Research Talk

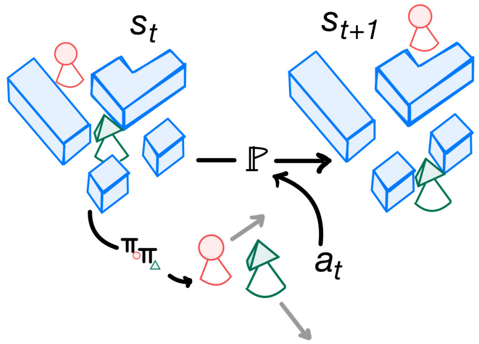
Riccardo Zamboni

06-08-2025

Outline

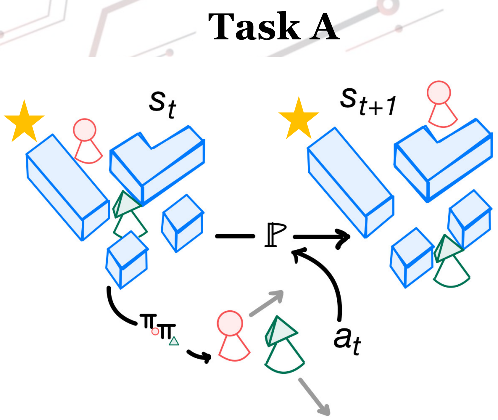
- Motivations
- Unsupervised Pre-Training: The Setting & Problem Formulation
- One Fun Fact
- Pre-training with Partial Observations
- Pre-training with Multiple Agents
- Future research directions

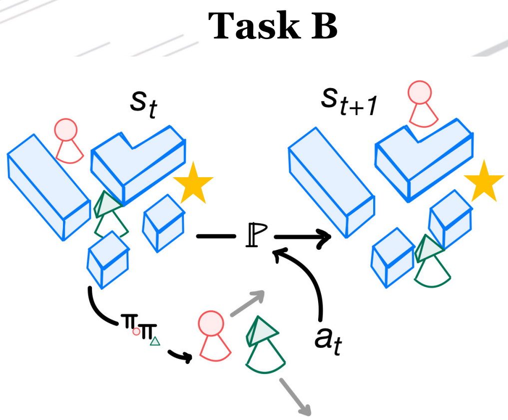
Motivations

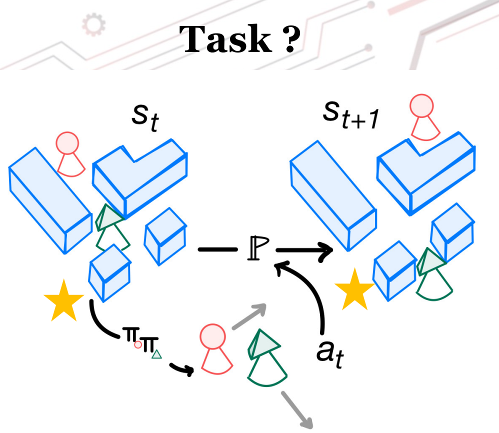


$$d^\pi(s = s_{t+1}) \in \Delta_{\mathcal{S}}$$

Motivations







Motivations

What if you happen to have a simulator, but the **task** is **mis-specified**, or **not fixed**, or yet **unknown**?

Motivations

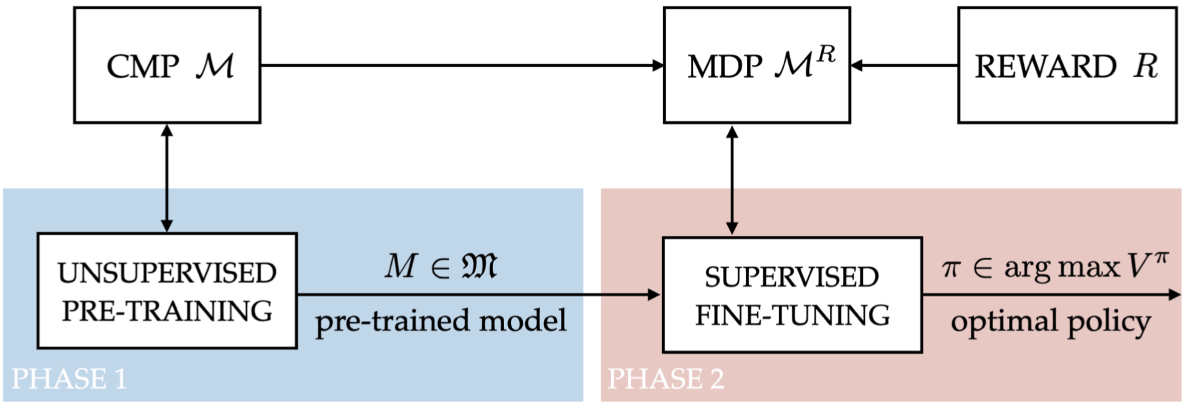
What if you happen to have a simulator, but the **task** is **mis-specified**, or **not fixed**, or yet **unknown**?

In RL, **unsupervised pre-training** [1, 2] is a solution:

Learn **something useful** no matter the task, to leverage later as soon as a task is provided.

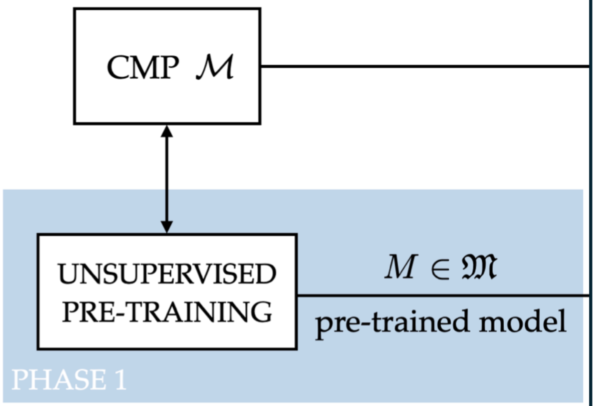
[1] Laskin et al., Unsupervised reinforcement learning benchmark, NeurIPS 2021
[2] Zisselmann et al. Explore to Generalize in Zero-Shot RL. NeurIPS 2023

Unsupervised Pre-Training: The Setting & Problem Formulation



[1] Laskin et al., Unsupervised reinforcement learning benchmark, NeurIPS 2021
[2] Zisselmann et al. Explore to Generalize in Zero-Shot RL. NeurIPS 2023

Unsupervised Pre-Training: The Setting & Problem Formulation

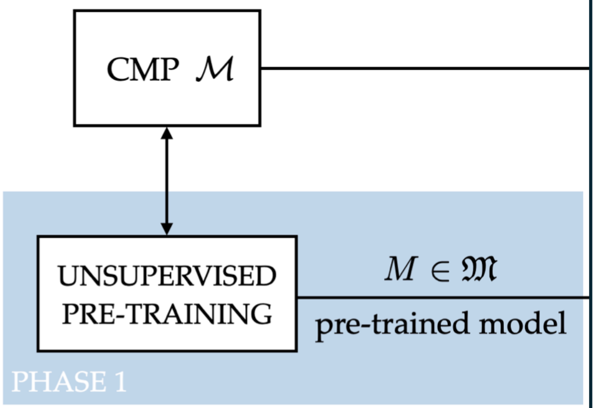


Unsupervised Pre-Training Objective

$$\max_{M \in \mathfrak{M}} \mathcal{F}_{\text{pre-train}}(M, \mathcal{M})$$

[1] Laskin et al., Unsupervised reinforcement learning benchmark, NeurIPS 2021
[2] Zisselmann et al. Explore to Generalize in Zero-Shot RL. NeurIPS 2023

Unsupervised Pre-Training: The Setting & Problem Formulation



Unsupervised Pre-Training Objective

$$\max_{M \in \mathfrak{M}} \mathcal{F}_{\text{pre-train}}(M, \mathcal{M})$$

Which model should we pre-train?

