A Model-Based Solution to the Offline MARL Coordination Problem

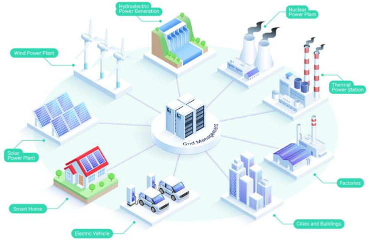
MARL Reading Group
10/24/2023

Paul Barde, Jakob Foerster, Derek Nowrouzezahrai, and Amy Zhang

Motivation

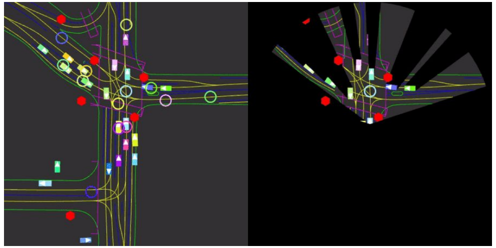
Motivation - **Offline** Multi-Agent

- Many **real-world problems** are multi-agent



Interactions are **costly** and **dangerous**

Simulations are **challenging**



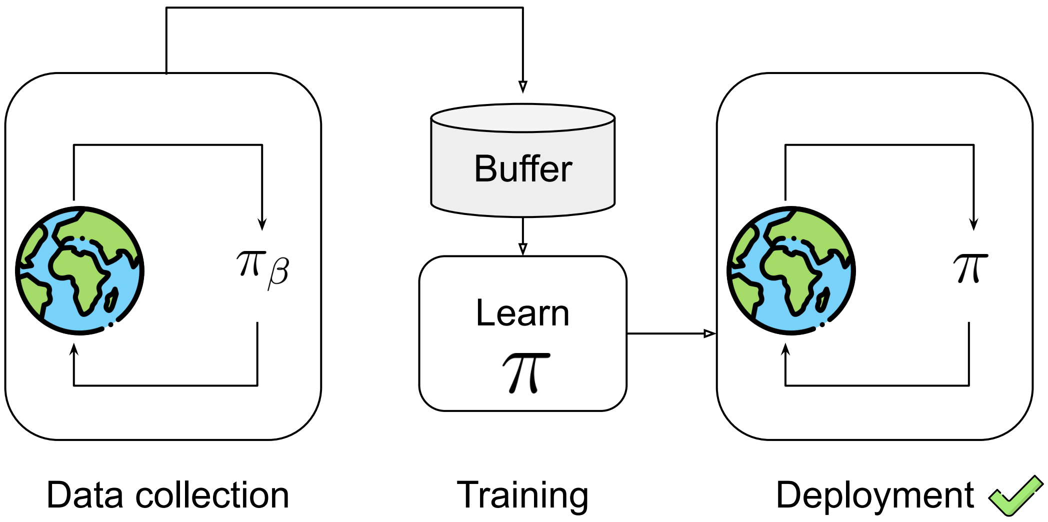
Leverage existing data

Vinitsky, Eugene, et al. "Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world." *Advances in Neural Information Processing Systems* 35 (2022)

Refresher

Refresher

Offline **Reinforcement Learning**



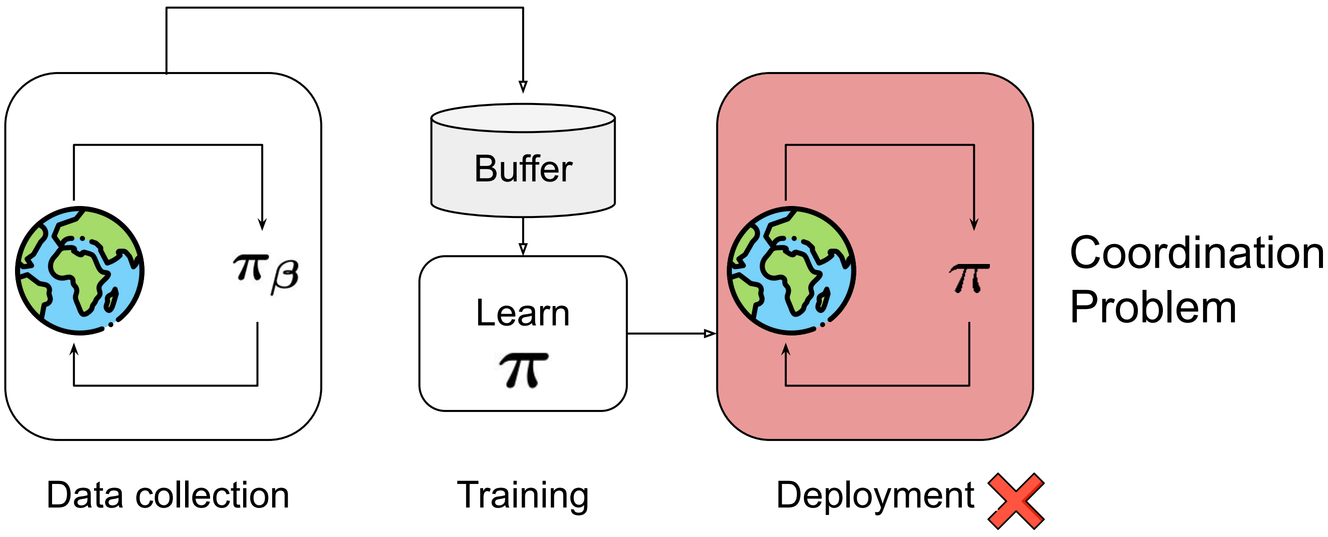
Levine, Sergey, et al. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems." *arXiv preprint arXiv:2005.01643* (2020).

Hypothesis

Offline **MARL**?

$$\pi_\beta \to \boldsymbol{\pi_\beta} \triangleq \prod_i \pi_\beta^i$$

$$\pi \to \boldsymbol{\pi} \triangleq \prod_i \pi^i$$



Definitions

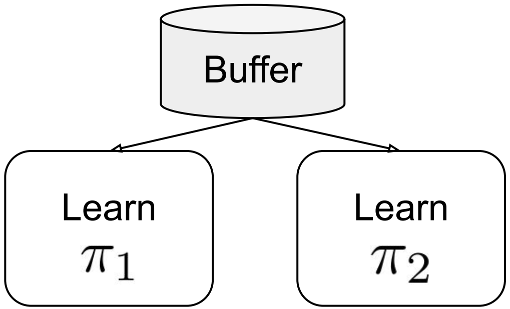
Coordination

Many different notions of coordination:

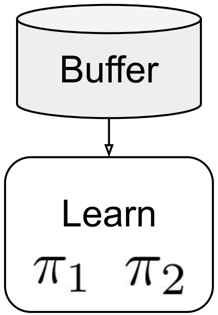
- Zero-Shot Coordination Hu, Hengyuan, et al. "“other-play” for zero-shot coordination." *International Conference on Machine Learning*. PMLR, 2020.
- Ad-Hoc Teamplay Cui, Brandon, et al. "Adversarial Diversity in Hanabi." *The Eleventh International Conference on Learning Representations*. 2022.
- Etc.

