilde{s}) \in \Delta_{\tilde{\mathcal{S}}}$. We refer to this distribution as *mixture* distribution, given that it is defined as a uniform mixture of the per-agent marginal distributions. Intuitively, it describes the average probability over all the agents to be in a common state $\tilde{s} \in \tilde{\mathcal{S}}$, in contrast with the joint distribution that describes the probability for them to be in a joint state s, or the marginals that describes the probability of each one of them separately.

Similarly to what happens for the joint distribution, one can define the empirical distribution induced by K episodes as $\tilde{d}_K(\tilde{s}) = \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} d_{K,i}(\tilde{s})$ and $\tilde{d}^\pi = \mathbb{E}_{\tilde{d}_K\sim p_K^\pi}[\tilde{d}_K]$. The mixture distribution allows for the definition of the *Mixture* objectives, in their infinite and finite trials formulations respectively:

$$\max_{\pi=(\pi^i\in\Pi^i)_{i\in[\mathcal{N}]}} \left\{\tilde{\zeta}_\infty(\pi) := \mathcal{F}(\tilde{d}^\pi)\right\} \qquad \max_{\pi=(\pi^i\in\Pi^i)_{i\in[\mathcal{N}]}} \left\{\tilde{\zeta}_K(\pi) := \mathop{\mathbb{E}}_{\tilde{d}_K\sim p_K^\pi} \mathcal{F}(\tilde{d}_K)\right\} \tag{3}$$

When this kind of objectives is employed in task-agnostic exploration, the entropy of the mixture distribution decomposes as $H(\tilde{d}^\pi) = \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} H(d_i^\pi) + \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} D_{\mathrm{KL}}(d_i^\pi||\tilde{d}^\pi)$ and one remarkable scenario arises: Agents follow policies possibly inducing lower disjoint entropies, but their induced marginal distributions are maximally different. Thus, the average entropy remains low, but the overall mixture entropy is high due to diversity (i.e., high values of the KL divergences). This scenario has been referred to in [12] as the *clustering* scenario and, in the following, we will provide additional evidences why this scenario is particularly relevant.

4 A Formal Characterization of Multi-Agent task-agnostic exploration

In the previous section, we provided a principled problem formulation of multi-agent task-agnostic exploration through an array of different objectives. Here, we address the second research question:

(**Q2**) *How are different formulations related? Do crucial theoretical differences emerge?*

First of all, we show that all the entropy infinite-trials objectives can be linked through the following:

Lemma 4.1 (Entropy Mismatch)**.** *For every cMG $\mathcal{M}_H$, for a fixed (joint) policy $\pi = (\pi^i)_{i\in\mathcal{N}}$ the infinite-trials objectives are ordered according to:*

$$\frac{H(d^\pi)}{|\mathcal{N}|} \leqslant \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} H(d_i^\pi) \leqslant H(\tilde{d}^\pi) \leqslant \sup_{i\in[\mathcal{N}]} H(d_i^\pi) + \log(|\mathcal{N}|) \leqslant H(d^\pi) + \log(|\mathcal{N}|)$$

The full derivation of these bounds is reported in Appendix B. This set of bounds characterizes the performances over infinite-trials objective for the same policy as a function of the number of agents. In particular, disjoint objectives generally provide poor approximations of the joint objective, while the mixture objective is guaranteed to be a rather good lower bound to the joint entropy as well, since its over-estimation scales logarithmically with the number of agents.

It is still an open question how hard it is to actually optimize for these objectives. Here, we provide a rather positive answer, here stated informally and extensively discussed in Appendix B.1:

Fact 4.1 ((Informal) Sufficiency of Independent Policy Gradient)**.** *Under proper assumptions, for every cMG $\mathcal{M}_\mathcal{F}$, independent Policy Gradient over infinite trials non-disjoint objectives (see Eqs.* (1),(3)*) via centralized-information policies of the form $\pi = (\pi^i \in \Delta_{\mathcal{S}}^{\mathcal{A}^i})_{i\in[\mathcal{N}]}$ converges* fast.

This result suggests that PG should be generally enough for the infinite-trials optimization. However, cMDP theory has outlined that optimizing for infinite-trials objectives might actually lead to extremely poor performance as soon as the policies are deployed over just a handful of trials, i.e. in almost any practical scenario [29]. We show that this property transfers to cMGs, with interesting additional takeouts:

Theorem 4.2 (Finite-Trials Mismatch in cMGs)**.** *For every cMG $\mathcal{M}_\mathcal{F}$ equipped with a L-Lipschitz function $\mathcal{F}$, let $K \in \mathbb{N}^+$ be a number of evaluation episodes/trials, and let $\delta \in (0, 1]$ be a confidence level, then for any (joint) policy $\pi = (\pi^i \in \Pi^i)_{i\in[\mathcal{N}]}$, it holds that the mismatches $err(\pi) := |\zeta_K(\pi) - \zeta_\infty(\pi)|, err^i(\pi) := |\zeta_K^i(\pi) - \zeta_\infty^i(\pi)|, \tilde{err}(\pi) := |\tilde{\zeta}_K(\pi) - \tilde{\zeta}_\infty(\pi)|$ are bounded as:*

$$err(\pi) \leqslant LT\sqrt{\frac{2|\mathcal{S}|\log(2T/\delta)}{K}}, \quad err^i(\pi) \leqslant LT\sqrt{\frac{2|\tilde{\mathcal{S}}|\log(2T/\delta)}{K}}, \quad \tilde{err}(\pi) \leqslant LT\sqrt{\frac{2|\tilde{\mathcal{S}}|\log(2T/\delta)}{|\mathcal{N}|K}}.$$

In general, this set of bounds confirms that for any given policy, infinite and finite trials performances might be extremely different, and thus optimizing the infinite-trials objective might lead to unpredictable performance at deployment, whenever this is done over a handful of trials. This property is inherently linked to the *convex* nature of cMGs, and compared to cMDPs [29], in multi-agent settings the result portraits a more nuanced scene: *(i)* The mismatch still scales with the cardinality of the support of the state distribution, yet, for joint objectives, this quantity scales very poorly in the number of agents.[4] Thus, even though optimizing infinite-trials joint objectives might be rather easy *in theory* as Fact 4.1 suggests, it might result in poor performances *in practice*. On the other hand, the quantity is independent of the number of agents for disjoint and mixture objectives. *(ii)* Looking at mixture objectives, the mismatch scales sub-linearly with the number of agents $\mathcal{N}$. In some sense, the number of agents has the same role as the number of trials: The more the agents the less the deployment mismatch, and at the limit, with $\mathcal{N} \to \infty$, the mismatch vanishes completely.[5]

5 An Algorithm for Multi-Agent task-agnostic exploration in Practice

As stated before, a core drive of this work is addressing practical scenarios, where only a handful of trials can be drawn while interacting with the environment. Yet, Th. 4.2 implies that optimizing for infinite-trials objectives, as with PG updates in Fact 4.1, might result in poor performance at deployment. As a result, here we address the third research question, that is:

(**Q3**) *Can we explicitly pre-train a policy for task-agnostic exploration in practical multi-agent scenarios?*

Remarkably, it is possible to directly optimize the single-trial objective in multi-agent cases with decentralized algorithms: We introduce *Trust Region Pure Exploration* (TRPE), the first decentralized algorithm that explicitly addresses single-trial objectives in cMGs, with task-agnostic exploration as a special case. TRPE takes inspiration from trust-region based methods as TRPO [36].

Definition 5.1 (Surrogate Function over a Single Trial)**.** *For every cMG $\mathcal{M}_{\mathcal{F}}$ equipped with a L-Lipschitz function $\mathcal{F}$, let d_1 be a general single-trial distribution $d_1 = \{d_1, d_{1,i}, \tilde{d}_1\}$, then for any per-agent deviation over policies $\pi = (\pi^i, \pi^{-i})$, $\tilde{\pi} = (\tilde{\pi}^i, \pi^{-i})$, it is possible to define a per-agent* Surrogate Function *$\mathcal{L}^i(\tilde{\pi}/\pi)$ of the form $\mathcal{L}^i(\tilde{\pi}/\pi) = \mathbb{E}_{d_1 \sim p_1^{\pi}} \rho^i_{\tilde{\pi}/\pi} \mathcal{F}(d_1)$, where ρ^i is the per-agent importance-weight coefficient $\rho^i_{\tilde{\pi}/\pi} = p_1^{\tilde{\pi}}/p_1^{\pi} = \prod_{t \in [T]} \frac{\tilde{\pi}^i(\mathbf{a}^i[t] \| \mathbf{s}^i[t])}{\pi^i(\mathbf{a}^i[t] \| \mathbf{s}^i[t])}$.*

From this definition, it follows that the trust-region algorithmic blueprint of [36] can be directly applied to single-trial formulations, with a parametric space of stochastic differentiable policies for each agent $\Theta = \{\pi^i_{\theta^i} : \theta^i \in \Theta^i \subseteq \mathbb{R}^q\}$.

