Efficient Model-based Multi-agent Reinforcement Learning via Optimistic Equilibrium Computation

Pier Giuseppe Sessa, Maryam Kamgarpour, Andreas Krause

International Conference on Machine Learning (ICML), 2022

sycamore lab
SYSTEMS CONTROL AND MULTIAGENT OPTIMIZATION RESEARCH
LAS Learning & Adaptive Systems
ETH zürich
EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Setup: Episodic MA-RL

- General-sum N-players Markov game, with horizon H:
  - Continuous action and state spaces: $\mathscr{A}^i, i = 1,\ldots, N, \ \ \mathscr{S}$
  - Π^i = space of all policies $\pi^i : \mathscr{S} \to \mathscr{A}^i$

- Environment **transition function** $f : \Pi_{i=1}^N \mathscr{A}^i \times \mathscr{S} \to \mathscr{S}$ is a-priori **unknown** and can only be learned via interaction rounds

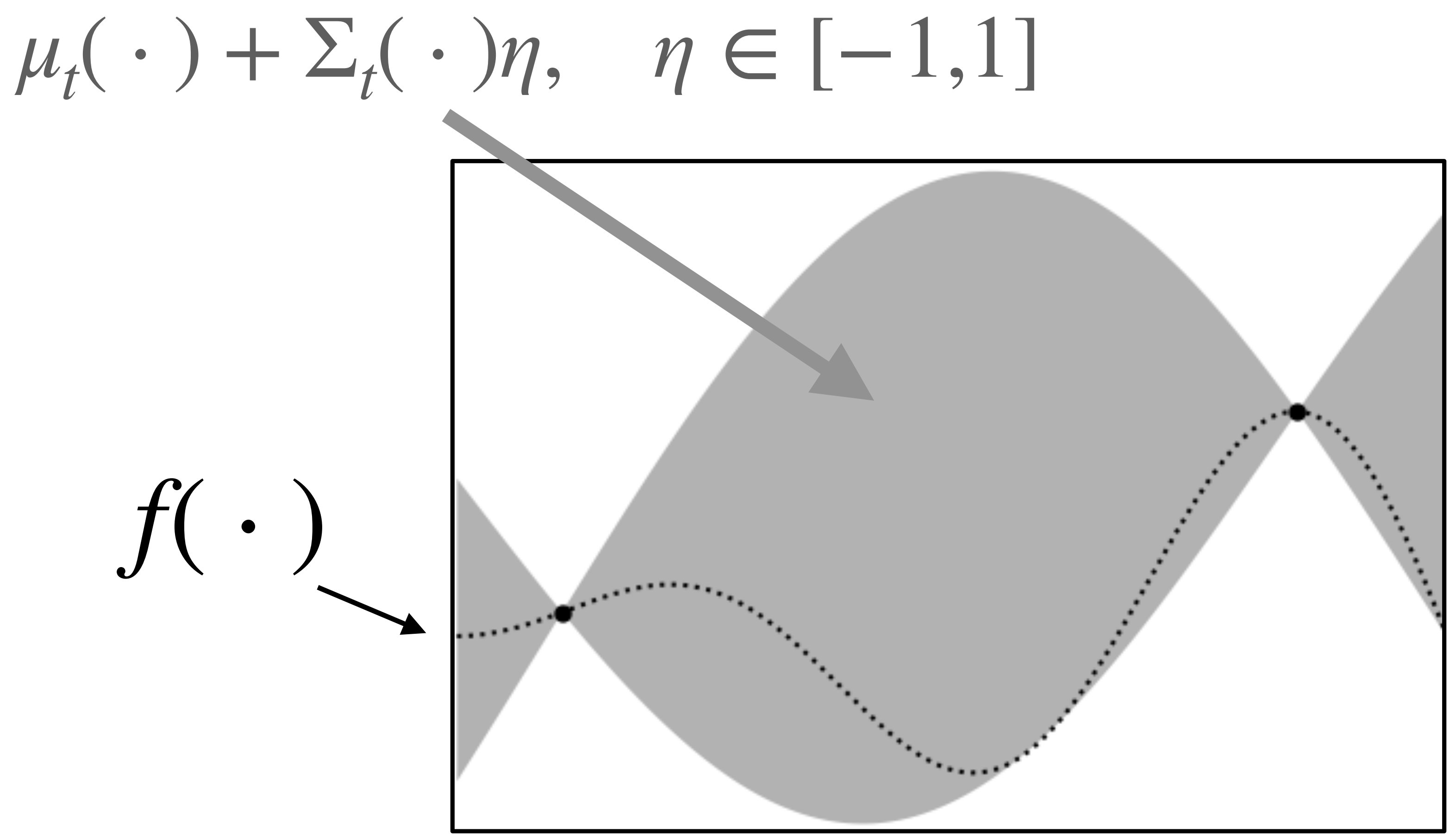
- At each round t:
  - agents play using policies $\{\pi_t^i, i = 1,\ldots N\}$
  - we observe H transitions $\{(\mathbf{a}_h, s_h), s_{h+1}\}$