Offline Coordination

“Agents **trained offline (together)** perform well together at **deployment**.”



Independent
Learners

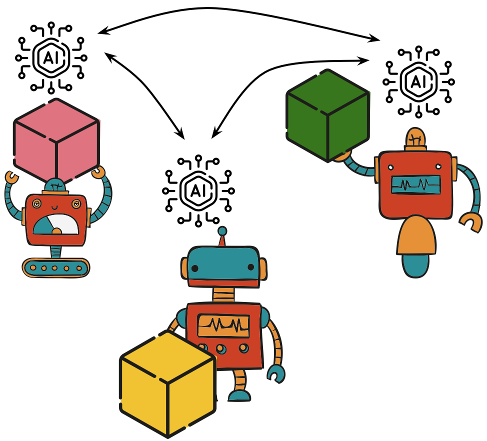


Agents **share information** during **training**

Centralized Training
Decentralized Execution
(CTDE)

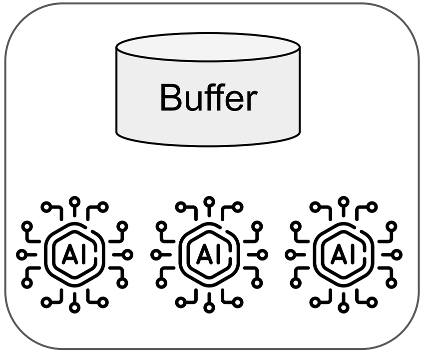
Side Note

CTDE assumption is trivial for **offline** learning.



Online learning →Physical interactions →**Embodied** learners

Sharing information is **communication intensive**.

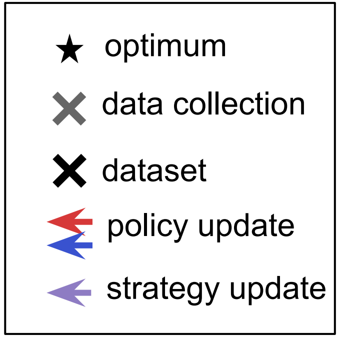


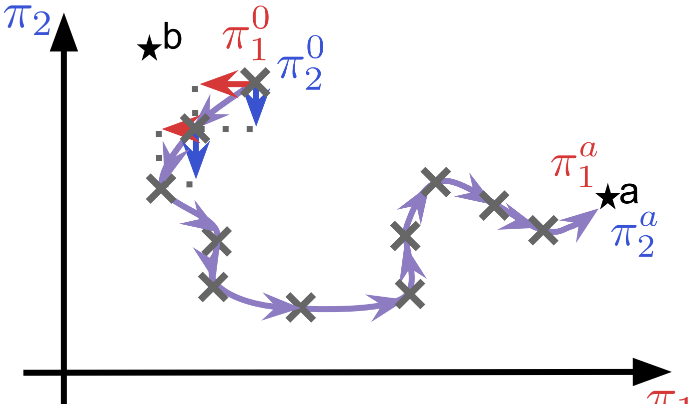
Offline learning→ No physical interaction → **Virtual** learners

Sharing information **is trivial.**

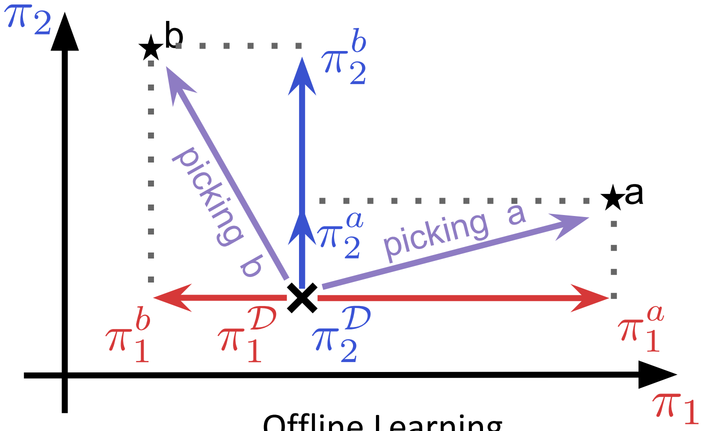
Offline Coordination Problem

The Offline Coordination Problem





Online Learning



Offline Learning

Strategy Agreement (SA)
Strategy FineTuning (SFT)

NO interactions during learning

Hypotheses

Hypotheses

(H1) : Current **offline MARL methods (model-free) fail** at Offline Coordination

- Strategy Agreement (SA)
- Strategy Fine-Tuning (SFT)

(H2) : It comes from the absence of **agent-to-agent interactions** during learning

→ **Model-Based** approaches can fix this.

Experiments

The Baselines

- Implicit Q-learning (IQL)
  - **Single-agent → centralized execution by controlling joint action**

    → **Upper bound on Strategy-Agreement** since centralized execution bypasses it.
- CTDE Learners
  - MA-IQL: CTDE extension to IQL
  - MA-TD3+BC: CTDE Twin Delayed DDPG + Behavioral Cloning regularization
- Independent Learners
  - ITD3+BC : Independent Twin Delayed DDPG + Behavioral Cloning regularization
  - ICQL : ITD3 + Regularization on Q-values (favors dataset transitions)
  - OMAR : ICQL + zero-order optimization (random shooting)
  - IBC: (Vanilla) Independent Behavior Cloning (Imitation Learning)

Model-free

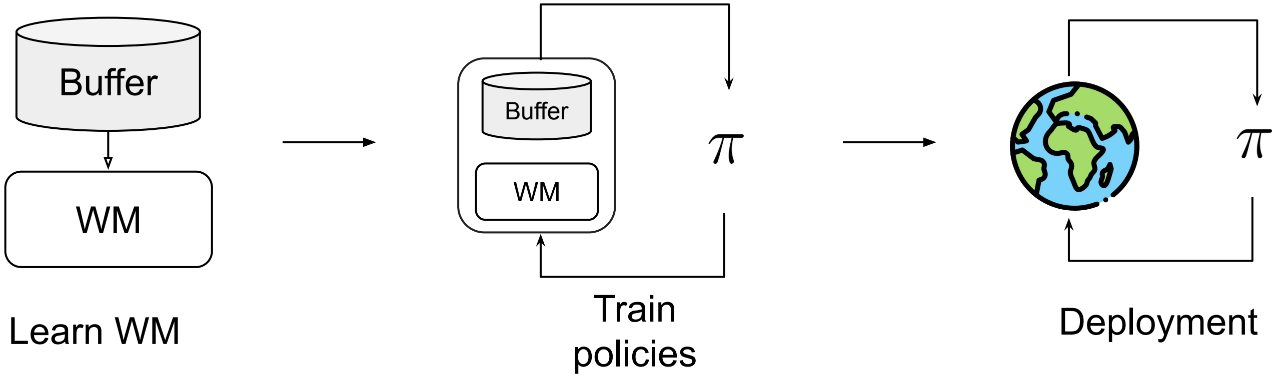
Our Method - MOMA-PPO

Model-Based Offline Multi-Agent Proximal Policy Optimization (MOMA-PPO)