In practice, KL-divergence is employed for greater scalability provided a trust-region threshold δ, we address the following optimization problem for each agent:

$$\max_{\tilde{\theta}^i \in \Theta^i} \mathcal{L}^i(\tilde{\theta}^i/\theta^i) \text{ s.t. } D_{\mathrm{KL}}(\pi^i_{\tilde{\theta}^i} \| \pi^i_{\theta^i}) \leqslant \delta$$

where we simplified the notation by letting $\mathcal{L}^i(\tilde{\theta}^i/\theta^i) := \mathcal{L}^i(\pi^i_{\tilde{\theta}^i}, \pi^{-i}_{\theta^{-i}}/\pi_\theta)$.

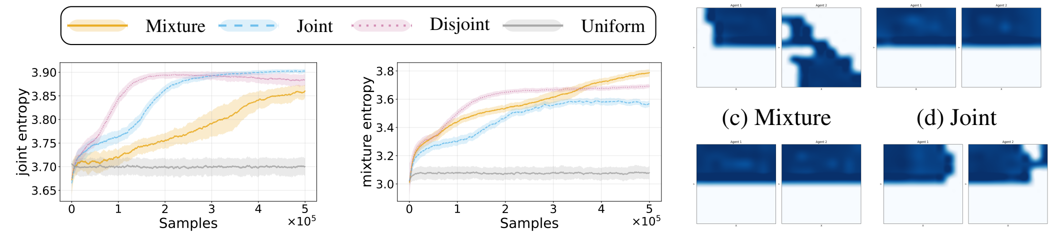
Algorithm: Trust Region Pure Exploration (TRPE)

> **Input**: exploration horizon T, trajectories N, trust-region threshold δ, learning rate η
> Initialize $\boldsymbol{\theta} = (\theta^i)_{i \in [\mathcal{N}]}$
> **for** epoch $= 1, 2, \dots$ until convergence **do**
>   Collect N trajectories with $\pi_{\boldsymbol{\theta}} = (\pi^i_{\theta^i})_{i \in [\mathcal{N}]}$
>   **for** agent $i = 1, 2, \dots$ **concurrently do**
>   Set datasets $\mathcal{D}^i = \{(\mathbf{s}^i_n, \mathbf{a}^i_n), \zeta^n_1\}_{n \in [N]}$
> $h = 0$, $\theta^i_h = \theta^i$
>   **while** $D_{\mathrm{KL}}(\pi^i_{\theta^i_h} \| \pi^i_{\theta^i_0}) \leqslant \delta$ **do**
>   Compute $\hat{\mathcal{L}}^i(\theta^i_h/\theta^i_0)$ via IS.
> $\theta^i_{h+1} = \theta^i_h + \eta \nabla_{\theta^i_h} \hat{\mathcal{L}}^i(\theta^i_h/\theta^i_0)$
> $h \leftarrow h + 1$
>   **end while**
> $\theta^i \leftarrow \theta^i_h$
>   **end for**
> **end for**
> **Output**: joint policy $\pi_{\boldsymbol{\theta}} = (\pi^i_{\theta^i})_{i \in [\mathcal{N}]}$

The main idea then follows from noticing that the surrogate function in Def. 5.1 consists of an Importance Sampling (IS) estimator [30], and it is then possible to optimize it in a fully decentralized and off-policy manner [22, 24]. More specifically, given a pre-specified objective of interest $\zeta_1 \in \{\zeta_1^\infty, \zeta_1^i, \tilde{\zeta}_1\}$, agents sample N trajectories

[4] Indeed, in the case of product state-spaces $\mathcal{S} = \times_{i \in [\mathcal{N}]} \mathcal{S}_i$ the cardinality scales exponentially with the number of agents $|\mathcal{N}|$.

[5] In this scenario, all the bounds of Lemma 4.1 linking different objectives become vacuous.



(a) Joint Entropy (b) Mixture Entropy (c) Mixture (d) Joint (e) Disjoint (f) Uniform

Figure 1: Single-trial Joint and Mixture Entropy induced by different objective optimization along a $T = 50$ horizon. (*Right*) State Distributions of two agents induced by different learned policies. We report the average and 95% confidence interval over 4 runs.

$\{(\mathbf{s}_n, \mathbf{a}_n)\}_{n\in[N]}$ following a (joint) policy with parameters $\boldsymbol{\theta}_0 = (\theta^i_0, \theta^{-i}_0)$. They then compute the values of the objective for each trajectory, building separate datasets $\mathcal{D}^i = \{(\mathbf{s}^i_n, \mathbf{a}^i_n), \zeta^n_1\}_{n\in[N]}$ and using it to compute the Monte-Carlo approximation of the Surrogate Function, namely $\hat{\mathcal{L}}^i(\theta^i_h/\theta^i_0) = \frac{1}{N}\sum_{n\in[N]} \rho^{i,n}_{\theta^i_h/\theta^i_0}\zeta^n_1$, $\rho^{i,n}_{\theta^i_h/\theta^i_0} = \prod_{t\in[T]} \pi^i_{\theta^i_h}(\mathbf{a}^i_n[t]|\mathbf{s}^i_n[t])/\pi^i_{\theta^i_0}(\mathbf{a}^i_n[t]|\mathbf{s}^i_n[t])$, and ζ^n_1 is the plug-in estimator of the entropy based on the empirical measure d_1 [31]. Finally, at each off-policy iteration h, each agent updates its parameter via gradient ascent $\theta^i_{h+1} \leftarrow \theta^i_h + \eta\nabla_{\theta^i_h}\hat{\mathcal{L}}^i(\theta^i_h/\theta^i_0)$ until the trust-region boundary is reached, i.e., when it holds $D_{\mathrm{KL}}(\pi^i_{\overline{\theta}^i}\|\pi^i_{\theta^i}) > \delta$. The pseudo-code of TRPE is reported in Algorithm 1. We remark that even though TRPE is applied to task-agnostic exploration, the algorithmic blueprint does not explicitly require the function $\mathcal{F}$ to be the entropy function and thus it is of independent interest. The interested reader can found some comments of its potential limitations in Appendix C.

6 Empirical Corroboration

In this section, we address the last research question, that is:

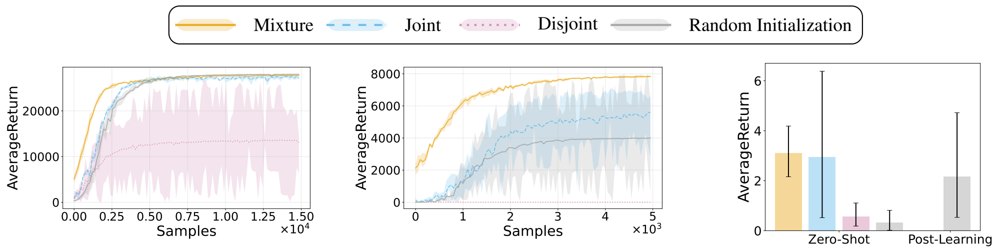
(**Q4**) *Do crucial differences emerge in practice? Does this have an impact on downstream tasks learning?*

by providing empirical corroboration of the findings discussed so far. Especially, we aim to answer the following questions: (**Q4.1**) Is Algorithm 1 actually capable of optimizing finite-trials objectives? (**Q4.2**) Do different objectives enforce different behaviors, as expected from Section 3? (**Q4.3**) Does the *clustering* behavior of mixture objectives play a crucial role? If yes, when and why?

Experimental Domains. The experiments were performed with the aim to illustrate essential features of task-agnostic exploration suggested by the theoretical analysis, and for this reason the domains were selected for being challenging while keeping high interpretability. The first is a notoriously difficult multi-agent exploration task called *secret room* [MPE, 20],[6] referred to as Env. (**i**). In such task, two agents are required to reach a target while navigating over two rooms divided by a door. In order to keep the door open, at least one agent have to remain on a switch. Two switches are located at the corners of the two rooms. The hardness of the task then comes from the need of coordinated exploration, where one agent allows for the exploration of the other. The second is a simpler exploration task yet over a high dimensional state-space, namely a 2-agent instantiation of *Reacher* [MaMuJoCo, 32], referred to as Env. (**ii**). Each agent corresponds to one joint and equipped with decentralized-information policies. In order to allow for the use of plug-in estimator of the entropy [31], each state dimension was discretized over 10 bins.

task-agnostic exploration. As common for the unsupervised RL framework [8, 13, 18, 25], Algorithm 1 was first tested in her ability to optimize for task-agnostic exploration objectives, thus in environments *without rewards*. In Figure 1, we report the results for a short, and thus more challenging, exploration horizon ($T = 50$) over Env. (**i**), as it is far more interpretable. Other experiments with longer horizons or over Env. (**ii**) can be found in Appendix C. Interestingly, at this challenging exploration regime, when looking at the joint entropy in Figure 1a, joint and disjoint objectives perform rather well compared to mixture ones in terms of induced joint entropy, while they fail to address mixture entropy explicitly, as seen in Figure 1b. On the other hand mixture-based

[6] We highlight that all previous efforts in this task employed centralized-information policies. On the other hand, we are interested on the role of the entropic feedback in fostering coordination rather than full-state conditioning, thus we employed decentralized-information policies.