- Dynamic regret of agent-i:

distance from
best-response

$$R^i(T) := \sum_{t=1}^{T} \boxed{\max_{\pi \in \Pi^i} \mathbb{E}_{\pi_t^{-i}}\left[V^i(\pi, \pi_t^{-i})\right] - \mathbb{E}_{\pi_t^1,\ldots,\pi_t^N}\left[V^i(\pi_t^1, \ldots, \pi_t^N)\right]}$$

H-MARL algorithm



1) Obtain calibrated model for environment's transition function (e.g. via RKHS regression, deep ensembles, …):

2) Build <u>optimistic value functions</u> for the agents as:

auxiliary function

$$\mathrm{UCB}_t^i(\boldsymbol{\pi}) = \max_{\eta(\cdot)\in[-1,1]^p} \mathbb{E}\left[\sum_{h=0}^{H-1} r^i(s_h, \mathbf{a}_h)\right] \tag{2}$$

$$\text{s.t.} \quad \mathbf{a}_h = \boldsymbol{\pi}(s_h)$$

plausible states' trajectory according to learned model

$$s_h = \mu_t(s_{h-1}, \mathbf{a}_{h-1}) + \beta_t \cdot \Sigma_t(s_{h-1}, \mathbf{a}_{h-1})\eta(s_{h-1}, \mathbf{a}_{h-1}) + w_h \,.$$

- We propose a practical implementation via sampling of η

H-MARL algorithm

Algorithm 1 The H-MARL algorithm

Require: Agents' policy spaces $\Pi^1, \ldots, \Pi^N$.

1: **for** $t = 1, \ldots, T$ **do**

2: $\mathcal{P}_t \leftarrow$ Find-CCE$\big(\text{UCB}^1_{t-1}(\cdot), \ldots, \text{UCB}^N_{t-1}(\cdot)\big)$, with $\text{UCB}^i_{t-1}(\cdot)$ defined in Eq. (2).

3: Episode rollout using policies

$$\boldsymbol{\pi}_t = (\pi^1_t, \ldots, \pi^N_t) \sim \mathcal{P}_t$$

4: Update transition model $\mu_t(\cdot,\cdot)$, $\Sigma_t(\cdot,\cdot)$, using observed H transitions.

Compute equilibrium of **optimistic hallucinated** game:

- Can simulate it arbitrarily often, e.g. using model-free approaches
- In practice, could use independent DQN learning, MADDPG, etc.