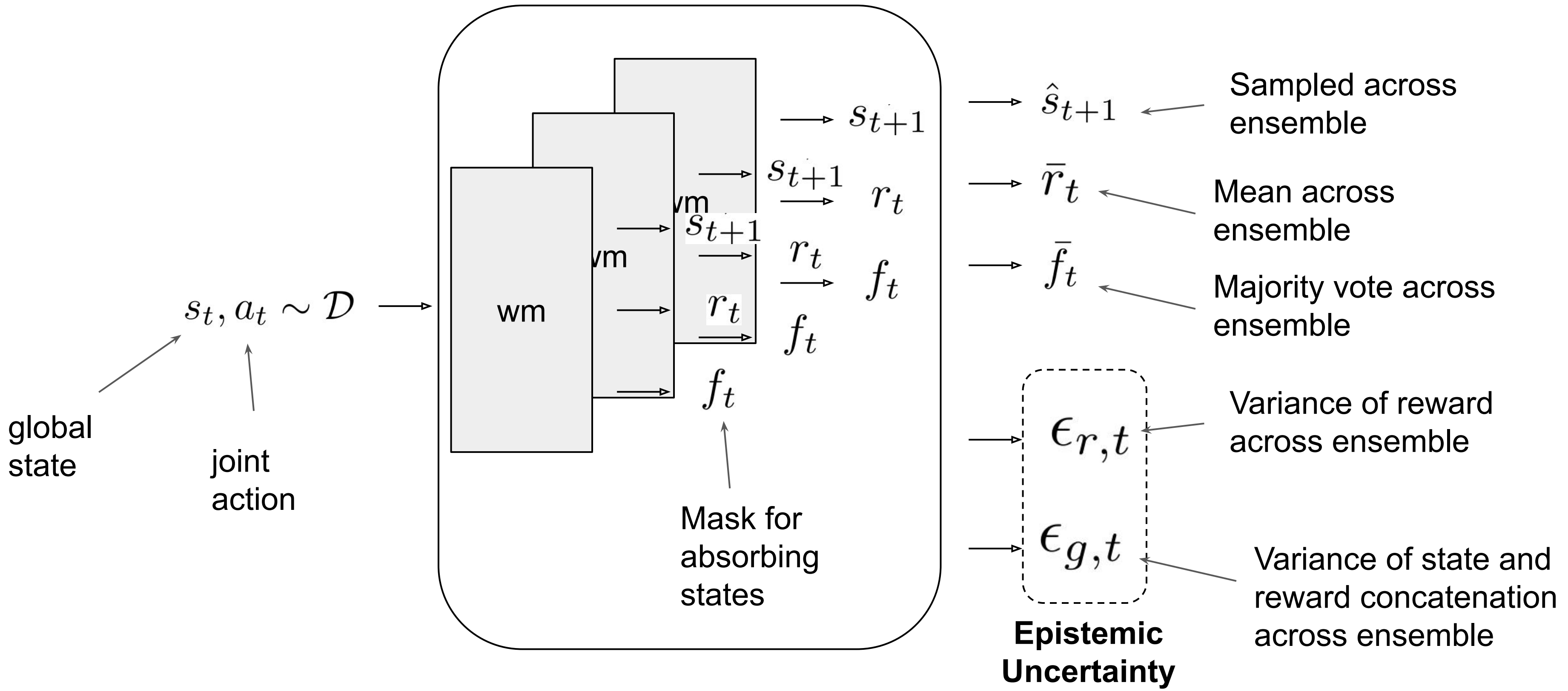
- Dyna-like approach: use model to generate training data
- CTDE
- Based on Multi-Agent PPO

Idea:

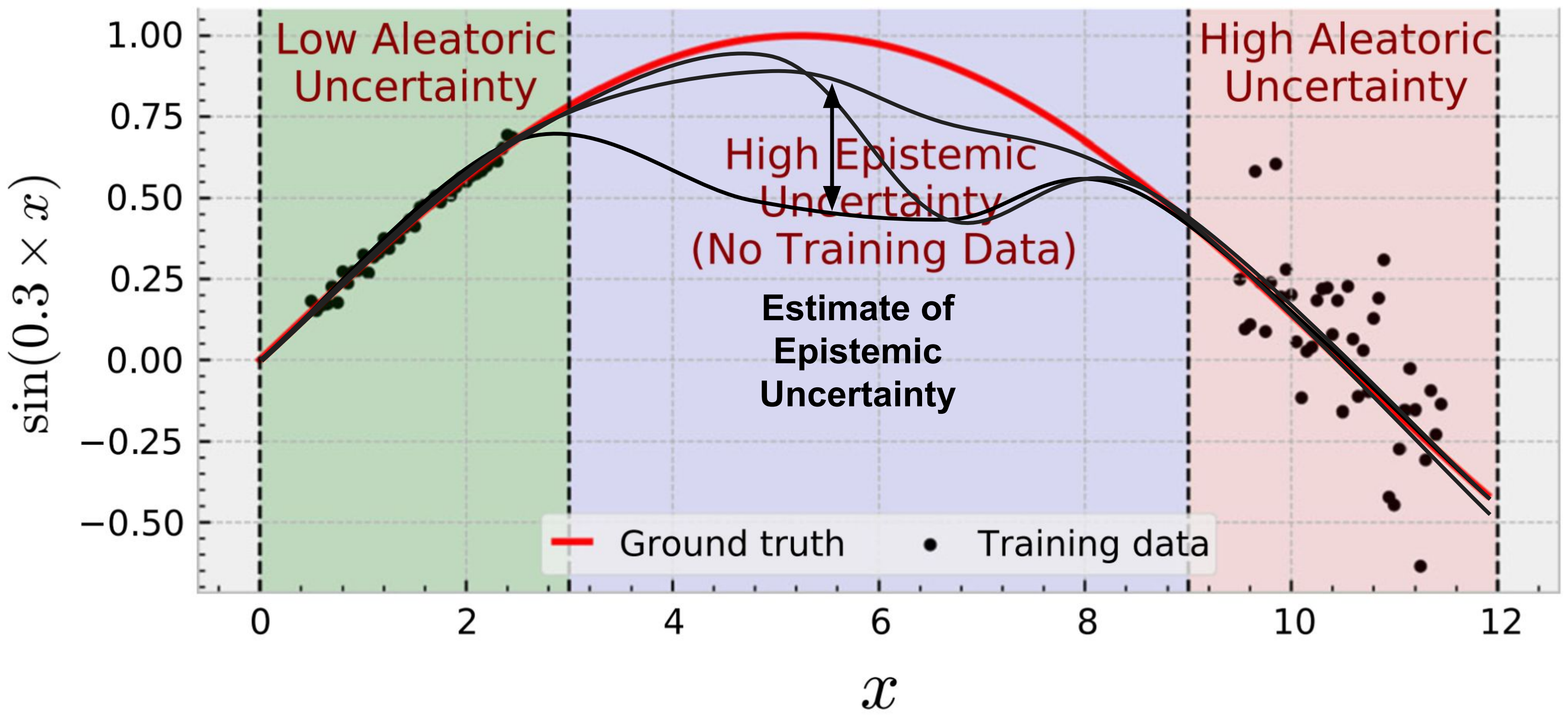
1. Learn a **centralized world-model** on the dataset
2. Use it to generate **synthetic rollouts** train PPO policies