(a) MA-TRPO + TRPE Pre-Training (Env. (**i**), $T = 150$)

(b) MA-TRPO + TRPE Pre-Training (Env. (**i**), $T = 50$)

(c) MA-TRPO + TRPE Pre-Training (Env. (**ii**), $T = 100$)

Figure 2: Effect of pre-training in sparse-reward settings. (*Left*) Policies initialized with either Uniform or TRPE pre-trained policies. (*Right*) Policies initialized with either Zero-Mean or TRPE pre-trained policies. We report the average and 95% c.i. over 4 runs over worst-case goals.

objectives result in optimizing both mixture *and* joint entropy effectively, as one would expect by the bounds in Th. 4.1. By looking at the actual state visitation induced by the trained policies, the difference between the objectives is apparent. While optimizing joint objectives, agents exploit the high-dimensionality of the joint space to induce highly entropic distributions even without exploring the space uniformly via coordination (Fig. 1d); the same outcome happens in disjoint objectives, with which agents focus on over-optimizing over a restricted space loosing any incentive for coordinated exploration (Fig. 1e). On the other hand, mixture objectives enforce a clustering behavior (Fig. 1e) and result in a better efficient exploration.

Policy Pre-Training via Task-Agnostic Exploration. More interestingly, we tested the effect of pre-training policies via task-agnostic exploration as a way to alleviate the well-known hardness of *sparse-reward* settings. In order to do so, we employed a multi-agent counterpart of the TRPO algorithm [36] with different pre-trained policies. First, we investigated the effect on the learning curve in the hard-exploration task of Env. (**i**) under long horizons ($T = 150$), with a *worst-case goal* set on the opposite corner of the closed room. Pre-training via mixture objectives still lead to a faster learning compared to initializing the policy with a uniform distribution. On the other hand, joint objective pre-training did not lead to substantial improvements over standard initializations. More interestingly, when extremely short horizons were taken into account ($T = 50$) the difference became appalling, as shown in Fig. 2a: pre-training via mixture-based objectives leaded to faster learning and higher performances, while pre-training via disjoint objectives turned out to be even *harmful* (Fig. 2b).[7] Finally, we tested the zero-shot capabilities of policy pre-training on the simpler but high dimensional exploration task of Env. (**ii**), where the goal was sampled randomly between *worst-case positions* at the boundaries of the region reachable by the arm. As shown in Fig. 3p, both joint and mixture were able to guarantee zero-shot performances via pre-training compatible with MA-TRPO after learning over $2e4$ samples, while disjoint objectives were not. On the other hand, pre-training with joint objectives showed an extremely high-variance, leading to worst-case performances not better than the ones of random initialization. Mixture objectives on the other hand showed higher stability in guaranteeing compelling zero-shot performance.

Takeaways. Overall, all of the experimental questions were answered: (**Q4.1**) Algorithm 1 is indeed able to optimize for finite-trial objectives; (**Q4.2**) **Mixture objectives enforce coordination**, essential when high efficiency is required, while joint or disjoint objectives may fail to lead to relevant solutions because of under or over optimization; (**Q4.3**) **The efficient coordination** through mixture objectives enforces the ability of **pre-training via task-agnostic exploration** to lead to **faster and better training** and even **zero-shot generalization**.

7 Conclusions and Perspectives

In this paper, we introduce a principled framework for unsupervised pre-training in MARL via task-agnostic exploration. First, we show formalize it via several different objectives, linked through performance bounds. Moreover, we theoretically characterize how the problem, even when tractable in theory, is non-trivial in practice. Then, we design a practical algorithm and we use it in a set of experiments to confirm the expected superiority of mixture objectives in practice, due to their ability to enforce efficient exploration over short horizons. Future works can build over our results in many directions, for instance by developing scalable algorithms for continuous domains.

[7] This is probably due to pre-training almost deterministic policies, as shown in Fig. 4 in Appendix C.

References

[1] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. Bellemare. Reincarnating Reinforcement Learning: Reusing prior computation to accelerate progress. *Advances in neural information processing systems*, 35:28955–28971, 2022.

[2] S. V. Albrecht, F. Christianos, and L. Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024. URL `https://www.marl-book.com`.

[3] D. P. Bertsekas and J. N. Tsitsiklis. Introduction to probability (athena scientific, belmont, ma). *EKLER Ek A: Sıralı Istatistik Ek B: Integrallerin Sayısal Hesabı Ek B*, 1, 2002.

[4] A. Breitenmoser, M. Schwager, J.-C. Metzger, R. Siegwart, and D. Rus. Voronoi coverage of non-convex environments with a group of networked robots. In *2010 IEEE international conference on robotics and automation*, pages 4982–4989. IEEE, 2010.

[5] V. De Paola, R. Zamboni, M. Mutti, and M. Restelli. Enhancing diversity in parallel agents: A maximum state entropy exploration story. *arXiv preprint arXiv:2505.01336*, 2025.

[6] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. Benchmarking deep Reinforcement Learning for continuous control. In *International conference on machine learning*, pages 1329–1338. PMLR, 2016.

[7] I. Gemp, A. Haupt, L. Marris, S. Liu, and G. Piliouras. Convex Markov Games: A framework for creativity, imitation, fairness, and safety in multiagent learning, 2025. URL `https://arxiv.org/abs/2410.16600`.

[8] E. Hazan, S. Kakade, K. Singh, and A. Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691, 2019.

[9] H. Hu, A. Lerer, A. Peysakhovich, and J. Foerster. "Other-Play" for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR, 2020.

[10] S. Iqbal and F. Sha. Coordinated exploration via intrinsic rewards for multi-agent Reinforcement Learning. *arXiv preprint arXiv:1905.12127*, 2019.

[11] M. B. Johanson, E. Hughes, F. Timbers, and J. Z. Leibo. Emergent bartering behaviour in Multi-Agent Reinforcement Learning, 2022. URL `https://arxiv.org/abs/2205.06760`.

[12] A. Kolchinsky and B. Tracey. Estimating mixture entropy with pairwise distances. *Entropy*, 19 (7):361, July 2017. ISSN 1099-4300. doi: 10.3390/e19070361. URL `http://dx.doi.org/10.3390/e19070361`.

[13] M. Laskin, D. Yarats, H. Liu, K. Lee, A. Zhan, K. Lu, C. Cang, L. Pinto, and P. Abbeel. URLB: Unsupervised Reinforcement Learning benchmark. *Advances in Neural Information Processing Systems (Datasets & Benchmarks)*, 2021.

[14] L. Lee, B. Eysenbach, E. Parisotto, E. Xing, S. Levine, and R. Salakhutdinov. Efficient exploration via state marginal matching, 2020. URL `https://arxiv.org/abs/1906.05274`.

[15] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras. Global convergence of Multi-Agent Policy Gradient in Markov Potential Games, 2021. URL `https://arxiv.org/abs/2106.01969`.

[16] M. L. Littman. Markov Games as a framework for multi-agent Reinforcement Learning. In W. W. Cohen and H. Hirsh, editors, *Machine Learning Proceedings 1994*, pages 157–163. Morgan Kaufmann, San Francisco (CA), 1994. ISBN 978-1-55860-335-6. doi: https://doi.org/10.1016/B978-1-55860-335-6.50027-1. URL `https://www.sciencedirect.com/science/article/pii/B9781558603356500271`.

[17] H. Liu and P. Abbeel. APS: Active pretraining with successor features. In *International Conference on Machine Learning*, 2021.

[18] H. Liu and P. Abbeel. Behavior from the void: unsupervised active pre-training. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

[19] I.-J. Liu, U. Jain, R. A. Yeh, and A. Schwing. Cooperative exploration for multi-agent deep Reinforcement Learning. In *International conference on machine learning*, pages 6826–6836. PMLR, 2021.

[20] I.-J. Liu, U. Jain, R. A. Yeh, and A. Schwing. Cooperative exploration for multi-agent deep reinforcement learning. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6826–6836. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/liu21j.html`.

[21] J. McLurkin and D. Yamins. Dynamic task assignment in robot swarms. In *Robotics: Science and Systems*, volume 8. Cambridge, USA, 2005.

[22] A. M. Metelli, M. Papini, N. Montali, and M. Restelli. Importance sampling techniques for policy optimization. *Journal of Machine Learning Research*, 21(141):1–75, 2020. URL `http://jmlr.org/papers/v21/20-124.html`.

[23] R. Mirsky, I. Carlucho, A. Rahman, E. Fosong, W. Macke, M. Sridharan, P. Stone, and S. V. Albrecht. A survey of ad hoc teamwork research. In *European conference on multi-agent systems*, pages 275–293. Springer, 2022.

[24] M. Mutti and M. Restelli. An intrinsically-motivated approach for learning highly exploring and fast mixing policies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (04):5232–5239, Apr. 2020. doi: 10.1609/aaai.v34i04.5968. URL `https://ojs.aaai.org/index.php/AAAI/article/view/5968`.

[25] M. Mutti, L. Pratissoli, and M. Restelli. Task-agnostic exploration via policy gradient of a non-parametric state entropy estimate. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9028–9036, 2021.

[26] M. Mutti, R. De Santi, P. De Bartolomeis, and M. Restelli. Challenging common assumptions in convex Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35: 4489–4502, 2022.

[27] M. Mutti, R. De Santi, and M. Restelli. The importance of non-Markovianity in maximum state entropy exploration. In *International Conference on Machine Learning*, pages 16223–16239, 2022.