- Transition Models
- Representations
- Data-Sets
- Policy Spaces
- Policies

[1] Laskin et al., Unsupervised reinforcement learning benchmark, NeurIPS 2021
[2] Zisselmann et al. Explore to Generalize in Zero-Shot RL. NeurIPS 2023

Unsupervised Pre-Training: The Setting & Problem Formulation

Unsupervised Pre-Training Objective

$$\max_{M \in \mathfrak{M}} \mathcal{F}_{\text{pre-train}}(M, \mathcal{M})$$

Unsupervised Pre-Training: The Setting & Problem Formulation

Unsupervised Pre-Training Objective

$$\max_{M \in \mathfrak{M}} \mathcal{F}_{\text{pre-train}}(M, \mathcal{M})$$

We pre-train policies.

State Entropy Maximization

$$\mathcal{F}_{\text{pre-train}} = H(d^\pi)$$

$$H(d^\pi) := - \mathbb{E}_{s \sim d^\pi} \log d^\pi(s)$$

$$d^\pi(s) = \frac{1}{T} \sum_{t \in [T]} Pr(s_t = s | \pi, \mu)$$

Unsupervised Pre-Training: The Setting & Problem Formulation

Unsupervised Pre-Training Objective

$$\max_{M \in \mathfrak{M}} \mathcal{F}_{\text{pre-train}}(M, \mathcal{M})$$

We pre-train policies.

State Entropy Maximization

$$\mathcal{F}_{\text{pre-train}} = H(d^\pi)$$

$$H(d^\pi) := - \mathop{\mathbb{E}}_{s \sim d^\pi} \log d^\pi(s)$$

$$d^\pi(s) = \frac{1}{T} \sum_{t \in [T]} Pr(s_t = s | \pi, \mu)$$

Policy pre-training in **MDPs** allows for zero-shot **generalization** [2]. task-misspecification **robustness** [3]

[2] Zisselmann et al. Explore to Generalize in Zero-Shot RL. NeurIPS 2023
[3] Ashlag et al. State Entropy Regularization for Robust Reinforcement Learning, under-review 2025

One Fun Fact about State Entropy Maximization

(Standard) RL Objective:	**VS**	**Convex RL** Objective:
$\max_{d^\pi \in \Delta_{\mathcal{S}}} \langle d^\pi, r \rangle$		$\max_{d^\pi \in \Delta_{\mathcal{S}}} \mathcal{F}(d^\pi)$

One Fun Fact about State Entropy Maximization

(Standard) RL Objective:

$$\max_{d^\pi \in \Delta_\mathcal{S}} \langle d^\pi, r \rangle$$

VS

Convex RL Objective:

$$\max_{d^\pi \in \Delta_\mathcal{S}} \mathcal{F}(d^\pi)$$

Apprenticeship Learning, Inverse RL, Constrained RL, Imitation Learning, Diverse Skill Discovery are **all instances of convex RL** [4].
(I claim RLHF as well, prove me wrong)

[4] Mutti et al., Convex Reinforcement Learning in Finite Trials. JMLR 2023

One Fun Fact about State Entropy Maximization

(Standard) RL Objective:	**VS**	**Convex RL** Objective:
$\max_{d^\pi \in \Delta_\mathcal{S}} \langle d^\pi, r \rangle$		$\max_{d^\pi \in \Delta_\mathcal{S}} \mathcal{F}(d^\pi)$

Apprenticeship Learning, Inverse RL, Constrained RL, Imitation Learning, Diverse Skill Discovery are **all instances of convex RL** [4].

But Convex RL is **hard**: non-Markovian rewards and no Bellman Operators, number of trials matters.

[4] Mutti et al., Convex Reinforcement Learning in Finite Trials. JMLR 2023

One Fun Fact about State Entropy Maximization

One Hardness of Convex RL resides in the **number of trials** [4]:

Finite-Trials State Distribution:

$$d_K(s) = \frac{1}{KT} \sum_{k,t \in [K,T]} 1(\mathbf{s}_k[t] = s)$$

VS

Infinite-Trials State Distribution:

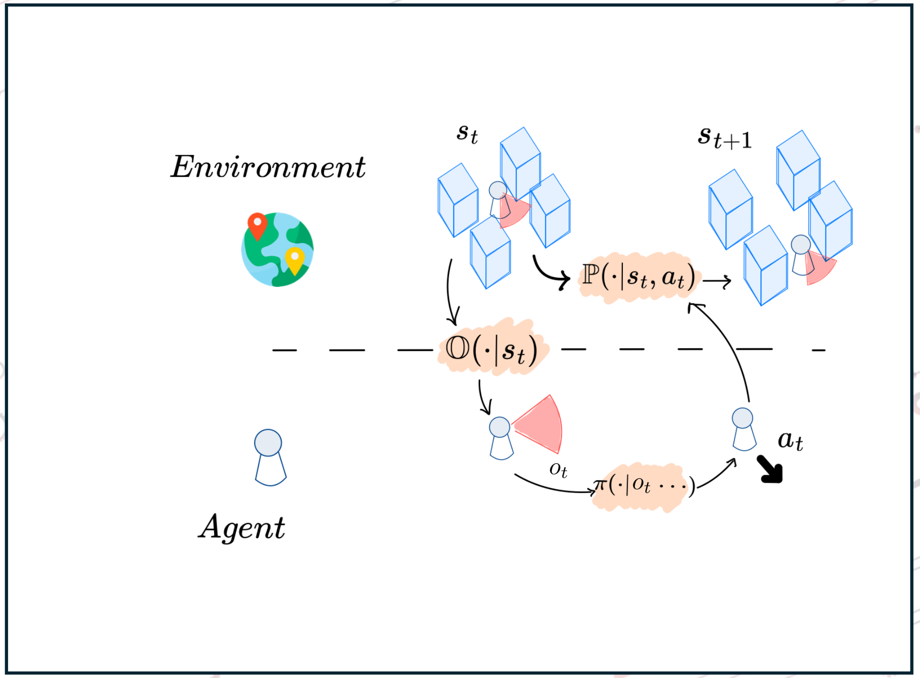
$$d^\pi(s) = \mathbb{E}_{d_K \sim p_K^\pi}[d_K(s)]$$

$$\mathcal{F}(d^\pi) \neq \mathbb{E}_{d_K \sim p_K^\pi}[\mathcal{F}(d_K)]$$

[4] Mutti et al., Convex Reinforcement Learning in Finite Trials. JMLR 2023

Pre-Training with Partial Observations

$$\mathbb{M} := (\mathcal{S}, \mathcal{O}, \mathbb{O}, \mathcal{A}, \mathbb{P}, \mu, T)$$



- $\mathcal{S}$ State Space
- $\mathcal{O}$ Observation Space
- $\mathbb{O} : \mathcal{S} \to \Delta(\mathcal{O})$ Observation Matrix
- $\mathcal{A}$ Action Space
- $\pi : \mathcal{I} \to \Delta(\mathcal{A})$ Policy
- $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ Transition Matrix
- μ Initial State Distribution
- T Episode Horizon ($t \in [T]$)

where $\mathcal{I} \in \{\mathcal{O}, \mathcal{O}^T\}$

[5] Åström, Optimal control of Markov processes with incomplete state information, 1965

Pre-Training with Partial Observations

In **Partially Observable** Environments:

- **Observations jeopardize pre-training** [A] and agents need to **regularize** with respect to the **observation quality** to counteract the mismatch.