MOMA-PPO - World Model **Ensemble**



Refresher - Epistemic Uncertainty



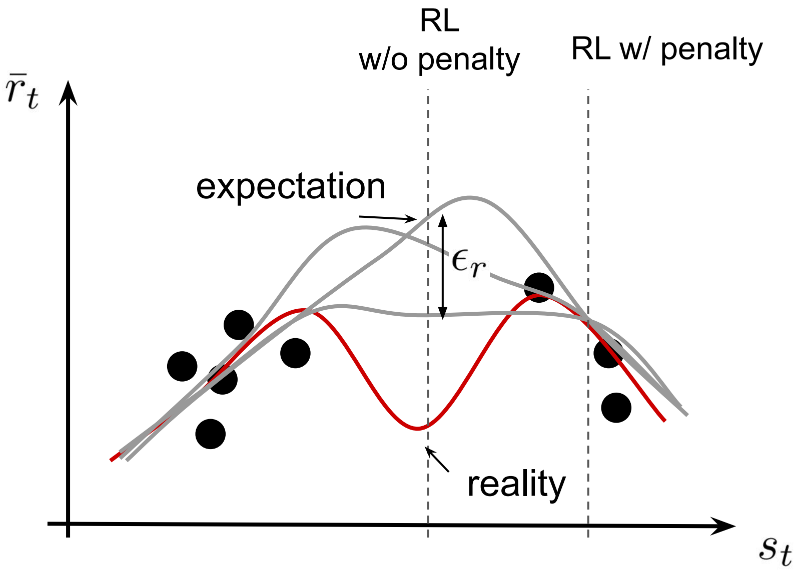
Tuna, Omer Faruk, Ferhat Ozgur Catak, and M. Taner Eskil. "Exploiting epistemic uncertainty of the deep learning models to generate adversarial samples." *Multimedia Tools and Applications* 81.8 (2022): 11479-11500.

MOMA-PPO - World-Model use

Prevent RL algorithm to exploit model's errors

- Epistemic uncertainty penalized reward

$$\tilde{r}_t = \bar{r}_t - \lambda_r \epsilon_r - \lambda_g \epsilon_g$$

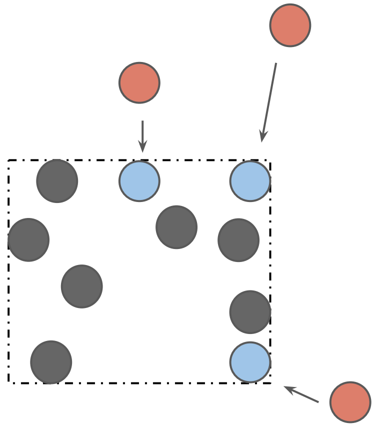


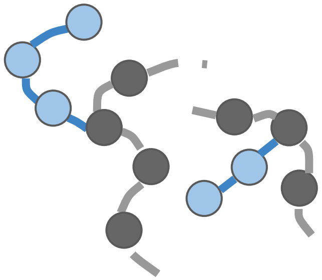
MOMA-PPO - World-Model use

Avoid “unfeasible” data

→Stay close to dataset

- In terms of values → bounding box clipping
- In terms of rollouts
  - Generate from dataset
  - Generate for few steps
  - Early termination (based on WM uncertainty)

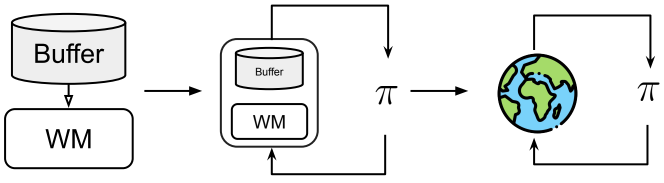




Our Method - MOMA-PPO

In a nutshell

1. Learn a centralized world-model (WM) on the dataset
   - World-model **ensemble** to compute **epistemic uncertainty**
2. Use it to train PPO policies by generating rollouts
   - Sample state in dataset
   - Query current policies for actions
   - Generate transition with WM
   - Clip values to dataset
   - Terminate rollout if its length or uncertainty is above thresholds
   - Penalize reward for uncertainty