[28] M. Mutti, M. Mancassola, and M. Restelli. Unsupervised Reinforcement Learning in multiple environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[29] M. Mutti, R. D. Santi, P. D. Bartolomeis, and M. Restelli. Convex Reinforcement Learning in Finite Trials. *Journal of Machine Learning Research*, 24(250):1–42, 2023. URL `http://jmlr.org/papers/v24/22-1514.html`.

[30] A. B. Owen. *Monte Carlo theory, methods and examples*. `https://artowen.su.domains/mc/`, 2013.

[31] L. Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, June 2003. ISSN 0899-7667. doi: 10.1162/089976603321780272. URL `https://doi.org/10.1162/089976603321780272`.

[32] B. Peng, T. Rashid, C. Schroeder de Witt, P.-A. Kamienny, P. Torr, W. Böhmer, and S. Whiteson. FACMAC: Factored multi-agent centralised policy gradients. *Advances in Neural Information Processing Systems*, 34:12208–12221, 2021.

[33] J. Perolat, B. De Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony, et al. Mastering the game of stratego with model-free multiagent Reinforcement Learning. *Science*, 378(6623):990–996, 2022.

[34] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[35] M. Samvelyan, T. Rashid, C. S. De Witt, G. Farquhar, N. Nardelli, T. G. Rudner, C.-M. Hung, P. H. Torr, J. Foerster, and S. Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.

[36] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel. Trust region policy optimization, 2017. URL `https://arxiv.org/abs/1502.05477`.

[37] Y. Seo, L. Chen, J. Shin, H. Lee, P. Abbeel, and K. Lee. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*, pages 9443–9454. PMLR, 2021.

[38] D. Tiapkin, D. Belomestny, D. Calandriello, E. Moulines, R. Munos, A. Naumov, P. Perrault, Y. Tang, M. Valko, and P. Menard. Fast rates for maximum entropy exploration. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34161–34221. PMLR, 23–29 Jul 2023.

[39] M. Toquebiau, N. Bredeche, F. Benamar, and J.-Y. Jun. Joint intrinsic motivation for coordinated exploration in multi-agent deep Reinforcement Learning. *arXiv preprint arXiv:2402.03972*, 2024.

[40] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. 2003. URL `https://api.semanticscholar.org/CorpusID:12164823`.

[41] P. Xu, J. Zhang, and K. Huang. Population-based diverse exploration for sparse-reward multi-agent tasks. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 283–291, 2024.

[42] H. Yang, D. Shi, C. Zhao, G. Xie, and S. Yang. Ciexplore: Curiosity and influence-based exploration in multi-agent cooperative scenarios with sparse rewards. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2321–2330, 2021.

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

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

[45] T. Zahavy, B. O'Donoghue, G. Desjardins, and S. Singh. Reward is enough for convex MDPs. *Advances in Neural Information Processing Systems*, 34:25746–25759, 2021.

[46] R. Zamboni, D. Cirino, M. Restelli, and M. Mutti. How to explore with belief: State entropy maximization in POMDPs. In *International Conference on Machine Learning*, 2024.

[47] R. Zamboni, D. Cirino, M. Restelli, and M. Mutti. The limits of pure exploration in POMDPs: When the observation entropy is enough. *RLJ*, 2:676–692, 2024.

[48] J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang. Variational policy gradient method for Reinforcement Learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

[49] S. Zhang, J. Cao, L. Yuan, Y. Yu, and D.-C. Zhan. Self-motivated multi-agent exploration. *arXiv preprint arXiv:2301.02083*, 2023.

[50] T. Zhang, P. Rashidinejad, J. Jiao, Y. Tian, J. E. Gonzalez, and S. Russell. MADE: Exploration via maximizing deviation from explored regions. *Advances in Neural Information Processing Systems*, 34:9663–9680, 2021.

A Related Works

Below, we summarize the most relevant work investigating related concepts.

Entropic Functionals in MARL. A large plethora of works on both swarm robotics [21, 4] and multi-agent intrinsic motivation [among others, 10, 42, 50, 49, 41, 39] investigated the effects of employing entropic-like functions to boost exploration and performances in down-stream tasks. Importantly, these works are of empirical nature, and they do not investigate the theoretical properties of cMGs or multi-agent task-agnostic exploration, nor they propose algorithms able to pre-train policies *without access* to extrinsic rewards.[8] Finally, while a similar notion of cMGs was proposed in [7], their contributions are focused on the existence of equilibria and performances of centralized algorithms over infinite-trials objectives.

task-agnostic exploration and Policy Optimization. Entropy maximization in MDPs was first introduced in [8] and then investigated extensively in various works [e.g., 24, 25, 27, 28, 18, 17, 37, 43, 38, 47, 46, 5]. Finally, we considered an IS policy gradient estimator inspired by the work of [22], yet with other forms of IS estimators, such as non-parametric k-NN estimators proposed in [25].

B Proofs of the Main Theoretical Results

In this Section, we report the full proofing steps of the Theorems and Lemmas in the main paper. In general, thought this paper, we make the following assumption for the objective function $\mathcal{F}$:

Assumption B.1 (Lipschitz). *A function $\mathcal{F} : \mathcal{A} \to \mathbb{R}$ is Lipschitz-continuous for some constant $L < \infty$, or L-Lipschitz for short, if it holds $|\mathcal{F}(x) - \mathcal{F}(y)| \leqslant L\|x - y\|_1,\ \forall (x, y) \in \mathcal{A}^2$.*

Here follow the proofs of the main results.

Lemma 4.1 (Entropy Mismatch). *For every cMG $\mathcal{M}_H$, for a fixed (joint) policy $\pi = (\pi^i)_{i\in\mathcal{N}}$ the infinite-trials objectives are ordered according to:*

$$\frac{H(d^\pi)}{|\mathcal{N}|} \leqslant \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} H(d_i^\pi) \leqslant H(\tilde{d}^\pi) \leqslant \sup_{i\in[\mathcal{N}]} H(d_i^\pi) + \log(|\mathcal{N}|) \leqslant H(d^\pi) + \log(|\mathcal{N}|)$$

Proof. The bounds follow directly from simple yet fundamental relationships between entropies of joint, marginal and mixture distributions which can be found in [31, 12], in particular:

$$\begin{aligned}\frac{1}{|\mathcal{N}|}H(d^\pi) \leqslant \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} H(d_i^\pi) \overset{\text{(a)}}{\leqslant} H(\tilde{d}^\pi) &\overset{\text{(b)}}{\leqslant} \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} H(d_i^\pi) + \log(|\mathcal{N}|)\\ &\overset{\text{(c)}}{\leqslant} \sup_{i\in[\mathcal{N}]} H(d_i^\pi) + \log(|\mathcal{N}|) \leqslant H(d^\pi) + \log(|\mathcal{N}|)\end{aligned}$$

where step (a) and (b) use the fact that $\tilde{d}^\pi(s) := \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} d_i^\pi(s)$ is a uniform mixture over the agents, whose distribution over the weights has entropy $\log(|\mathcal{N}|)$, so as we can apply the bounds from [12]. Step (c) uses the fact that $H(d^\pi) = \sum_{i\in[\mathcal{N}]} H(d_i^\pi|d_{<i}^\pi)$, then taking the supremum as first i it follows that $\sup_{i\in[\mathcal{N}]} H(d_i^\pi) = H(d^\pi) - \sum_{j\in[\mathcal{N}]>i} H(d_j^\pi|d_{<j}^\pi, d_i^\pi) \leqslant H(d^\pi)$ due to non-negativity of entropy. □

Theorem 4.2 (Finite-Trials Mismatch in cMGs). *For every cMG $\mathcal{M}_\mathcal{F}$ equipped with a L-Lipschitz function $\mathcal{F}$, let $K \in \mathbb{N}^+$ be a number of evaluation episodes/trials, and let $\delta \in (0, 1]$ be a confidence level, then for any (joint) policy $\pi = (\pi^i \in \Pi^i)_{i\in[\mathcal{N}]}$, it holds that the mismatches $err(\pi) := |\zeta_K(\pi) - \zeta_\infty(\pi)|, err^i(\pi) := |\zeta_K^i(\pi) - \zeta_\infty^i(\pi)|, \tilde{err}(\pi) := |\tilde{\zeta}_K(\pi) - \tilde{\zeta}_\infty(\pi)|$ are bounded as:*

$$err(\pi) \leqslant LT\sqrt{\frac{2|\mathcal{S}|\log(2T/\delta)}{K}}, \quad err^i(\pi) \leqslant LT\sqrt{\frac{2|\tilde{\mathcal{S}}|\log(2T/\delta)}{K}}, \quad \tilde{err}(\pi) \leqslant LT\sqrt{\frac{2|\tilde{\mathcal{S}}|\log(2T/\delta)}{|\mathcal{N}|K}}.$$

[8]The interested reader can refer to [25, 18] for an extensive investigation of the fundamental differences between intrinsic motivation and task-agnostic exploration.