- When learning via a **latent model** [B], learning should explicitly avoid **hallucinatory effects** of the **latent representation.**

[A] **Zamboni** et al. The Limits of Pure Exploration in POMDPs: When the Observation Entropy is Enough. RLC 2024
[B] **Zamboni** et al. How to explore with belief: state entropy maximization in POMDPs . ICML 2024

Pre-Training with Partial Observations

Maximum State Entropy (**MSE**)	vs	Maximum Observation Entropy (**MOE**)
$\max_{\pi \in \Pi} H(d_{\mathcal{S}}^{\pi})$		$\max_{\pi \in \Pi} H(d_{\mathcal{O}}^{\pi})$

[A] **Zamboni** et al. The Limits of Pure Exploration in POMDPs: When the Observation Entropy is Enough. RLC 2024

Pre-Training with Partial Observations

Maximum State Entropy (**MSE**)

$\max_{\pi\in\Pi} H(d^{\pi}_{\mathcal{S}})$

VS

Maximum Observation Entropy (**MOE**)

$\max_{\pi\in\Pi} H(d^{\pi}_{\mathcal{O}})$

$$\log\left(\frac{1}{\sigma_{\max}(\mathbb{O}^{\circ-1})}\right) \leq H(d^{\pi}_{\mathcal{S}}) - H(d^{\pi}_{\mathcal{O}}) \leq \log(\sigma_{\max}(\mathbb{O}))$$

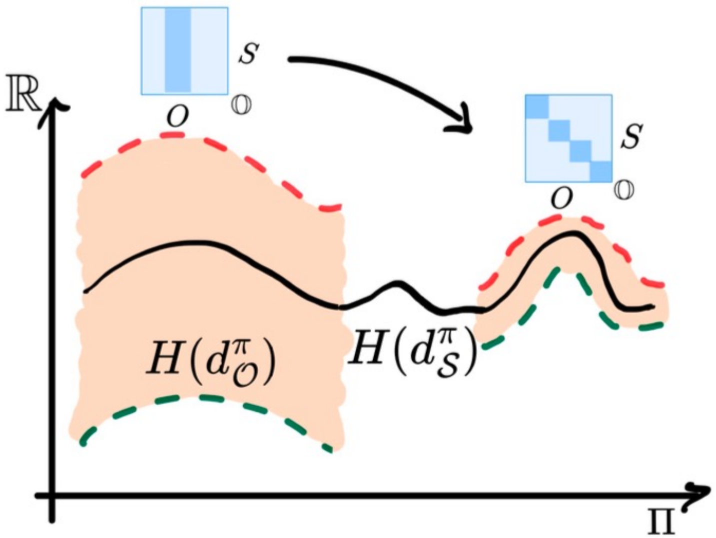
$\sigma_{\max}(A) := ||A||_2 = \sqrt{\lambda_{\max}(A^{\star}A)}$ Maximum Singular Value

$A^{\circ-1}_{ij} = \frac{1}{A_{ij}} \; \forall i,j$ Hadamard Inverse

[A] **Zamboni** et al. The Limits of Pure Exploration in POMDPs: When the Observation Entropy is Enough. RLC 2024

Pre-Training with Partial Observations

$$\log\left(\frac{1}{\sigma_{\max}(\mathbb{O}^{\circ-1})}\right) \leq H(d^\pi_{\mathcal{S}}) - H(d^\pi_{\mathcal{O}}) \leq \log(\sigma_{\max}(\mathbb{O}))$$



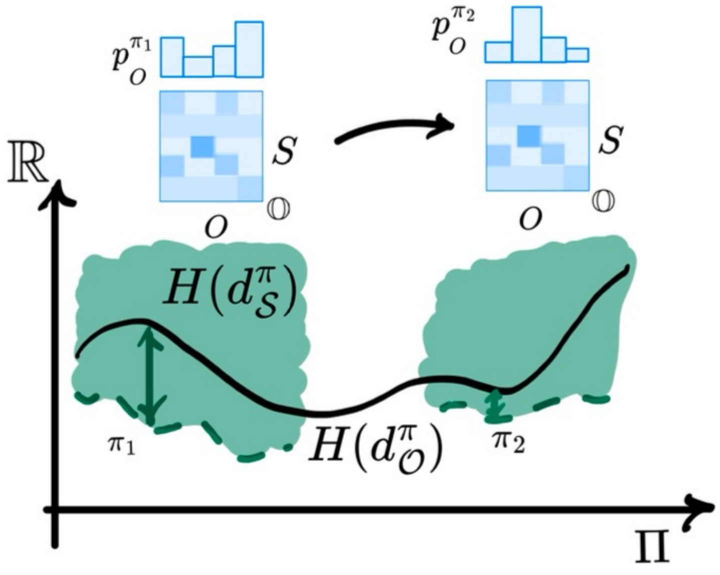
Pro: Bidirectional Bound.
Cons:

- Opaque dependency on $\mathbb{O}$.
- Independent of the policy.

[A] **Zamboni** et al. The Limits of Pure Exploration in POMDPs: When the Observation Entropy is Enough. RLC 2024

Pre-Training with Partial Observations

$$H(d_{\mathcal{S}}^{\pi}) \geq H(d_{\mathcal{O}}^{\pi}) - H(S|O,\pi) + \log(\sigma_{\max}(\mathbb{O}))$$



$$H(S|O,\pi) := \mathbb{E}_{o\sim d_{\mathcal{O}}^{\pi}}[H(\mathbb{O}(o|\cdot))]$$

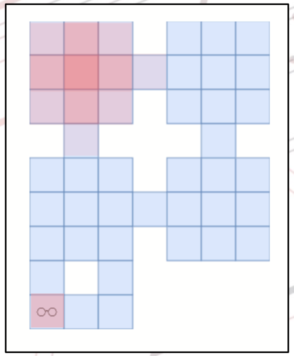
Pro:

- Implicit Dependency on the policy.
- Accessible in POMDPs.

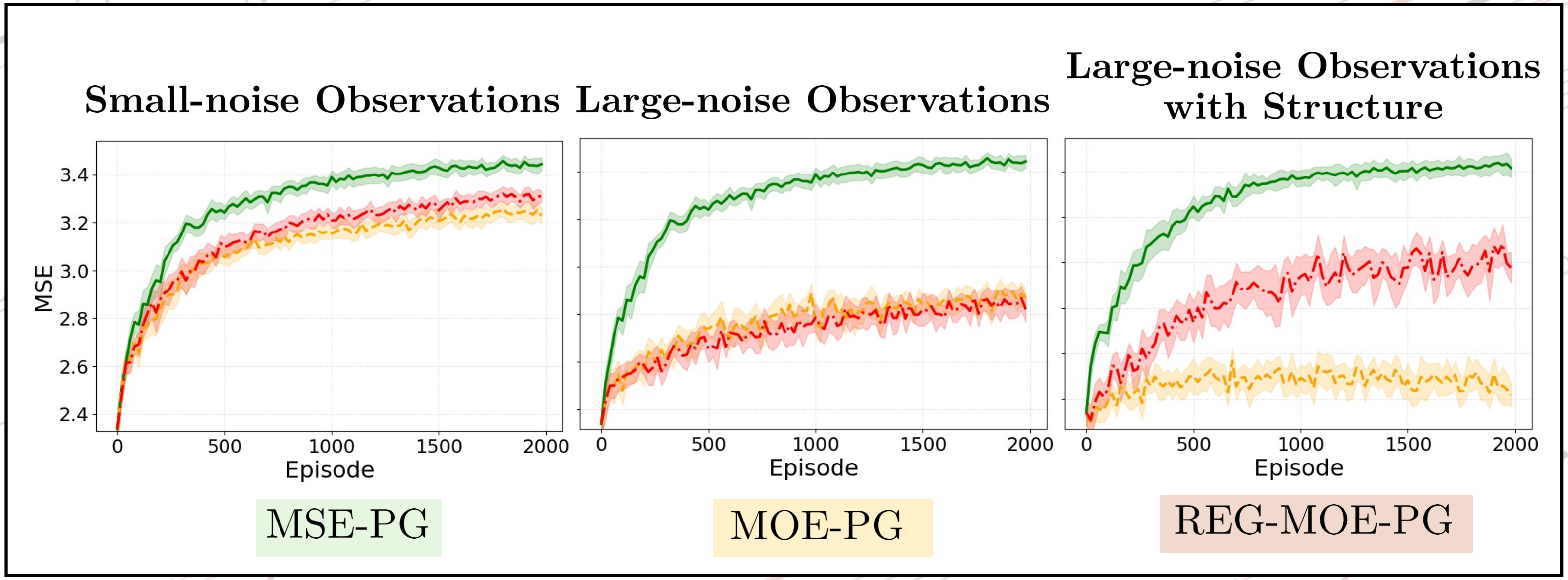
Cons: Lower-Bound only.