Learn WM

Train policies

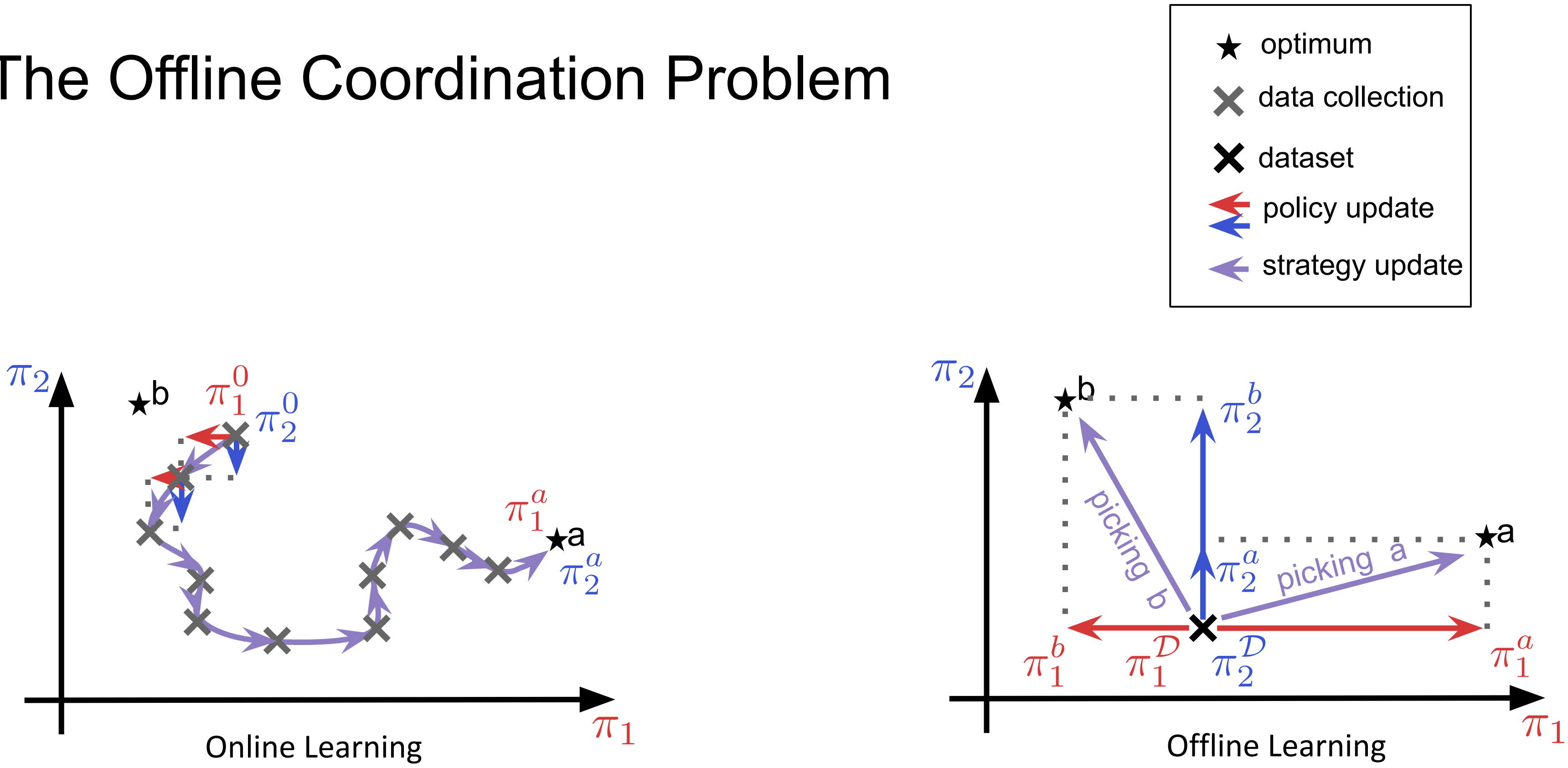
Deployment

Recap

Recap - The Methods

- Model-Free
  - Implicit Q-learning (IQL): single-agent → **centralized execution**
  - **CTDE**-learners:
    - MA-IQL
    - MA-TD3+BC
  - **Independent**-learners
    - IBC
    - ITD3+BC
    - ICQL
    - OMAR
- Model-Based
  - **MOMA-PPO**

The Offline Coordination Problem



Strategy Agreement (SA)
Strategy FineTuning (SFT)

NO interactions during learning

Tasks

Tasks

- Coordination Game (strategy agreement)
    - Three datasets:
        - Favorable: agents go right most of the time
        - Neutral: both act at random
        - Unfavorable: agents go in opposite direction most of the time (agent 1 goes right, agent 2 left)

		a^2	
		$\leftarrow$	$\rightarrow$
a^1	$\leftarrow$	1,1	0,0
	$\rightarrow$	0,0	1,1

    - All datasets have full coverage so **centralized critic** can learn

$$Q(\rightarrow, \rightarrow) = Q(\leftarrow, \leftarrow) = 1 \text{ while } Q(\rightarrow, \leftarrow) = Q(\leftarrow, \rightarrow) = 0$$

Yet, **decentralized actors** still cannot figure out whether to go left or right.

Tasks

Centralized actors and strategy agreement

$$\underbrace{Q(\rightarrow,\rightarrow) = Q(\leftarrow,\leftarrow) = 1}_{\boldsymbol{\pi}(\rightarrow,\rightarrow) = \boldsymbol{\pi}(\leftarrow,\leftarrow) = 0.5} \text{ while } \underbrace{Q(\rightarrow,\leftarrow) = Q(\leftarrow,\rightarrow) = 0}_{\boldsymbol{\pi}(\rightarrow,\leftarrow) = \boldsymbol{\pi}(\leftarrow,\rightarrow) = 0.}$$

Centralized actor controls joint action so always coordinated

Tasks

Decentralized actors and strategy agreement

$$Q(\rightarrow,\rightarrow) = Q(\leftarrow,\leftarrow) = 1 \text{ while } Q(\rightarrow,\leftarrow) = Q(\leftarrow,\rightarrow) = 0$$

$$\boldsymbol{\pi}(\rightarrow,\rightarrow) = \boldsymbol{\pi}(\leftarrow,\leftarrow) = 0.5 \qquad \boldsymbol{\pi}(\rightarrow,\leftarrow) = \boldsymbol{\pi}(\leftarrow,\rightarrow) = 0.$$

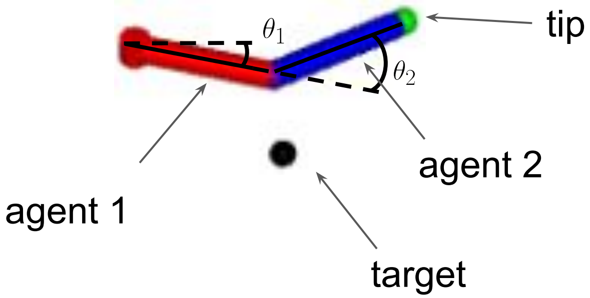
$$\pi_1(\rightarrow)\pi_2(\rightarrow) = \pi_1(\leftarrow)\pi_2(\leftarrow) = 0.5 \qquad \pi_1(\rightarrow)\pi_2(\leftarrow) = \pi_1(\leftarrow)\pi_2(\rightarrow) = 0.$$

$$\pi_1(\rightarrow) = \pi_1(\leftarrow) = \pi_2(\leftarrow) = \pi_2(\rightarrow) = 0.5$$

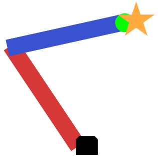
Decentralized actors need to break symmetry
→ coordination occurs by chance (half of the time)

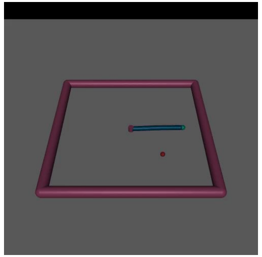
Tasks

- Two Agent reacher (strategy agreement)

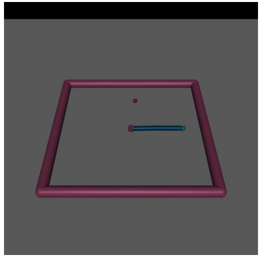


- Dataset is a mix of expert demonstrations

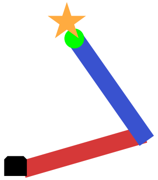




Clockwise experts

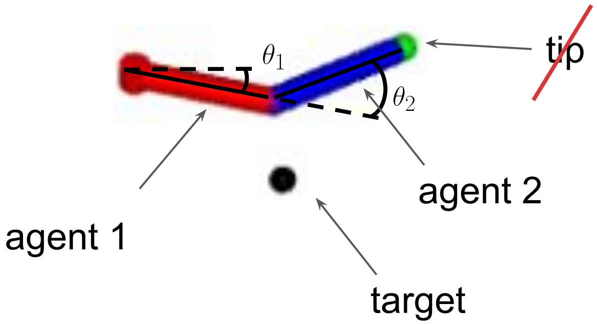


Counter-clockwise experts



Tasks

- Two Agent reacher (strategy agreement)