Proof. For the general proof structure, we adapt the steps of [26] for cMDPs to the different objectives possible in cMGs. Let us start by considering joint objectives, then:

$$\begin{aligned}\big|\zeta_K(\pi) - \zeta_\infty(\pi)\big| &= \Big|\mathop{\mathbb{E}}_{d_K\sim p_K^\pi}[\mathcal{F}(d_K)] - \mathcal{F}(d^\pi)\Big| \leqslant \mathop{\mathbb{E}}_{d_K\sim p_K^\pi}[|\mathcal{F}(d_K) - \mathcal{F}(d^\pi)|] \\ &\overset{(a)}{\leqslant} \mathop{\mathbb{E}}_{d_K\sim p_K^\pi}[L\,\|d_K - d^\pi\|_1] \leqslant L \mathop{\mathbb{E}}_{d_K\sim p_K^\pi}[\|d_K - d^\pi\|_1] \\ &\overset{(b)}{\leqslant} L \mathop{\mathbb{E}}_{d_K\sim p_K^\pi}\left[\max_{t\in[T]}\|d_{K,t} - d_t^\pi\|_1\right],\end{aligned}$$

where in step (a) we apply the Lipschitz assumption on $\mathcal{F}$ to write and in step (b) we apply a maximization over the episode's step by noting that $d_K = \frac{1}{T}\sum_{t\in[T]} d_{K,t}$ and $d^\pi = \frac{1}{T}\sum_{t\in[T]} d_t^\pi$. We then apply bounds in high probability

$$\begin{aligned}Pr\Big(\max_{t\in[T]}\|d_{K,t} - d_t^\pi\|_1 \geqslant \epsilon\Big) &\leqslant Pr\Big(\bigcup_t \|d_{K,t} - d_t^\pi\|_1 \geqslant \epsilon\Big) \\ &\overset{(c)}{\leqslant} \sum_t Pr\Big(\|d_{K,t} - d_t^\pi\|_1 \geqslant \epsilon\Big) \\ &\leqslant T\; Pr\Big(\|d_{K,t} - d_t^\pi\|_1 \geqslant \epsilon\Big),\end{aligned}$$

with $\epsilon > 0$ and in step (c) we applied a union bound. We then consider standard concentration inequalities for empirical distributions [40] so to obtain the final bound

$$Pr\left(\|d_{K,t} - d_t^\pi\|_1 \geqslant \sqrt{\frac{2|\mathcal{S}|\log(2/\delta')}{K}}\right) \leqslant \delta'. \tag{4}$$

By setting $\delta' = \delta/T$, and then plugging the empirical concentration inequality, we have that with probability at least $1-\delta$

$$\big|\zeta_K(\pi) - \zeta_\infty(\pi)\big| \leqslant LT\sqrt{\frac{2|\mathcal{S}|\log(2T/\delta)}{K}},$$

which concludes the proof for joint objectives.

The proof for disjoint objectives follows the same rational by bounding each per-agent term separately and after noticing that due to Assumption 3.1, the resulting bounds get simplified in the overall averaging. As for mixture objectives, the only core difference is after step (b), where $\tilde{d}_K$ takes the place of d_K and $\tilde{d^\pi}$ of d^π. The remaining steps follow the same logic, out of noticing that the empirical distribution with respect to $\tilde{d^\pi}$ is taken with respect $|\mathcal{N}|K$ samples in total. Both the two bounds then take into account that the support of the empirical distributions have size $|\tilde{\mathcal{S}}|$ and not $|\mathcal{S}|$. □

B.1 Policy Gradient in cMGs with Infinite-Trials Formulations.

In this Section, we analyze policy search for the infinite-trials joint problem ζ_∞ of Eq. (1), via projected gradient ascent over parametrized policies, providing in Th. B.6 the formal counterpart of Fact 4.1 in the Main paper. As a side note, all of the following results hold for the (infinite-trials) mixture objective $\tilde{\zeta}_\infty$ of Eq. (3). We will consider the class of parametrized policies with parameters $\theta_i \in \Theta_i \subset \mathbb{R}^d$, with the joint policy then defined as $\pi_\theta, \theta \in \Theta = \times_{i\in[\mathcal{N}]}\Theta_i$. Additionally, we will focus on the computational complexity only, by assuming access to the exact gradient. The study of statistical complexity surpasses the scope of the current work. We define the **(independent) Policy Gradient Ascent** (PGA) update as:

$$\theta_i^{k+1} = \mathop{\arg\max}_{\theta_i\in\Theta_i} \zeta_\infty(\pi_{\theta^k}) + \big\langle\nabla_{\theta_i}\zeta_\infty(\pi_{\theta^k}), \theta_i - \theta_i^k\big\rangle - \frac{1}{2\eta}\|\theta_i - \theta_i^k\|^2 = \Pi_{\Theta_i}\big\{\theta_i^k + \eta\nabla_{\theta_i}\zeta_\infty(\pi_{\theta^k})\big\} \tag{5}$$

where $\Pi_{\Theta_i}\{\cdot\}$ denotes Euclidean projection onto Θ_i, and equivalence holds by the convexity of Θ_i. The classes of policies that allow for this condition to be true will be discussed shortly.

In general the overall proof is built of three main steps, shared with the theory of Potential Markov Games [15]: (i) prove the existence of well behaved stationary points; (ii) prove that performing independent policy gradient is equivalent to perform joint policy gradient; (iii) prove that the (joint) PGA update converges to the stationary points via single-agent like analysis. In order to derive the subsequent convergence proof, we will make the following assumptions:

Assumption B.1. Define the quantity $\lambda(\theta) := d^{\pi_\theta}$, then:
(i). $\lambda(\cdot)$ forms a bijection between Θ and $\lambda(\Theta)$, where Θ and $\lambda(\Theta)$ are closed and convex.
(ii). The Jacobian matrix $\nabla_\theta \lambda(\theta)$ is Lipschitz continuous in Θ.
(iii). Denote $g(\cdot) := \lambda^{-1}(\cdot)$ as the inverse mapping of $\lambda(\cdot)$. Then there exists $\ell_\theta > 0$ s.t. $\|g(\lambda) - g(\lambda')| \leqslant \ell_\theta \|\lambda - \lambda'\|$ for some norm $\|\cdot\|$ and for all $\lambda, \lambda' \in \lambda(\Theta)$.

Assumption B.2. There exists $L > 0$ such that the gradient $\nabla_\theta \zeta_\infty(\pi_\theta)$ is L-Lipschitz.

Assumption B.3. The agents have access to a gradient oracle $\mathcal{O}(\cdot)$ that returns $\nabla_{\theta_i} \zeta_\infty(\pi_\theta)$ for any deployed joint policy π_θ.

On the Validity of Assumption B.1. This set of assumptions enforces the objective $\zeta_\infty(\pi_\theta)$ to be well-behaved with respect to θ even if non-convex in general, and will allow for a rather strong result. Yet, the assumptions are known to be true for directly parametrized policies over the whole support of the distribution d^π [48], and as a result they implicitly require agents to employ policies conditioned over the full state-space $\mathcal{S}$. Fortunately enough, they also guarantee Θ to be convex.

Lemma B.4 (**(i)** Global optimality of stationary policies [48])**.** *Suppose Assumption B.1 holds, and $\mathcal{F}$ is a concave, and continuous function defined in an open neighborhood containing $\lambda(\Theta)$. Let θ^* be a first-order stationary point of problem* (1)*, i.e.,*

$$\exists u^* \in \hat{\partial}(\mathcal{F} \circ \lambda)(\theta^*), \quad s.t. \quad \langle u^*, \theta - \theta^* \rangle \leqslant 0 \qquad for \qquad \forall \theta \in \Theta. \tag{6}$$

Then θ^* is a globally optimal solution of problem (1)*.*

This result characterizes the optimality of stationary points for Eq. (1). Furthermore, we know from [15] that stationary points of the objective are Nash Equilibria.

Lemma B.5 (**(ii)** Projection Operator [15])**.** *Let $\theta := (\theta_1, ..., \theta_\mathcal{N})$ be the parameter profile for all agents and use the update of Eq.* (5) *over a non-disjoint infinite-trials objective. Then, it holds that*

$$\Pi_\Theta\{\theta^k + \eta \nabla_\theta \zeta_\infty(\pi_{\theta^k})\} = \Big(\Pi_{\Theta_i}\{\theta_i^k + \eta \nabla_{\theta_i} \zeta_\infty(\pi_{\theta^k})\}\Big)_{i \in [\mathcal{N}]}$$

This result will only be used for the sake of the convergence analysis, since it allows to analyze independent updates as joint updates over a single objective. The following Theorem is the formal counterpart of Fact 4.1 and it is a direct adaptation to the multi-agent case of the single-agent proof by [48], by exploiting the previous result.