[A] **Zamboni** et al. The Limits of Pure Exploration in POMDPs: When the Observation Entropy is Enough. RLC 2024

Pre-Training with Partial Observations

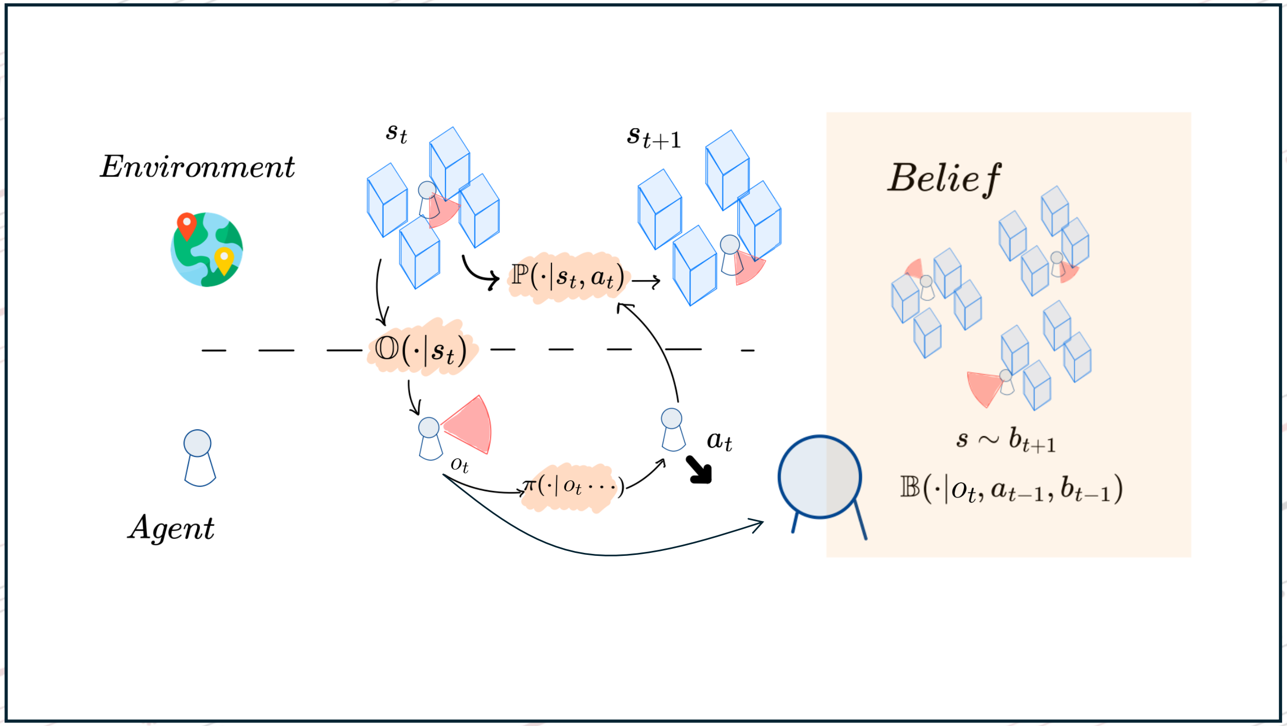


$\mathbb{X}(\cdot|s) \in \mathcal{N}(s, \sigma^2)$



[A] **Zamboni** et al. The Limits of Pure Exploration in POMDPs: When the Observation Entropy is Enough. RLC 2024

Pre-Training with Partial Observations



$\mathcal{S}$ State Space
$\mathcal{O}$ Observation Space
$\mathbb{O} : \mathcal{S} \to \Delta(\mathcal{O})$ Observation Matrix
$\mathcal{A}$ Action Space
$\pi : \mathcal{I} \to \Delta(\mathcal{A})$ Policy
$\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ Transition Matrix
$b \in \mathcal{B} \subseteq \Delta(\mathcal{S})$ Belief Model
$\mathbb{B} : \mathcal{X} \times \mathcal{A} \times \mathcal{B} \to \mathcal{B}$ Model Update
$= \frac{\mathbb{O}(o|\cdot) \sum_{s'} \mathbb{P}(\cdot|s', a) b(s')}{\sum_{s'} \mathbb{O}(o|s') \sum_{s''} \mathbb{P}(s''|s', a) b(s')}$
μ Initial State Distribution
T Episode Horizon ($t \in [T]$)

where $\mathcal{I} \in \{\mathcal{O}, \mathcal{S}, \mathcal{B}, \mathcal{O}^T, \mathcal{S}^T, \mathcal{B}^T\}$

[6] Avalos et al., The Wasserstein Believer. ICLR 2024

Pre-Training with Partial Observations

Maximum Believed Entropy (**MBE**)

$$H(d^{\pi}_{\tilde{\mathcal{S}}}) := \mathop{\mathbb{E}}_{\mathbf{b} \sim p^{\pi}_{\mathcal{B}}} \mathop{\mathbb{E}}_{d_{\tilde{\mathcal{S}}} \sim \mathbf{b}} H(d_{\tilde{\mathcal{S}}})$$

[B] **Zamboni** et al. How to explore with belief: state entropy maximization in POMDPs . ICML 2024

Pre-Training with Partial Observations

Maximum Believed Entropy (**MBE**)

$$H(d^{\pi}_{\tilde{\mathcal{S}}}) := \mathop{\mathbb{E}}_{\mathbf{b}\sim p^{\pi}_{\mathcal{B}}} \mathop{\mathbb{E}}_{d_{\tilde{\mathcal{S}}}\sim \mathbf{b}} H(d_{\tilde{\mathcal{S}}})$$

Pro:

- Learned Model
- Non-Markovianity

Learning over the latent model can be exploited to build **degenerate** (i.e. highly entropic) **representations**.

Cons: Hallucinations

[B] **Zamboni** et al. How to explore with belief: state entropy maximization in POMDPs . ICML 2024

Pre-Training with Partial Observations

$$H(d^{\pi}_{\tilde{\mathcal{S}}}) := \mathop{\mathbb{E}}_{\mathbf{b}\sim p^{\pi}_{\mathcal{B}}} \mathop{\mathbb{E}}_{d_{\tilde{\mathcal{S}}}\sim\mathbf{b}} H(d_{\tilde{\mathcal{S}}})$$

$$\geq \qquad H(\mathbf{b}) = \sum_{t\in[T]} H(\mathbf{b}_t)$$

$$H_{\beta}(d^{\pi}_{\tilde{\mathcal{S}}}) := \mathop{\mathbb{E}}_{\mathbf{b}\sim p^{\pi}_{\mathcal{B}}} \Big[\mathop{\mathbb{E}}_{d_{\tilde{\mathcal{S}}}\sim\mathbf{b}} H(d_{\tilde{\mathcal{S}}}) - \beta H(\mathbf{b}) \Big]$$

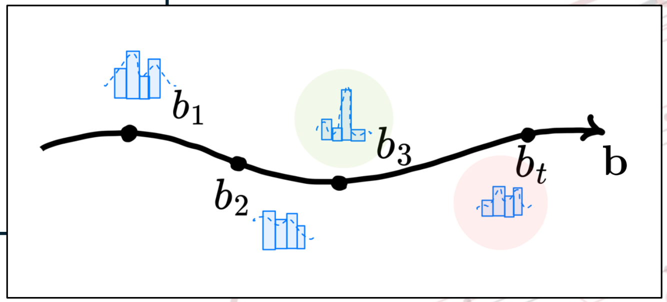
[B] **Zamboni** et al. How to explore with belief: state entropy maximization in POMDPs . ICML 2024