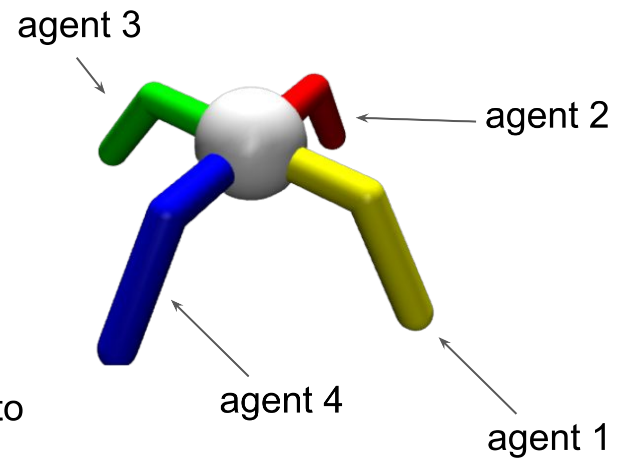


Agents must derive tip position from joint angles

- Partial observability
    - Full Observability (FO): every agent sees everything
    - Partial Observability (PO)
        - Independent: agent sees target and **joint it controls**
        - Leader-only: both agents observe the two joints, but only the **red agent** observes the **target's position**

Tasks

- Four Agent Ant (strategy fine-tuning)
  - Agent decomposition as in MAMuJoCo
- Datasets from D4RL
  - Random : randomly initialized policy
  - Medium : policy at mid-performance training
  - Full-replay : full experience replay buffer used to train to expert performance
  - Expert : expert demonstrations
- Partial observability
  - FO: all the agents observe the full robot
  - PO: each agent observes the limb it controls, **only agent 1** observe torso information (**velocity, heading, etc.**).



Results

Coordination Game

	IQL	MAIQL	IBC	MOMA-PPO
fav.	**1. ± 0.**	**1. ± 0.**	**1. ± 0.**	**1. ± 0.**
neutral	**1. ± 0.**	**0.9 ± 0.1**	0.55 ± 0.11	**1. ± 0.**
unfav.	**1. ± 0.**	0. ± 0.	0. ± 0.	**1. ± 0.**

Two-Agent Reacher

Tasks \ Algorithms		model-free							model-based (ours)
		centralized	CTDE		independent learners				CTDE
		IQL	MAIQL	MATD3+BC	IBC	ITD3+BC	ICQL	IOMAR	**MOMA-PPO**
FO	all-observant	**1.07 ± 0.01**	0.96 ± 0.05	1.04 ± 0.01	1.02 ± 0.01	0.78 ± 0.00	0.48 ± 0.06	0.73 ± 0.01	**1.07 ± 0.01**
PO	independent		**0.92 ± 0.04**	0.59 ± 0.03	0.76 ± 0.04	0.30 ± 0.11	0.46 ± 0.04	0.45 ± 0.02	**0.95 ± 0.06**
	leader-only		0.80 ± 0.05	0.73 ± 0.02	0.84 ± 0.02	0.48 ± 0.04	0.31 ± 0.05	0.39 ± 0.02	**1.00 ± 0.01**

Four-Agent Ant

Tasks \ Algorithms		model-free							model-based (ours)
		centralized	CTDE		independent learners				CTDE
		IQL	MAIQL	MAIQL-ft	IBC	ITD3+BC	ICQL	IOMAR	**MOMA-PPO**
FO	ant-random	0.12 ± 0.00	0.28 ± 0.01	0.28 ± 0.03	0.31 ± 0.00	0.22 ± 0.02	0.08 ± 0.00	0.08 ± 0.00	**0.52 ± 0.07**
	ant-medium	0.97 ± 0.02	0.85 ± 0.02	0.81 ± 0.02	0.84 ± 0.01	1.04 ± 0.00	0.88 ± 0.12	1.10 ± 0.03	**1.29 ± 0.06**
	ant-full-replay	1.22 ± 0.02	0.77 ± 0.21	0.95 ± 0.13	1.20 ± 0.01	1.33 ± 0.01	1.21 ± 0.02	1.30 ± 0.00	**1.42 ± 0.07**
	ant-expert	1.26 ± 0.01	1.24 ± 0.00	1.06 ± 0.07	1.24 ± 0.00	1.25 ± 0.02	0.73 ± 0.15	1.16 ± 0.01	**1.49 ± 0.01**
PO	ant-random		0.31 ± 0.00	0.34 ± 0.04	0.31 ± 0.00	0.31 ± 0.00	0.17 ± 0.02	0.21 ± 0.02	**0.42 ± 0.05**
	ant-medium		0.14 ± 0.02	0.11 ± 0.01	0.17 ± 0.01	0.22 ± 0.05	0.09 ± 0.02	0.06 ± 0.01	**0.54 ± 0.19**
	ant-full-replay		0.18 ± 0.02	-0.07 ± 0.10	0.21 ± 0.02	0.20 ± 0.01	0.09 ± 0.01	0.11 ± 0.02	**0.46 ± 0.10**
	ant-expert		-0.16 ± 0.01	-0.23 ± 0.02	0.05 ± 0.04	0.16 ± 0.00	0.11 ± 0.03	0.10 ± 0.01	**0.18 ± 0.00**

Results – Offline coordination

- MA - Model-Free methods **(Fail)**
  - Fail at Strategy Agreement (SA)
  - Fail at Strategy Fine-Tuning (SFT)

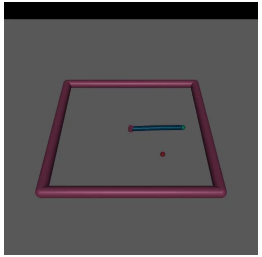
- Fully Centralized - Model-Free (Mixed)
  - Bypasses SA
  - **Fails at SFT!**

- Model-Based method (MOMA-PPO) **(Success)**
  - Solves both SA and SFT

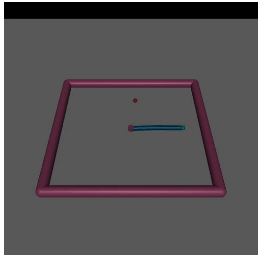
- **Model-Based > Model-Free (even fully centralized!)**

Illustrative Rollouts

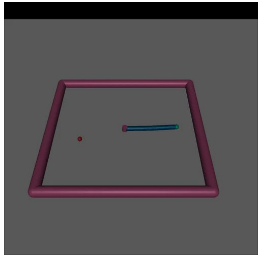
- Partially Observable Two Agent Reacher (strategy agreement)



Clockwise experts

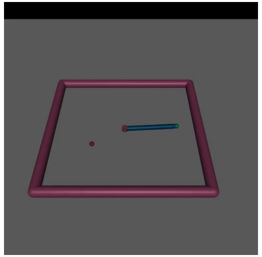


Counter-clockwise experts



ITD3+BC failure

Agents picked different strategies

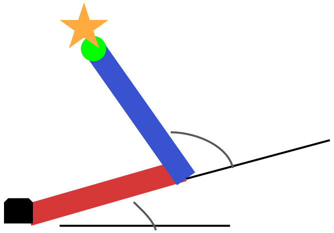


MOMA-PPO success

Both clockwise and counter-clockwise → Improves on experts

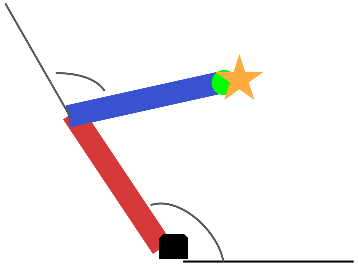
Illustrative Rollouts

Better than expert?