Theorem B.6 (**(iii)** Convergence rate of independent PGA to stationary points (Formal Fact 4.1))**.** *Let Assumptions B.1 and B.2 hold. Denote $D_\lambda := \max_{\lambda, \lambda' \in \lambda(\Theta)} \|\lambda - \lambda'\|$ as defined in Assumption B.1(iii). Then the independent policy gradient update* (5) *with $\eta = 1/L$ satisfies for all k with respect to a stationary (joint) policy π_{θ^*} the following*

$$\zeta_\infty(\pi_{\theta^*}) - \zeta_\infty(\pi_{\theta^k}) \leqslant \frac{4L\ell_\theta^2 D_\lambda^2}{k+1}.$$

Proof. First, the Lipschitz continuity in Assumption B.2 indicates that

$$\big|\zeta_\infty(\lambda(\theta)) - \zeta_\infty(\lambda(\theta^k)) - \langle \nabla_\theta \zeta_\infty(\lambda(\theta^k)), \theta - \theta^k \rangle\big| \leqslant \frac{L}{2}\|\theta - \theta^k\|^2.$$

Consequently, for any $\theta \in \Theta$ we have the ascent property:

$$\zeta_\infty(\lambda(\theta)) \geqslant \zeta_\infty(\lambda(\theta^k)) + \langle \nabla_\theta \zeta_\infty(\lambda(\theta^k)), \theta - \theta^k \rangle - \frac{L}{2}\|\theta - \theta^k\|^2 \geqslant \zeta_\infty(\lambda(\theta)) - L\|\theta - \theta^k\|^2. \tag{7}$$

The optimality condition in the policy update rule (5) coupled with the result of Lemma B.5 allows us to follow the same rational as [48]. We will report their proof structure after this step for completeness.

$$\begin{aligned}
\zeta_\infty(\lambda(\theta^{k+1})) &\geqslant \zeta_\infty(\lambda(\theta^k)) + \langle \nabla_\theta \zeta_\infty(\lambda(\theta^k)), \theta^{k+1} - \theta^k \rangle - \frac{L}{2}\|\theta^{k+1} - \theta^k\|^2 \\
&= \max_{\theta \in \Theta} \zeta_\infty(\lambda(\theta^k)) + \langle \nabla_\theta \zeta_\infty(\lambda(\theta^k)), \theta - \theta^k \rangle - \frac{L}{2}\|\theta - \theta^k\|^2 \\
&\overset{\text{(a)}}{\geqslant} \max_{\theta \in \Theta} \zeta_\infty(\lambda(\theta)) - L\|\theta - \theta^k\|^2 \\
&\overset{\text{(b)}}{\geqslant} \max_{\alpha \in [0,1]} \left\{ \zeta_\infty(\lambda(\theta_\alpha)) - L\|\theta_\alpha - \theta^k\|^2 : \theta_\alpha = g(\alpha\lambda(\theta^*) + (1-\alpha)\lambda(\theta^k)) \right\}. \qquad (8)
\end{aligned}$$

where step (a) follows from (7) and step (b) uses the convexity of $\lambda(\Theta)$. Then, by the concavity of ζ_∞ and the fact that the composition $\lambda \circ g = id$ due to Assumption B.1(i), we have that:

$$\zeta_\infty(\lambda(\theta_\alpha)) = \zeta_\infty(\alpha\lambda(\theta^*) + (1-\alpha)\lambda(\theta^k)) \geqslant \alpha\zeta_\infty(\lambda(\theta^*)) + (1-\alpha)\zeta_\infty(\lambda(\theta^k)).$$

Moreover, due to Assumption B.1(iii) we have that:

$$\begin{aligned}
\|\theta_\alpha - \theta^k\|^2 &= \|g(\alpha\lambda(\theta^*) + (1-\alpha)\lambda(\theta^k)) - g(\lambda(\theta^k))\|^2 \qquad (9) \\
&\leqslant \alpha^2 \ell_\theta^2 \|\lambda(\theta^*) - \lambda(\theta^k)\|^2 \\
&\leqslant \alpha^2 \ell_\theta^2 D_\lambda^2.
\end{aligned}$$

From which we get

$$\begin{aligned}
&\zeta_\infty(\lambda(\theta^*)) - \zeta_\infty(\lambda(\theta^{k+1})) \\
&\quad \leqslant \min_{\alpha \in [0,1]} \left\{ \zeta_\infty(\lambda(\theta^*)) - \zeta_\infty(\lambda(\theta_\alpha)) + L\|\theta_\alpha - \theta^k\|^2 : \theta_\alpha = g(\alpha\lambda(\theta^*) + (1-\alpha)\lambda(\theta^k)) \right\} \\
&\quad \leqslant \min_{\alpha \in [0,1]} (1-\alpha)\big(\zeta_\infty(\lambda(\theta^*)) - \zeta_\infty(\lambda(\theta^k))\big) + \alpha^2 L \ell_\theta^2 D_\lambda^2. \qquad (10)
\end{aligned}$$

We define $\Lambda(\pi_\theta) := \lambda(\theta)$, then $\alpha_k = \frac{\zeta_\infty(\Lambda(\pi^*)) - \zeta_\infty(\Lambda(\pi^k))}{2L\ell_\theta^2 D_\lambda^2} \geqslant 0$, which is the minimizer of the RHS of (10) as long as it satisfies $\alpha_k \leqslant 1$. Now, we claim the following: If $\alpha_k \geqslant 1$ then $\alpha_{k+1} < 1$. Further, if $\alpha_k < 1$ then $\alpha_{k+1} \leqslant \alpha_k$. The two claims together mean that $(\alpha_k)_k$ is decreasing and all α_k are in $[0, 1)$ except perhaps α_0.

To prove the first of the two claims, assume $\alpha_k \geqslant 1$. This implies that $\zeta_\infty(\Lambda(\pi^*)) - \zeta_\infty(\Lambda(\pi^k)) \geqslant 2L\ell_\theta^2 D_\lambda^2$. Hence, choosing $\alpha = 1$ in (10), we get

$$\zeta_\infty(\lambda(\theta^*)) - \zeta_\infty(\lambda(\theta^k)) \leqslant L\ell_\theta^2 D_\lambda^2$$

which implies that $\alpha_{k+1} \leqslant 1/2 < 1$. To prove the second claim, we plug α_k into (10) to get

$$\zeta_\infty(\lambda(\theta^*)) - \zeta_\infty(\lambda(\theta^{k+1})) \leqslant \left(1 - \frac{\zeta_\infty(\lambda(\theta^*)) - \zeta_\infty(\lambda(\theta^k))}{4L\ell_\theta^2 D_\lambda^2}\right)(\zeta_\infty(\lambda(\theta^*)) - \zeta_\infty(\lambda(\theta^k))),$$

which shows that $\alpha_{k+1} \leqslant \alpha_k$ as required.

Now, by our preceding discussion, for $k = 1, 2, \dots$ the previous recursion holds. Using the definition of α_k, we rewrite this in the equivalent form

$$\frac{\alpha_{k+1}}{2} \leqslant \left(1 - \frac{\alpha_k}{2}\right) \cdot \frac{\alpha_k}{2}.$$

By rearranging the preceding expressions and algebraic manipulations, we obtain

$$\frac{2}{\alpha_{k+1}} \geqslant \frac{1}{\left(1 - \frac{\alpha_k}{2}\right) \cdot \frac{\alpha_k}{2}} = \frac{2}{\alpha_k} + \frac{1}{1 - \frac{\alpha_k}{2}} \geqslant \frac{2}{\alpha_k} + 1.$$

For simplicity assume that $\alpha_0 < 1$ also holds. Then, $\frac{2}{\alpha_k} \geqslant \frac{2}{\alpha_0} + k$, and consequenlty

$$\zeta_\infty(\lambda(\theta^*)) - \zeta_\infty(\lambda(\theta^k)) \leqslant \frac{\zeta_\infty(\lambda(\theta^*)) - \zeta_\infty(\lambda(\theta^0))}{1 + \frac{\zeta_\infty(\lambda(\theta^*)) - \zeta_\infty(\lambda(\theta^0))}{4L\ell_\theta^2 D_\lambda^2} \cdot k} \leqslant \frac{4L\ell_\theta^2 D_\lambda^2}{k}.$$

A similar analysis holds when $\alpha_0 > 1$. Combining these two gives that $\zeta_\infty(\Lambda(\pi^*)) - \zeta_\infty(\Lambda(\pi^k)) \leqslant \frac{4L\ell_\theta^2 D_\lambda^2}{k+1}$ no matter the value of α_0, which proves the result. □

B.2 The Use of Markovian and Non-Markovian Policies in cMGs with Finite-Trials Formulations.

The following result describes how in cMGs, as for cMDPs, Non-Markovian policies are the right policy class to employ to guarantee well-behaved results.

Lemma B.1 (Sufficiency of Disjoint Non-Markvoian Policies)**.** *For every cMG $\mathcal{M}$ there exist a joint policy $\pi^\star = (\pi^{\star,i})_{i\in\mathcal{N}}$, with $\pi^{\star,i} \in \Delta_{\mathcal{S}^T}^{\mathcal{A}^i}$ being a deterministic Non-Markovian policy, that is a Nash Equilibrium for non-Disjoint single-trial objectives, for $K = 1$.*

Proof. The proof builds over a straight reduction. We build from the original MG $\mathcal{M}$ a temporally extended Markov Game $\tilde{\mathcal{M}} = (\mathcal{N}, \tilde{\mathcal{S}}, \mathcal{A}, \mathbb{P}, r, \mu, T)$. A state $\tilde{s}$ is defined for each history that can be induced, i.e., $\tilde{s} \in \tilde{\mathcal{S}} \iff \mathbf{s} \in \mathcal{S}^T$. We keep the other objects equivalent, where for the extended transition model we solely consider the last state in the history to define the conditional probability to the next history. We introduce a common reward function across all the agents $r : \tilde{\mathcal{S}} \to \mathbb{R}$ such that $r(\tilde{s}) = H(d(\tilde{s}))$ for joint objectives and $r(\tilde{s}) = (1/N)\sum_{i\in[\mathcal{N}]} H(d_i(\tilde{s}_i))$ for mixture objectives, for all the histories of length T and 0 otherwise. We now know that according to [Theorem 3.1, 15] there exists a deterministic Markovian policy $\tilde{\pi}^\star = (\tilde{\pi}^i)_{i\in\mathcal{N}}, \tilde{\pi}^i \in \Delta_{\tilde{\mathcal{S}}}^{\mathcal{A}_i}$ that is a Nash Equilibrium for $\tilde{\mathcal{M}}$. Since $\tilde{s}$ corresponds to the set of histories of the original game, $\tilde{\pi}^\star$ maps to a non-Markovian policy in it. Finally, it is straightforward to notice that the NE of $\tilde{\pi}^\star$ for $\tilde{\mathcal{M}}$ implies the NE of $\tilde{\pi}^\star$ for the original cMG $\mathcal{M}$. □