Pre-Training with Partial Observations

$$H(d^{\pi}_{\tilde{\mathcal{S}}}) := \mathop{\mathbb{E}}_{\mathbf{b}\sim p^{\pi}_{\mathcal{B}}} \mathop{\mathbb{E}}_{d_{\tilde{\mathcal{S}}}\sim\mathbf{b}} H(d_{\tilde{\mathcal{S}}})$$

$$\geq \qquad H(\mathbf{b}) = \textstyle\sum_{t\in[T]} H(\mathbf{b}_t)$$

$$H_{\beta}(d^{\pi}_{\tilde{\mathcal{S}}}) := \mathop{\mathbb{E}}_{\mathbf{b}\sim p^{\pi}_{\mathcal{B}}} \Big[\mathop{\mathbb{E}}_{d_{\tilde{\mathcal{S}}}\sim\mathbf{b}} H(d_{\tilde{\mathcal{S}}}) - \beta H(\mathbf{b}) \Big]$$



[B] **Zamboni** et al. How to explore with belief: state entropy maximization in POMDPs . ICML 2024

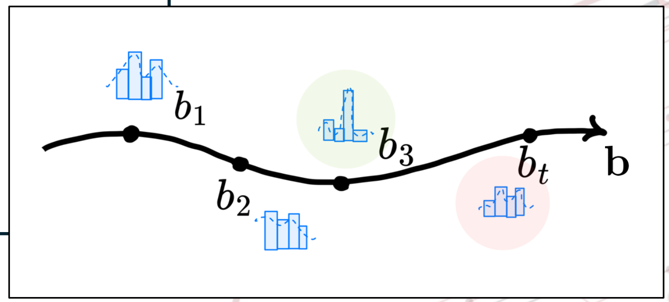
Pre-Training with Partial Observations

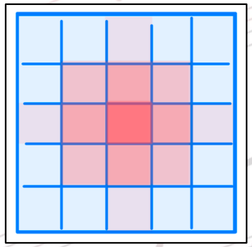
$$H(d^{\pi}_{\tilde{\mathcal{S}}}) := \mathop{\mathbb{E}}_{\mathbf{b}\sim p^{\pi}_{\mathcal{B}}} \mathop{\mathbb{E}}_{d_{\tilde{\mathcal{S}}}\sim \mathbf{b}} H(d_{\tilde{\mathcal{S}}})$$

$$\geq$$

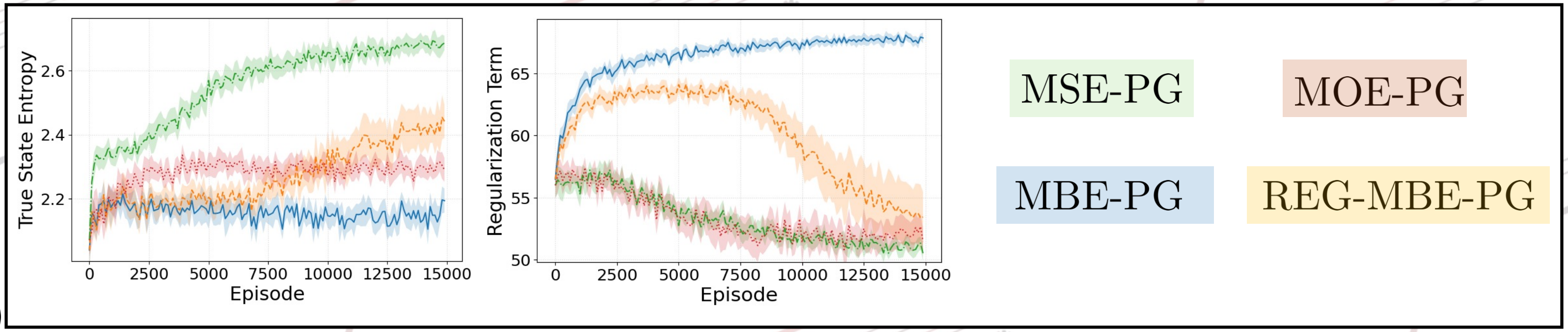
$$H(\mathbf{b}) = \textstyle\sum_{t\in[T]} H(\mathbf{b}_t)$$

$$H_{\beta}(d^{\pi}_{\tilde{\mathcal{S}}}) := \mathop{\mathbb{E}}_{\mathbf{b}\sim p^{\pi}_{\mathcal{B}}} \Big[\mathop{\mathbb{E}}_{d_{\tilde{\mathcal{S}}}\sim \mathbf{b}} H(d_{\tilde{\mathcal{S}}}) - \beta H(\mathbf{b}) \Big]$$





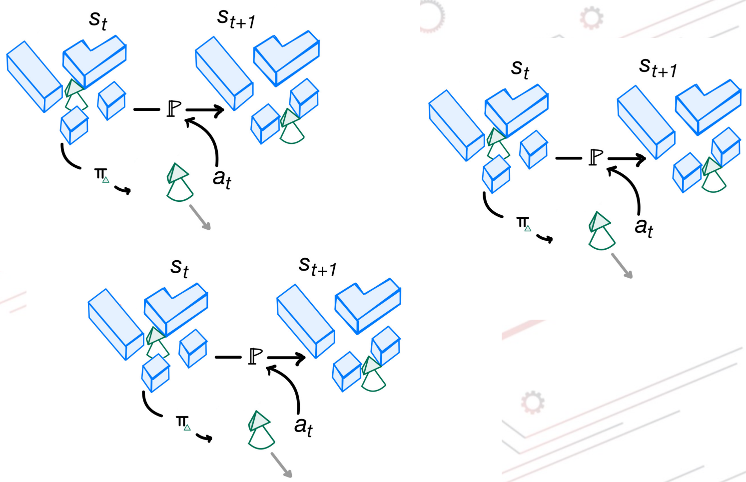
$\mathbb{X}(\cdot|s) \in \mathcal{N}(s, \sigma^2)$



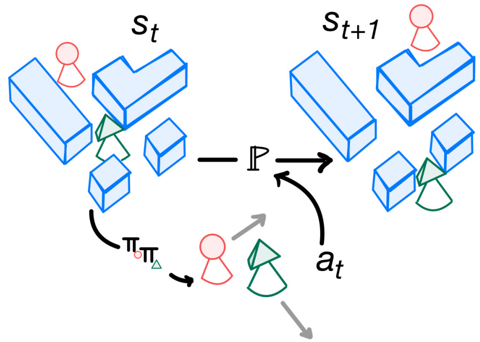
[B] **Zamboni** et al. How to explore with belief: state entropy maximization in POMDPs . ICML 2024

Pre-Training with Multiple Agents

Parallel MDPs [7]



Markov Games [8]



[7] Sucar. Parallel Markov Decision Processes. Advances in Probabilistic Graphical Models. 2007
[8] Littman. Markov games as a framework for multi-agent reinforcement learning. ICML 1994

Pre-Training with Multiple Agents

In **Multi-Agent** Environments:

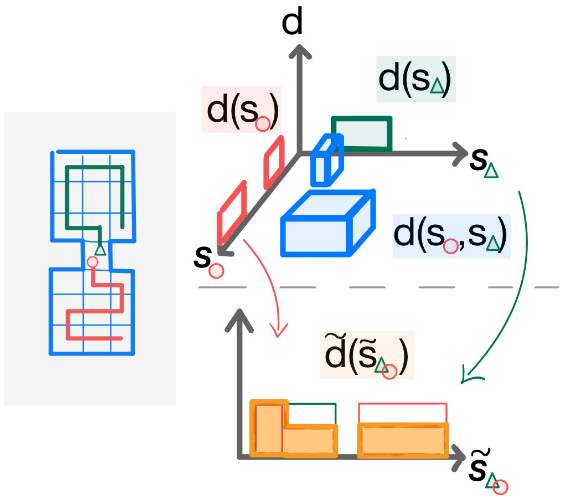
- When learning in **parallel environments** [C], **diversity collapse** should be explicitly avoided to have any advantages.