Counter-clockwise

Optimal

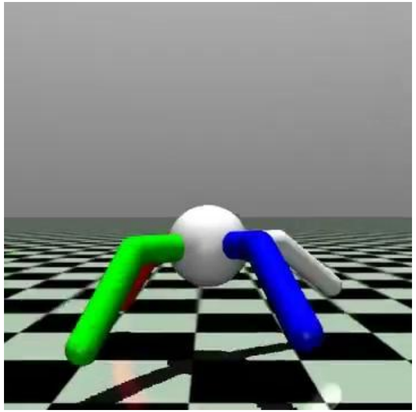


Clockwise

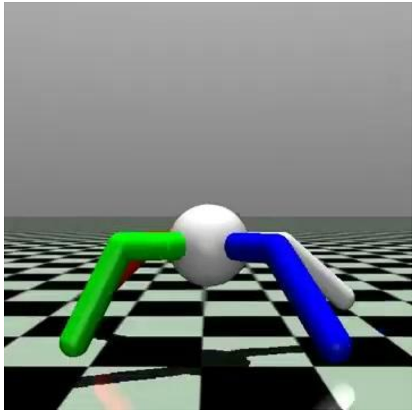
Suboptimal expert

Illustrative Rollouts

- Partially observable Four Agent Ant (strategy fine-tuning)



ITD3+BC: team failed to fine-tune and runs in circles



MOMA-PPO: white agent "steers" the team

Insights

- Strategy Fine-Tunining **occurs when you need to adapt** from the dataset
  - Suboptimal datasets
  - Partial observability
    - “Steering” behavior in partial observable ant
  - Otherwise, simple imitation is enough
    - BC performance is close to model-free performance

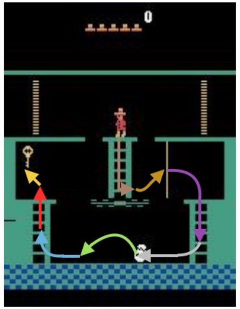
- Partial observability induces more **state ambiguity**
  - More difficult to break symmetry / course correct

- MOMA-PPO performance related to **dataset’s coverage/diversity** rather than **“expertness”**
  - Partial Observable Ant → random dataset > expert dataset
  - Best dataset is most likely the biggest one possible
    - Mix all the datasets you have (random, medium, expert, replay, etc.)
    - Train MOMA-PPO on it
    - Compare to model-free methods

Discussion on MOMA-PPO

Dataset Agents (data collectors)	$\neq$	MOMA-PPO agents
Real-world interactions		**Generated** interactions
Initial state distribution given by the **environment**		Initial state distribution given by the **dataset**

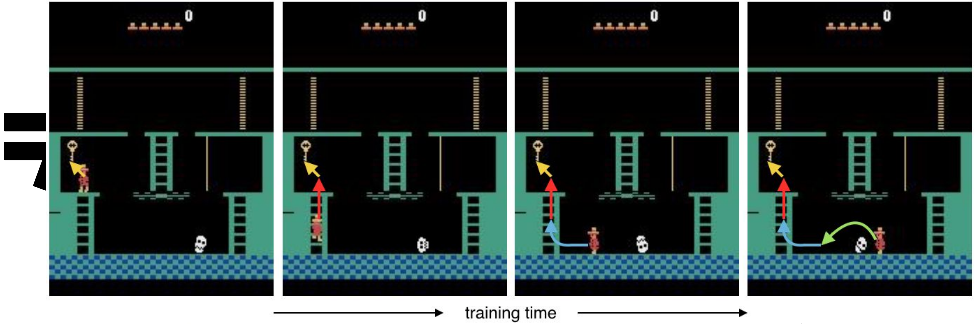
Discussion on MOMA-PPO



Dataset
Agents
data collectors)

al-world interactions

Initial state distribution given
by the **environment**



Initial state distribution
given by the **dataset**

Salimans, Tim, and Richard Chen. "Learning montezuma's revenge from a single demonstration." *arXiv preprint arXiv:1812.03381* (2018).

Discussion on MOMA-PPO

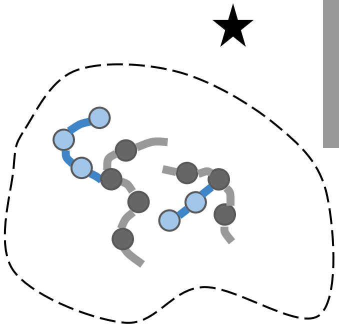
Dataset Agents (data collectors)	$\neq$	MOMA-PPO agents
Real-world interactions		**Generated** interactions
Initial state distribution given by the **environment**		Initial state distribution given by the **dataset**
Reward defined by the **task**		**Uncertainty averse** reward
Unconstrained exploration		Limited number of **steps away from dataset**

Discussion on MOMA-PPO

Dataset
(data collectors)
agents

$\neq$

MOMA-PPO agents

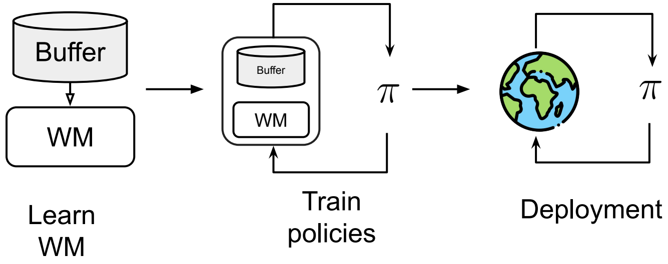


Limited number of **steps away from dataset**

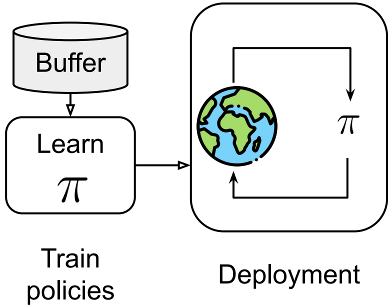
Discussion on MOMA-PPO

- Why PPO?
  - Online RL (PPO) > Offline RL (IQL, CQL)?
    → different learning paradigm

MOMA:



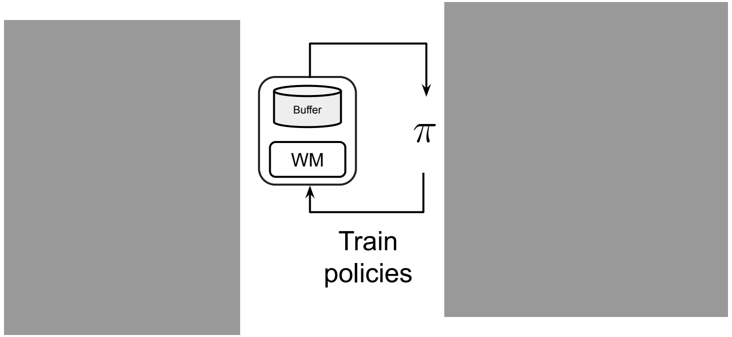
Offline Model-Free:



Discussion on MOMA-PPO

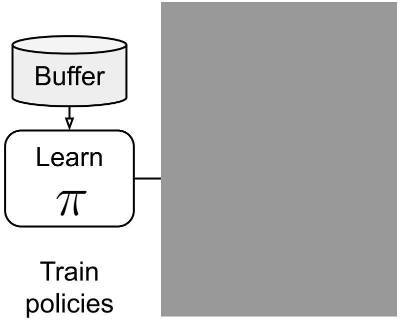
- Why PPO?
  - Online RL (PPO) > Offline RL (IQL, CQL)?
    → different learning paradigm

MOMA:



Interactively collected non-stationary data
→ **Online RL**

Offline Model-Free:



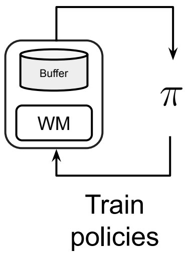
“Supervized” on static dataset

Discussion on MOMA-PPO

- Why PPO?
  - On-policy (PPO) > Off-policy (SAC, TD3)?

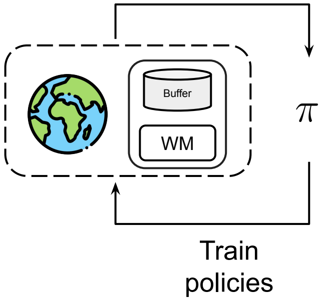
    → Model-based serves a different purpose

MOMA:



Adapt from dataset
Coordinate multiple agents
→ Robust and effective
→ **MA-PPO**

Online Dyna:



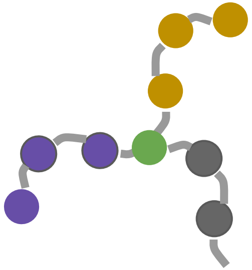
Sample efficiency wrt.
Real environment
→ **Off-policy**

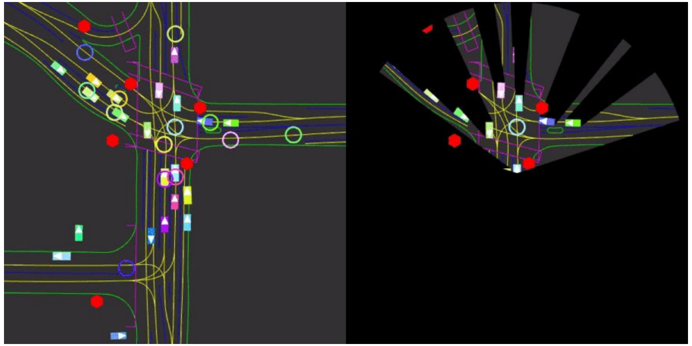
Yu, Chao, et al. "The surprising effectiveness of ppo in cooperative multi-agent games." *Advances in Neural Information Processing Systems* 35 (2022): 24611-24624.

Limitations

- MOMA-PPO takes **longer to train** (3-4 times wall-clock time) than the baselines
    - Need to **generate rollouts** through the world-model vs. supervised on **fixed dataset**

- MOMA-PPO performance still depends on the **wold-model's accuracy** and **generalization**
    - Strongly related to the **dataset's coverage** (but not necessarily dataset performance)
    - **World-model learning can be challenging**
        - Stochastic multi-modal environments
        - Complex dynamics (simulate road-users that we do not control)



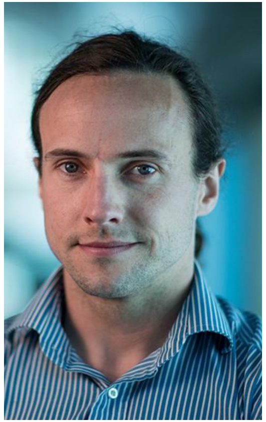


Future work

- Analyze more in-depth **model-free failure cases** (potential model-free solutions?)
- BC efficiency begs the question of how much of other methods' complexity is mandated in practice
- Reduce training time by using **synthetic interactions for coordination only** while mainly learning on dataset?
  - TD3 on dataset + (some) synthetic interactions
    - How to mix?
    - Monitor?
    - Avoid exploitation/extrapolation error while still improving on dataset?
    - Many trade-offs to investigate…
- Move to more complex tasks and environments
  - Stochastic
  - Observation/action space
  - Complex dynamics
  - Real world data …

Summary

- **Offline MARL** is an attractive solution to many **real world problems**.
- However, extending **current model-free offline RL** methods to the multi-agent setting **fail** at offline coordination problems.
- Our **model-based** approach solves this by restoring **agent-to-agent interactions** during learning.
    - Agents interact through the world-model



Jakob
Foerster

Derek
Nowrouzezahrai

Amy
Zhang

References and images

IMAGES:

- https://discover.rbcroyalbank.com/wp-content/uploads/banner-small-self-driving-cars_402x.jpg
- https://codibly.com/news-insights/what-are-smart-grids/
- https://www.vectorstock.com/royalty-free-vector/financial-stock-market-capital-markets-trading-vector-22947946
- https://www.newyorker.com/podcast/the-new-yorker-radio-hour/the-pandemic-at-three-who-got-it-right
- https://pngtree.com/so/road-accident

REFERENCES

- Vinitsky, Eugene, et al. "Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world." *Advances in Neural Information Processing Systems* 35 (2022)
- Levine, Sergey, et al. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems." *arXiv preprint arXiv:2005.01643* (2020).
- Hu, Hengyuan, et al. "“other-play” for zero-shot coordination." *International Conference on Machine Learning*. PMLR, 2020.
- Cui, Brandon, et al. "Adversarial Diversity in Hanabi." *The Eleventh International Conference on Learning Representations*. 2022.
- Tuna, Omer Faruk, Ferhat Ozgur Catak, and M. Taner Eskil. "Exploiting epistemic uncertainty of the deep learning models to generate adversarial samples." *Multimedia Tools and Applications* 81.8 (2022): 11479-11500.
- Yu, Chao, et al. "The surprising effectiveness of ppo in cooperative multi-agent games." *Advances in Neural Information Processing Systems* 35 (2022): 24611-24624.

Thank you!