The previous result implicitly asks for policies conditioned over the joint state space, as happened for infinite-trials objectives as well. Interestingly, finite-trials objectives allow for a further characterization of how an optimal Markovian policy would behave when conditioned on the per-agent states only:

Lemma B.7 (Behavior of Optimal Markovian Decentralized Policies)**.** *Let $\pi_{NM} = (\pi^i_{NM} \in \Delta_{\mathcal{S}^T}^{\mathcal{A}^i})_{i\in[\mathcal{N}]}$ an optimal deterministic non-Markovian centralized policy and $\bar{\pi}_M = (\bar{\pi}^i_M \in \Delta_{\mathcal{S}}^{\mathcal{A}^i})_{i\in[\mathcal{N}]}$ the optimal Markovian centralized policy, namely $\bar{\pi}_M = \arg\max_{\pi=(\pi^i\in\Delta_{\mathcal{S}}^{\mathcal{A}^i})_{i\in[\mathcal{N}]}} \zeta_1(\pi)$. For a fixed sequence $\mathbf{s}_t \in \mathcal{S}^t$ ending in state $s = (s_i, s_{-i})$, the variance of the event of the optimal Markovian decentralized policy $\pi_M = (\pi^i_M \in \Delta_{\mathcal{S}_i}^{\mathcal{A}^i})_{i\in[\mathcal{N}]}$ taking $a^* = \pi_{NM}(\cdot|\mathbf{s}_t) = \bar{\pi}_M(\cdot|s,t)$ in s_i at step t is given by*

$$\mathbb{V}\text{ar}\left[\mathcal{B}(\pi_M(a^*|s_i,t))\right] = \underset{\mathbf{s}\oplus s\sim p_t^{\pi_{NM}}}{\mathbb{V}\text{ar}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{NM}(a^*|\mathbf{s}\oplus s))\right]\right] + \underset{\mathbf{s}\oplus(\cdot,s_{-i})\sim p_t^{\bar{\pi}_M}}{\mathbb{V}\text{ar}}\left[\mathbb{E}\left[\mathcal{B}(\bar{\pi}_M(a^*|s_i,s_{-i},t))\right]\right].$$

where $\mathbf{s}\oplus s \in \mathcal{S}^t$ is any sequence of length t such that the final state is s, i.e., $\mathbf{s}\oplus s := (\mathbf{s}_{t-1} \in \mathcal{S}^{t-1})\oplus s$, and $\mathcal{B}(x)$ is a Bernoulli with parameter x.

Unsurprisingly, this Lemma shows that whenever the optimal Non-Markovian strategy for requires to adapt its decision in a joint state s according to the history that led to it, an optimal Markovian policy for the same objective must necessarily be a stochastic policy, additionally, whenever the optimal Markovian policy conditioned over per-agent states only will need to be stochastic whenever the optimal Markovian strategy conditioned on the full states randomizes its decision based on the joint state s.

Proof. Let us consider the random variable $A_i \sim \mathcal{P}_i$ denoting the event "the agent i takes action $a_i^* \in \mathcal{A}_i$". Through the law of total variance [3], we can write the variance of A given $s \in \mathcal{S}$ and $t \geqslant 0$ as

$$\begin{aligned}
\mathbb{V}\text{ar}\left[A|s,t\right] &= \mathbb{E}\left[A^2|s,t\right] - \mathbb{E}\left[A|s,t\right]^2 = \mathbb{E}_{\mathbf{s}}\left[\mathbb{E}\left[A^2|s,t,\mathbf{s}\right]\right] - \mathbb{E}_{\mathbf{s}}\left[\mathbb{E}\left[A|s,t,\mathbf{s}\right]\right]^2 \\
&= \mathbb{E}_{\mathbf{s}}\left[\mathbb{V}\text{ar}\left[A|s,t,\mathbf{s}\right] + \mathbb{E}\left[A|s,t,\mathbf{s}\right]^2\right] - \mathbb{E}_{\mathbf{s}}\left[\mathbb{E}_{\pi}\left[A|s,t,\mathbf{s}\right]\right]^2 \\
&= \mathbb{E}_{\mathbf{s}}\left[\mathbb{V}\text{ar}\left[A|s,t,\mathbf{s}\right]\right] + \mathbb{E}_{\mathbf{s}}\left[\mathbb{E}\left[A|s,t,\mathbf{s}\right]^2\right] - \mathbb{E}_{\mathbf{s}}\left[\mathbb{E}\left[A|s,t,\mathbf{s}\right]\right]^2 \\
&= \mathbb{E}_{\mathbf{s}}\left[\mathbb{V}\text{ar}\left[A|s,t,\mathbf{s}\right]\right] + \underset{\mathbf{s}}{\mathbb{V}\text{ar}}\left[\mathbb{E}\left[A|s,t,\mathbf{s}\right]\right].
\end{aligned} \tag{11}$$

Now let the conditioning event $\mathbf{s}$ be distributed as $\mathbf{s} \sim p_{t-1}^{\pi_{\mathrm{NM}}}$, so that the condition $s, t, \mathbf{s}$ becomes $\mathbf{s} \oplus s$ where $\mathbf{s} \oplus s = (s_0, a_0, s_1, \dots, s_t = s) \in \mathcal{S}^t$, and let the variable A be distributed according to $\mathcal{P}$ that maximizes the objective given the conditioning. Hence, we have that the variable A on the left hand side of (11) is distributed as a Bernoulli $\mathcal{B}(\bar{\pi}_{\mathrm{M}}(a^*|s,t))$, and the variable A on the right hand side of (12) is distributed as a Bernoulli $\mathcal{B}(\pi_{\mathrm{NM}}(a^*|\mathbf{s} \oplus s))$. Thus, we obtain

$$\mathbb{V}\mathrm{ar}\big[\mathcal{B}(\bar{\pi}_{\mathrm{M}}(a^*|s,t))\big] = \mathop{\mathbb{E}}_{\mathbf{s}\oplus s\sim p_t^{\pi_{\mathrm{NM}}}}\big[\mathbb{V}\mathrm{ar}\big[\mathcal{B}(\pi_{\mathrm{NM}}(a^*|\mathbf{s}\oplus s))\big]\big] + \mathop{\mathbb{V}\mathrm{ar}}_{\mathbf{s}\oplus s\sim p_t^{\pi_{\mathrm{NM}}}}\big[\mathbb{E}\big[\mathcal{B}(\pi_{\mathrm{NM}}(a^*|\mathbf{s}\oplus s))\big]\big]. \tag{12}$$

We know from Lemma B.1 that the policy π_{NM} is deterministic, so that $\mathbb{V}\mathrm{ar}\big[\mathcal{B}(\pi_{\mathrm{NM}}(a^*|\mathbf{s} \oplus s))\big] = 0$ for every $\mathbf{s} \oplus s$. We then repeat the same steps in order to compare the two different Markovian policies:

$$\mathbb{V}\mathrm{ar}\big[A|s_i,t\big] = \mathop{\mathbb{E}}_{s_{-i}}\Big[\mathbb{V}\mathrm{ar}\big[A|s_i,s_{-i},t\big]\Big] + \mathop{\mathbb{V}\mathrm{ar}}_{s_{-i}}\Big[\mathbb{E}\big[A|s_i,s_{-i},t\big]\Big].$$

Repeating the same considerations as before we get that we can use (12) to get:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big[\mathcal{B}(\pi_{\mathrm{M}}(a^*|s_i,t))\big] &= \mathop{\mathbb{E}}_{\mathbf{s}\oplus(\cdot,s_{-i})\sim p_t^{\bar{\pi}_{\mathrm{M}}}}\Big[\mathbb{V}\mathrm{ar}\big[\mathcal{B}(\bar{\pi}_{\mathrm{M}}(a^*|s_i,s_{-i},t))\big] + \mathbb{E}\big[\mathcal{B}(\bar{\pi}_{\mathrm{M}}(a^*|s_i,s_{-i},t))\big]\Big] \\ &= \mathop{\mathbb{V}\mathrm{ar}}_{\mathbf{s}\oplus s\sim p_t^{\pi_{\mathrm{NM}}}}\big[\mathbb{E}\big[\mathcal{B}(\pi_{\mathrm{NM}}(a^*|\mathbf{s}\oplus s))\big]\big] + \mathop{\mathbb{V}\mathrm{ar}}_{\mathbf{s}\oplus(\cdot,s_{-i})\sim p_t^{\bar{\pi}_{\mathrm{M}}}}\big[\mathbb{E}\big[\mathcal{B}(\bar{\pi}_{\mathrm{M}}(a^*|s_i,s_{-i},t))\big]\big].\end{aligned}$$

□

C Details on the Empirical Corroboration.