- When learning in **games** [D] over finite-trials, **curse of dimensionality** hinders the scalability of pre-training.

The answer to both these challenges is the use of **hybrid representation**.

[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025
[D] **Zamboni** et al. Towards Unsupervised Multi-Agent Reinforcement Learning. EXAIT @ ICML 2025 & Under-review

Pre-Training with Multiple Agents

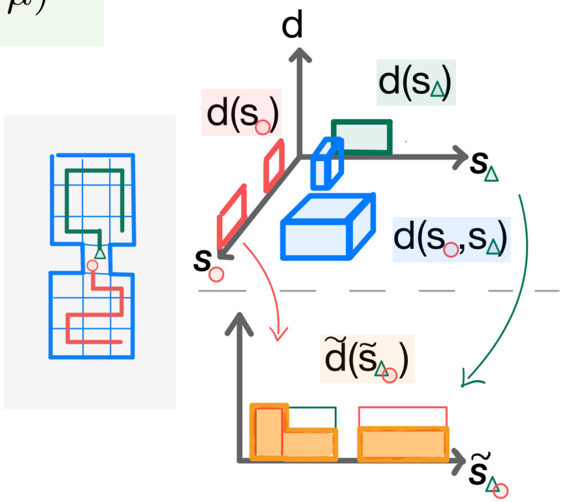


[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025
[D] **Zamboni** et al. Towards Unsupervised Multi-Agent Reinforcement Learning. EXAIT @ ICML 2025 & Under-review

Pre-Training with Multiple Agents

Marginal Distribution:

$$d_i^\pi(s_i) = \frac{1}{T}\sum_{t\in[T]} Pr(s_{t,i} = s_i|\pi,\mu)$$



Joint Distribution:

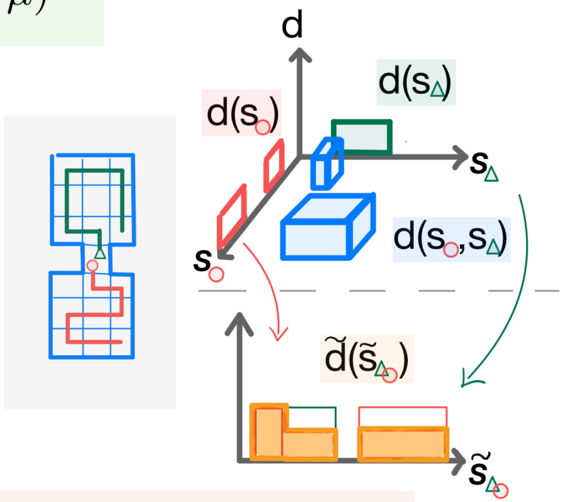
$$d^\pi(s) = \frac{1}{T}\sum_{t\in[T]} Pr(s_t = s|\pi,\mu)$$

[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025
[D] **Zamboni** et al. Towards Unsupervised Multi-Agent Reinforcement Learning. EXAIT @ ICML 2025 & Under-review

Pre-Training with Multiple Agents

Marginal Distribution:

$$d_i^\pi(s_i) = \frac{1}{T}\sum_{t\in[T]} Pr(s_{t,i} = s_i|\pi,\mu)$$



Joint Distribution:

$$d^\pi(s) = \frac{1}{T}\sum_{t\in[T]} Pr(s_t = s|\pi,\mu)$$

Mixture Distribution:

$$\tilde{d}^\pi(\tilde{s}) = \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} d_i^\pi(\tilde{s})$$

[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025
[D] **Zamboni** et al. Towards Unsupervised Multi-Agent Reinforcement Learning. EXAIT @ ICML 2025 & Under-review

Pre-Training with Multiple Agents

In **parallel** environments, the use of **mixture distributions** allows for:

- **Provably efficient learning** in infinite trials, via a **parallel** formulation of **Frank-Wolfe** [9]

[9] Hazan et al. Provably efficient Maximum Entropy Exploration. PMLR 2019
[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025

Pre-Training with Multiple Agents

In **parallel** environments, the use of **mixture distributions** allows for:

- **Provably efficient learning** in infinite trials, via a **parallel** formulation of **Frank-Wolfe** [9]

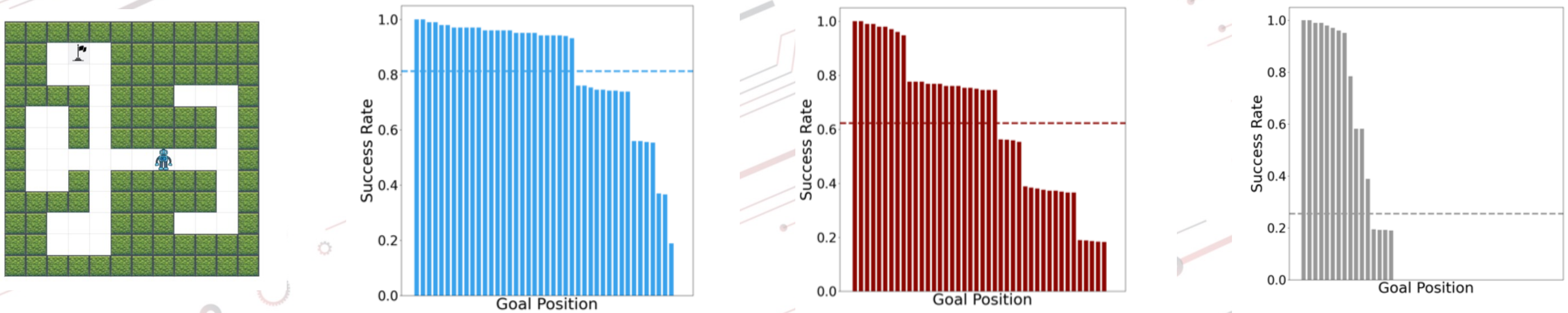
- In finite trials, optimizing the mixture entropy allows for **state distribution diversity**.

$$H(\tilde{d^\pi}) = \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} H(d_i^\pi) + \frac{1}{|\mathcal{N}|}\sum_{i\in[\mathcal{N}]} \mathrm{KL}(d_i^\pi \| \tilde{d^\pi})$$

[9] Hazan et al. Provably efficient Maximum Entropy Exploration. PMLR 2019
[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025

Pre-Training with Multiple Agents

Unsupervised parallel pre-training leads to **better data-collection** and **higher offline robustness**.



Success Rate of **Offline RL for different tasks**, with data collected with **parallel** or **non-parallel pre-trained** policies or **random** policies

[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025

Pre-Training with Multiple Agents

In **games**, the use of **mixture distributions** allows for:

- **Efficient Lower bounds** to the ideal objective

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

[D] **Zamboni** et al. Towards Unsupervised Multi-Agent Reinforcement Learning. EXAIT @ ICML 2025 & Under-review

Pre-Training with Multiple Agents

In **games**, the use of **mixture distributions** allows for:

- **Efficient Lower bounds** to the ideal objective

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

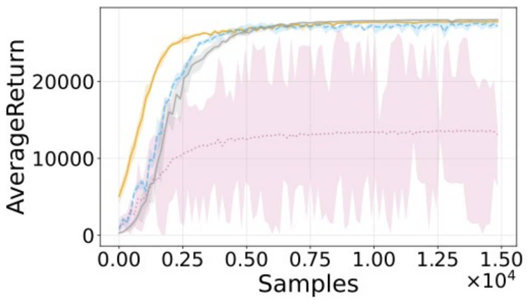
- **Faster concentration** of entropies

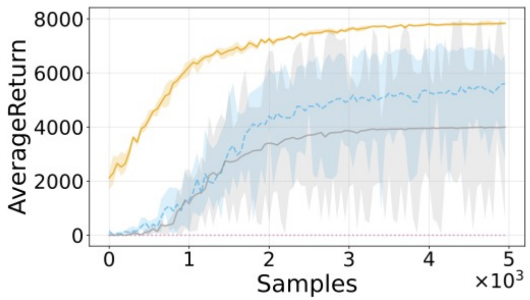
$$|H(d^\pi) - \mathbb{E}_{d_K \sim p_K^\pi} H(d_K)| \leq LT\sqrt{\frac{2|\mathcal{S}|\log(2T/\delta)}{K}} \quad \textbf{VS} \quad |H(\tilde{d}^\pi) - \mathbb{E}_{\tilde{d}_K \sim p_K^\pi} H(\tilde{d}_K)| \leq LT\sqrt{\frac{2|\tilde{\mathcal{S}}|\log(2T/\delta)}{|\mathcal{N}|K}}$$

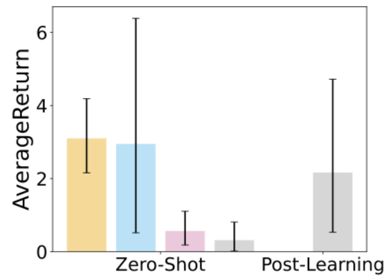
[D] **Zamboni** et al. Towards Unsupervised Multi-Agent Reinforcement Learning. EXAIT @ ICML 2025 & Under-review

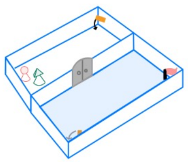
Pre-Training with Multiple Agents

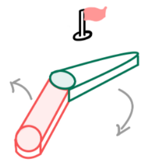
Unsupervised **multi-agent pre-training** leads to **faster learning** and **zero-shot performances** when done right.











Effect over **training dynamics** (left) and **zero-shot performances** (right) of **unsupervised policy pre-training**, with different objectives, **mixture**, **joint**, **disjoint** pre-training or **random** initialization.

[D] **Zamboni** et al. Towards Unsupervised Multi-Agent Reinforcement Learning. EXAIT @ ICML 2025 & Under-review

Future research directions

- **Scaling**, some **Scaling**, and then **Scaling**
- Unsupervised **Policy Space Compression**
- Dual optimization for **general convex** MDPs and MGs

References

[1] Laskin et al., Unsupervised reinforcement learning benchmark, NeurIPS 2021
[2] Zisselmann et al. Explore to Generalize in Zero-Shot RL. NeurIPS 2023
[3] Ashlag et al. State Entropy Regularization for Robust Reinforcement Learning, pre-print 2025
[4] Mutti et al., Convex Reinforcement Learning in Finite Trials. JMLR 2023
[5] Åström, Optimal control of Markov processes with incomplete state information, 1965
[6] Avalos et al., The Wasserstein Believer. ICLR 2024
[7] Sucar. Parallel Markov Decision Processes. Advances in Probabilistic Graphical Models. 2007
[8] Littman. Markov games as a framework for multi-agent reinforcement learning. ICML 1994
[9] Hazan et al. Provably efficient Maximum Entropy Exploration. PMLR 2019

[A] **Zamboni** et al. The Limits of Pure Exploration in POMDPs: When the Observation Entropy is Enough. RLC 2024
[B] **Zamboni** et al. How to explore with belief: state entropy maximization in POMDPs . ICML 2024
[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025
[D] **Zamboni** et al. Towards Unsupervised Multi-Agent Reinforcement Learning. EXAIT @ ICML 2025 & Under-review

Thank You

More References: Unsupervised Pre-Training

Approach	Pre-training	References
Low-rank or Block MDPs	Representations	[Misra et al., 2020], [Agarwal et al., 2020], [Modi et al., 2024]
Contrastive Loss	Representations	[Laskin et al., 2020, Luu et al., 2022], [Yu et al., 2025]
Reconstruction Loss	Representations	[Burda et al., 2019], [Anand et al., 2019], [Seo et al., 2022], [Meng et al., 2023]
Supervised Learning Loss	Representations	[Yuan et al., 2022, Yoon et al., 2023]
Reward-Free RL	Transition Model	[Jin et al., 2020], [Kaufmann et al., 2021], [Ménard et al., 2021], [Zhang et al., 2020d]
Task-Agnostic RL	Transition Model	[Zhang et al., 2020c]
Forward-Backward & Behavioral Foundation Models	Transition Model	[Touati and Ollivier, 2021, Tirinzoni et al., 2025, Sikchi et al., 2025]
World Models	Transition Model	[Ha and Schmidhuber, 2018], [Hafner et al., 2019], [Matsuo et al., 2022] [Hafner et al., 2023], [Pearce et al., 2024]
Curiosity	Transition Model	[Schmidhuber, 1991], [Pathak et al., 2017], [Burda et al., 2018]
Reward-Free Data Collection	Dataset	[Wang et al., 2020, Zanette et al., 2020]
ExORL	Dataset	[Yarats et al., 2022]
Explore2Offline	Dataset	[Lambert et al., 2022]
Count-Based	Dataset	[Bellemare et al., 2016]
Policy Space Compression	Policy Space	[Mutti et al., 2022c]
Policy Collection-Elimination	Policy Space	[Ye et al., 2023]
Mutual Information for Skill Discovery	Policy Space	[Gregor et al., 2017], [Eysenbach et al., 2018], [Hansen et al., 2019], [Sharma et al., 2019], [Campos et al., 2020], [Liu and Abbeel, 2021a], [He et al., 2022], [Zahavy et al., 2022]
Entropy Maximization	Policy	see Table 3.2
High-Level Hierarchical Policies	Policy	[Pertsch et al., 2021, Baker et al., 2022, Ramrakhya et al., 2023, Yuan et al., 2024]
Fine-Tuning Mechanisms	Policy	[Campos et al., 2021], [Pislar et al., 2021], [Xie et al., 2021], [Uchendu et al., 2023]

More References: State Entropy Maximization

Algorithm	Distribution	Space	Reference
MaxEnt	Discounted	State	[Hazan et al., 2019]
FW-AME	Stationary	State-Action	[Tarbouriech and Lazaric, 2019]
SMM	Marginal	State	[Lee et al., 2020]
IDE^{3}AL	Stationary	State	[Mutti and Restelli, 2020]
MEPOL	Marginal	State	[Mutti et al., 2021]
MaxRényi	Discounted	State-Action	[Zhang et al., 2021a]
GEM	Marginal	State	[Guo et al., 2021]
APT	Marginal	State	[Liu and Abbeel, 2021b]
RE3	Marginal	State	[Seo et al., 2021]
Proto-RL	Marginal	State	[Yarats et al., 2021]
MetaEnt	Discounted	State	[Zahavy et al., 2021]
RL-Explore-Ent	Discounted	State Trajectories	[Zahavy et al., 2021]
KME	Discounted	State	[Nedergaard and Cook, 2022]
FSC	Stationary	Observation Trajectories	[Savas et al., 2022]
CEM	Marginal	State	[Yang and Spaan, 2023]
$\eta\psi$-Learning	Discounted	State	[Jain et al., 2023]
ExpGen	Marginal	State	[Zisselman et al., 2023]
MOE	Marginal	Observation	[Zamboni et al., 2024b]
MBE	Marginal	Latent State	[Zamboni et al., 2024a]
TRPE	Marginal	State	[Zamboni et al., 2025]
PGL	Marginal	State	[Gemp et al., 2025]
PGPSE	Marginal	State	[De Paola et al., 2025]