All the experiments were performed over an Apple M2 chip (8-core CPU, 8-core GPU, 16-core Neural Engine) with 8 GB unified memory with a maximum time of execution of 24 hours.

Environments. The main empirical proof of concept was based on two environments. First, Env. (**i**), the so called *secret room* environment by [19]. In this environment, two agents operate within two rooms of a 10×10 discrete grid. There is one switch in each room, one in position $(1, 9)$ (corner of first room), another in position $(9, 1)$ (corner of second room). The rooms are separated by a door and agents start in the same room deterministically at positions $(1, 1)$ and $(2, 2)$ respectively. The door will open only when one of the switches is occupied, which means that the (Manhattan) distance between one of the agents and the switch is less than 1.5. The full state vector contains x, y locations of the two agents and binary variables to indicate if doors are open *but* per-agent policies are conditioned on their respective states only and the state of the door. For Sparse-Rewards Tasks, the goal was set to be deterministically at the worst case, namely $(9, 9)$ and to provide a positive reward to both the agents of 100 when reached, which means again that the (Manhattan) distance between one of the agents and the switch is less than 1.5, a reward of 0 otherwise. The second environment, Env. (**ii**), was the MaMuJoCo *reacher* environment [32]. In this environment, two agents operate the two linked joints and each space dimension is discretized over 10 bins. Per-agent policies were conditioned on their respective joint angles only. For Sparse-Rewards Tasks, the goal was set to be randomly at the worst case, namely on position $(\pm 0.21, \pm 0.21)$ on the boundary of the reachable area. Reaching the goal mean to have a tip position (not observable by the agents and not discretized) at a distance less that 0.05 and provides a positive reward to both the agents of 1 when reached, a reward of 0 otherwise.

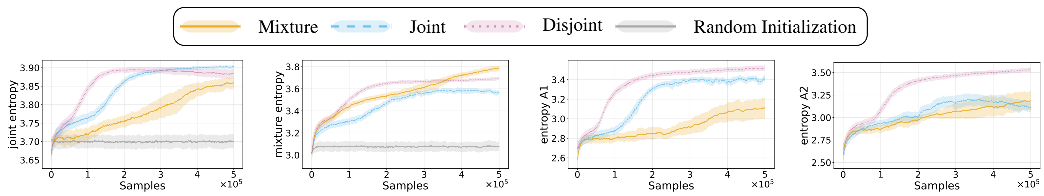
Class of Policies. In Env. (**i**), the policy was parametrized by a dense $(64, 64)$ Neural Network that takes as input the per-agent state features and outputs an action vector probabilities through a last soft-max layer. In Env. (**ii**), the policy was represented by a Gaussian distribution with diagonal covariance matrix. It takes as input the environment state features and outputs an action vector. The mean is state-dependent and is the downstream output of a a dense $(64, 64)$ Neural Network. The standard deviation is state-independent, represented by a separated trainable vector and initialized to -0.5. The weights are initialized via Xavier Initialization.

Trust Region Pure Exploration (TRPE). As outlined in the pseudocode of Algorithm 1, in each epoch a dataset of N trajectories is gathered for a given exploration horizon T, leading to the reported number of samples. Throughout the experiment the number of epochs e were set equal to $e = 10k$, the number of trajectories $N = 10$, the KL threshold $\delta = 6$, the maximum number of off-policy iterations set to $n_{\text{off,iter}} = 20$, the learning rate was set to $\eta = 10^{-5}$ and the number of seeds set equal to 4 due to the inherent low stochasticity of the environment.

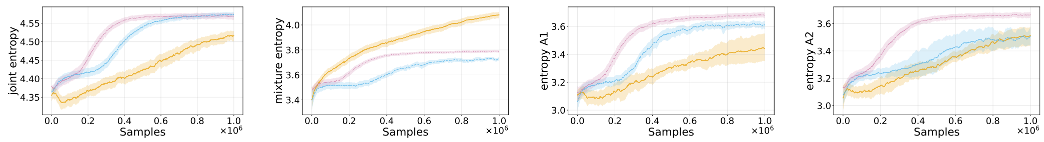
Limitations of TRPE The main limitations of the proposed methods are two. First, the Monte-Carlo estimation of single-trial objectives might be sample-inefficient in high-dimensional tasks. However, more efficient estimators of single-trial objectives remain an open question in single-agent convex RL as well, as the convex nature of the problem hinders the applicability of Bellman operators. Secondly, the plug-in estimator of the entropy is applicable to discrete spaces only, but designing scalable estimators of the entropy in continuous domains is usually a contribution *per se* [25].

Multi-Agent TRPO (MA-TRPO). We follow the same notation in [6]. Agents have independent critics $(64, 64)$ Dense networks and in each epoch a dataset of N trajectories is gathered for a given exploration horizon T for each agent, leading to the reported number of samples. Throughout the experiment the number of epochs e were set equal to $e = 100$, the number of trajectories building the batch size $N = 20$, the KL threshold $\delta = 10^{-4}$, the maximum number of off-policy iterations set to $n_{\text{off,iter}} = 20$, the discount was set to $\gamma = 0.99$.

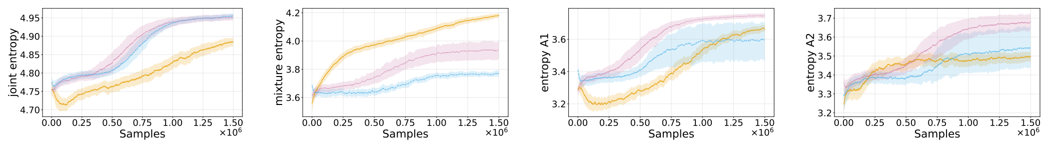
The Repository is made available at the following Repository.



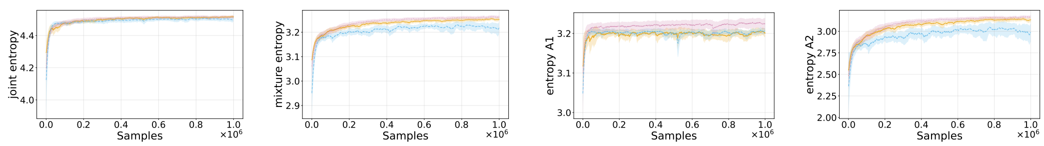
(a) TRPE Joint Entropy (Env. (**i**), $T = 50$).
(b) TRPE Mixture Entropy (Env. (**i**), $T = 50$).
(c) TRPE Entropy Agent 1 (Env. (**i**), $T = 50$).
(d) TRPE Entropy Agent 2 (Env. (**i**), $T = 50$).



(e) TRPE Joint Entropy (Env. (**i**), $T = 100$).
(f) TRPE Mixture Entropy (Env. (**i**), $T = 100$).
(g) TRPE Entropy Agent 1 (Env. (**i**), $T = 100$).
(h) TRPE Entropy Agent 2 (Env. (**i**), $T = 100$).

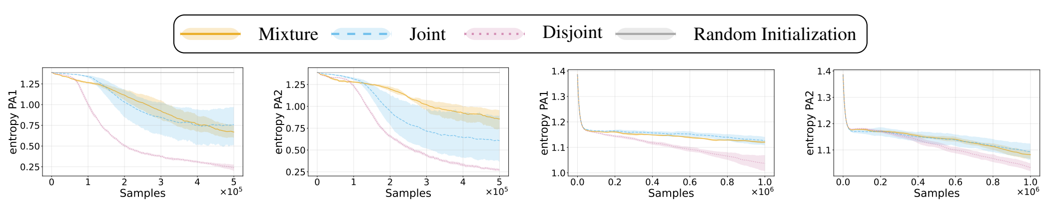


(i) TRPE Joint Entropy (Env. (**i**), $T = 150$).
(j) TRPE Mixture Entropy (Env. (**i**), $T = 150$).
(k) TRPE Entropy Agent 1 (Env. (**i**), $T = 150$).
(l) TRPE Entropy Agent 2 (Env. (**i**), $T = 150$).



(m) TRPE Joint Entropy (Env. (**ii**), $T = 100$).
(n) TRPE Mixture Entropy (Env. (**ii**), $T = 100$).
(o) TRPE Entropy Agent 1 (Env. (**ii**), $T = 100$).
(p) TRPE Entropy Agent 2 (Env. (**ii**), $T = 100$).

Figure 3: Full Visualization of Reported Experiments. Experiments with longer horizons highlight how **the easier the task, the less crucial the distinction between the objectives is.**



(a) Entropy of Agent 1 Policy in TRPE Training (**i**), $T = 50$).
(b) Entropy of Agent 2 Policy in TRPE Training (**i**), $T = 50$).
(c) Entropy of Agent 1 Policy in TRPE Training (**ii**), $T = 100$).
(d) Entropy of Agent 2 Policy in TRPE Training (**ii**), $T = 100$).

Figure 4: Policiy Entropy Insights for TRPO Pretraining in Env (**i**) and Env (**ii**). **Lower Entropic Policies with Disjoint Objectives might justify the difference in pre-training performance even if the performances in training are similar**.