One Fun Fact: Convex Objectives

Utility $\mathcal{F}$		Application	Infinite $\equiv$ Finite
$r \cdot d$	$r \in \mathbb{R}^S, d \in \Delta_\mathcal{S}$	RL	✓
$\|d - d_E\|_p^p$ $\mathrm{KL}(d\|\|d_E)$	$d, d_E \in \Delta_\mathcal{S}$	Imitation Learning	✗
$-d \cdot \log(d)$	$d \in \Delta_\mathcal{S}$	Pure Exploration	✗
$\mathrm{CVaR}_\alpha[r \cdot d]$ $r \cdot d - \mathbb{V}\mathrm{ar}[r \cdot d]$	$r \in \mathbb{R}^S, d \in \Delta_\mathcal{S}$	Risk-Averse RL	✗
$r \cdot d$, s.t. $\lambda \cdot d \leq c$	$r, \lambda \in \mathbb{R}^S, c \in \mathbb{R}, d \in \Delta_\mathcal{S}$	Linearly Constrained RL	✓
$-\mathbb{E}_z \mathrm{KL}(d_z\|\| \mathbb{E}_k d_k)$	$z \in \mathbb{R}^d, d_z, d_k \in \Delta_\mathcal{S}$	Diverse Skill Discovery	✗

[4] Mutti et al., Convex Reinforcement Learning in Finite Trials. JMLR 2023

Pre-Training with Partial Observations [A]

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}

[A] **Zamboni** et al. The Limits of Pure Exploration in POMDPs: When the Observation Entropy is Enough. RLC 2024

Pre-Training with Partial Observations [B]

Algorithm 1 **Reg-**PG for MaxEnt POMDPs

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

[B] **Zamboni** et al. How to explore with belief: state entropy maximization in POMDPs . ICML 2024

Pre-Training with Multiple Agents [C]

Algorithm 2 Parallel Frank-Wolfe.

1: **Input:** Step size η, number of iterations T, number of agents N, planning oracle tolerance $\varepsilon_1 > 0$, distribution estimation oracle tolerance $\varepsilon_0 > 0$.
2: Set $\{C_0^i = \{\pi_0^i\}\}_{i \in N}$ where π_0^i is an arbitrary policy, $\alpha_0^i = 1$.
3: **for** $t = 0, \dots, T-1$ **do**
4: Each agent call the state distribution oracle on $\pi_{\text{mix},t} = \frac{1}{N}\sum_i (\alpha_t^i, C_t^i)$:

$$\hat{d}^i_{\pi_{\text{mix},t}} = \textsc{DensityEst}\left(\pi_{\text{mix},t}, \varepsilon_0\right)$$

5: Define the reward function r_t^i for each agent i as

$$r_t^i(s) = \nabla H(\hat{d}^i_{\pi_{\text{mix},t}}) := \left.\frac{d\mathcal{H}(X)}{dX}\right|_{X=\hat{d}^i_{\pi_{\text{mix},t}}} .$$

6: Each agent computes the (approximately) optimal policy on r_t:

$$\pi_{t+1}^i = \textsc{ApproxPlan}\left(r_t^i, \varepsilon_1\right) .$$

7: Each agent updates

$$C_{t+1}^i = (\pi_0^i, \dots, \pi_t^i, \pi_{t+1}^i),$$
$$\alpha_{t+1}^i = ((1-\eta)\alpha_t^i, \eta).$$

8: **end for**
9: $\pi_{\text{mix},T} = \frac{1}{N}\sum_i (\alpha_T^i, C_T^i)$.

[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025

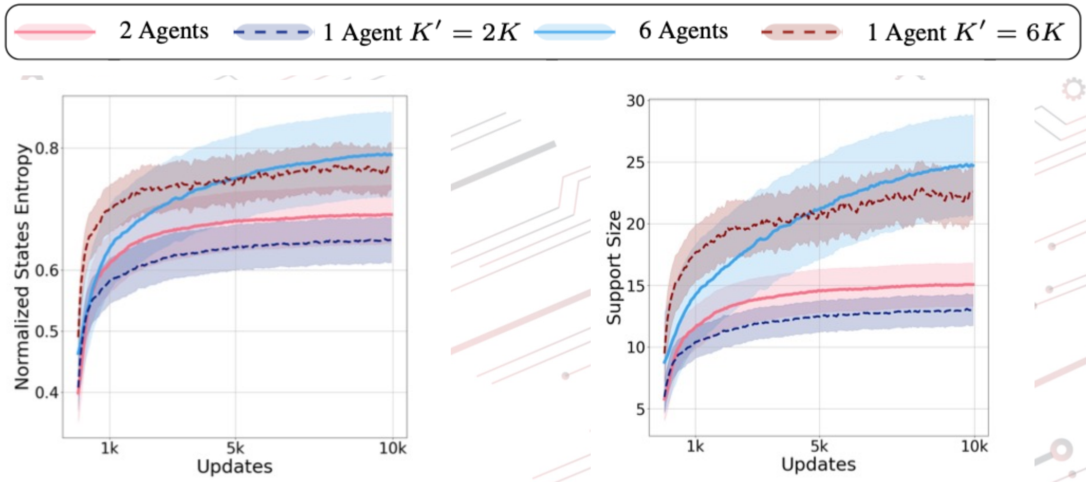
Pre-Training with Multiple Agents [C]

Algorithm 1: Policy Gradient for Parallel States Entropy maximization (**PGPSE**)

Input: Episodes N, Trajectories K, Batch Size B, Learning Rate α, Parameters $\theta = (\theta^i)_{i \in [m]}$
for $e \in \{1, \dots, N\}$ **do**
  for $itr \in \{1, \dots, B\}$ **do**
    for $k \in \{1, \dots, K\}$ **do**
      $\tau \sim \pi_\theta$ {Sample parallel trajectories}
      $\log \pi_{\theta_i} \leftarrow \sum_{t=1}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t)$
      $d_p(s) \leftarrow \frac{1}{km} \sum_{j,i,t=1}^{m,k,T} \mathbf{1}(s_{t,i,j} = s)$
      $\nabla_\theta \mathcal{J}(\theta) \mathrel{+}= \log \pi_{\theta_i} \cdot \mathcal{H}(d_p)$
    end for
  end for
  $\nabla_\theta \mathcal{J}(\theta) \leftarrow \frac{1}{B} \nabla_\theta \mathcal{J}(\theta)$
  $\theta \leftarrow \theta + \alpha \nabla_\theta \mathcal{J}(\theta)$
end for
Output: Policies $\pi_\theta = (\pi^i_{\theta^i})_{i \in [m]}$

[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025

Pre-Training with Multiple Agents [C]



[C] De Paola and **Zamboni**. Enhancing Diversity in Parallel Agents: A Maximum State Entropy Exploration Story. ICML 2025

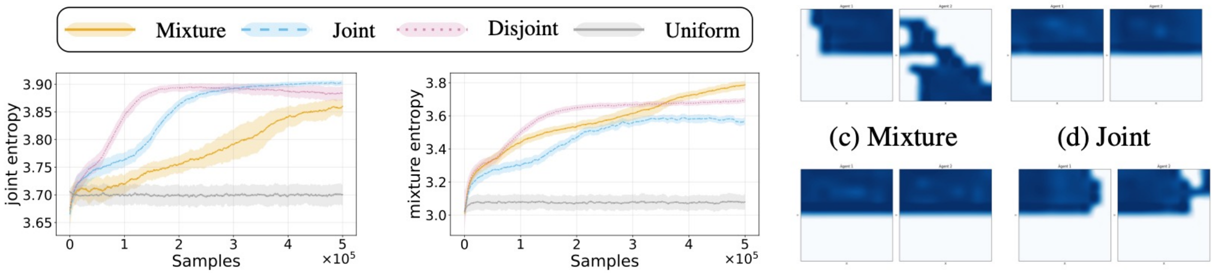
Pre-Training with Multiple Agents [D]

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 $\pi_{\boldsymbol{\theta}} = (\pi^i_{\theta^i})_{i \in [\mathcal{N}]}$

[D] **Zamboni** et al. Towards Unsupervised Multi-Agent Reinforcement Learning. EXAIT @ ICML 2025 & Under-review

Pre-Training with Multiple Agents [D]



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

[D] **Zamboni** et al. Towards Unsupervised Multi-Agent Reinforcement Learning. EXAIT @ ICML 2025 & Under-review