Thm: Each agent's dynamic regret is bounded, with prob. $1-\delta$, as:

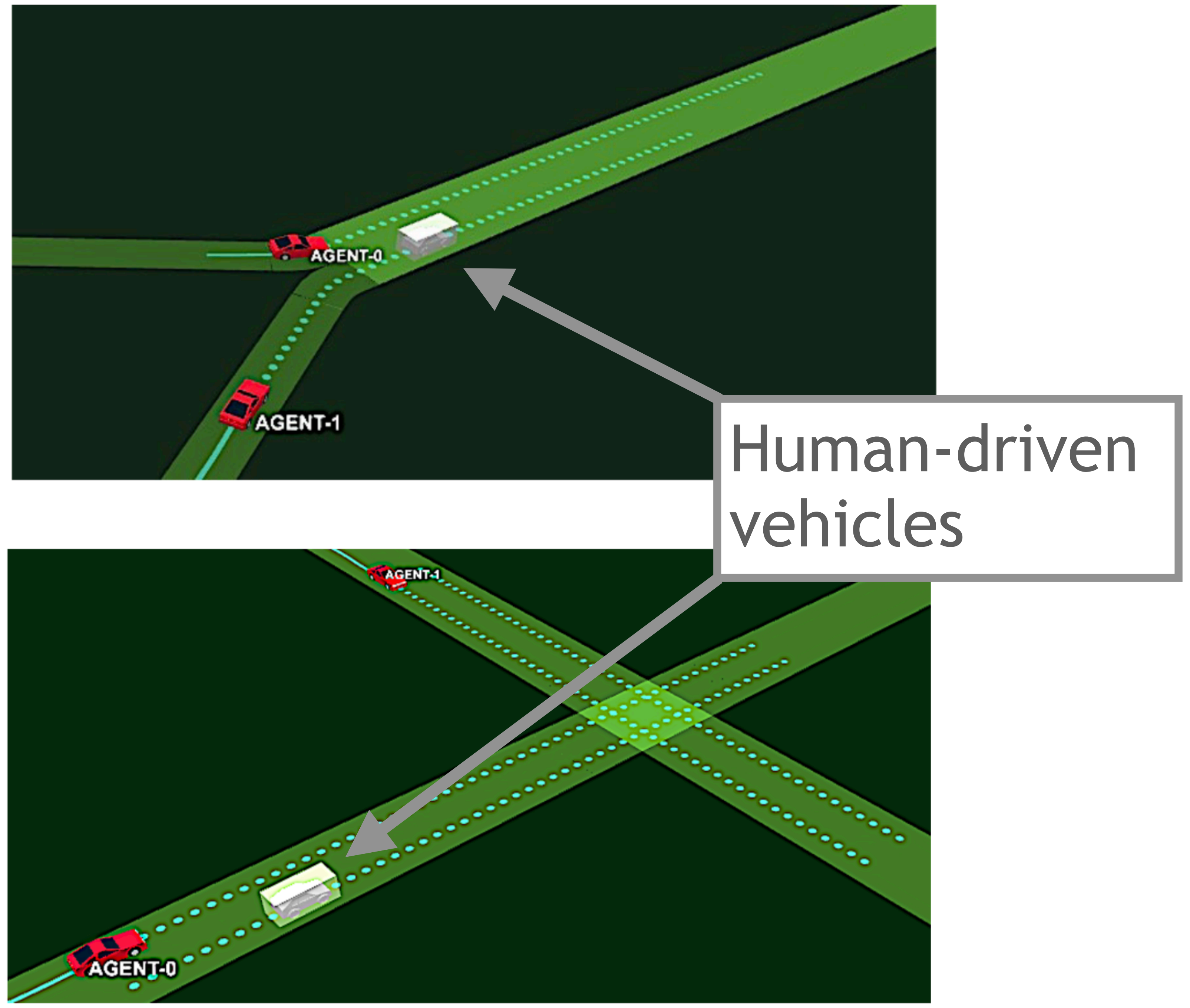
$$R^i(T) \leq \bar{L} H^{1.5} \sqrt{T\, I_T}$$

Lipschitz constant:
$\bar{L} = \mathcal{O}\big(N^{H/2} L_\pi^{H/2} (\bar{\beta}^H L_\sigma^H + L_f^H) + \log(1/\delta)\big)$

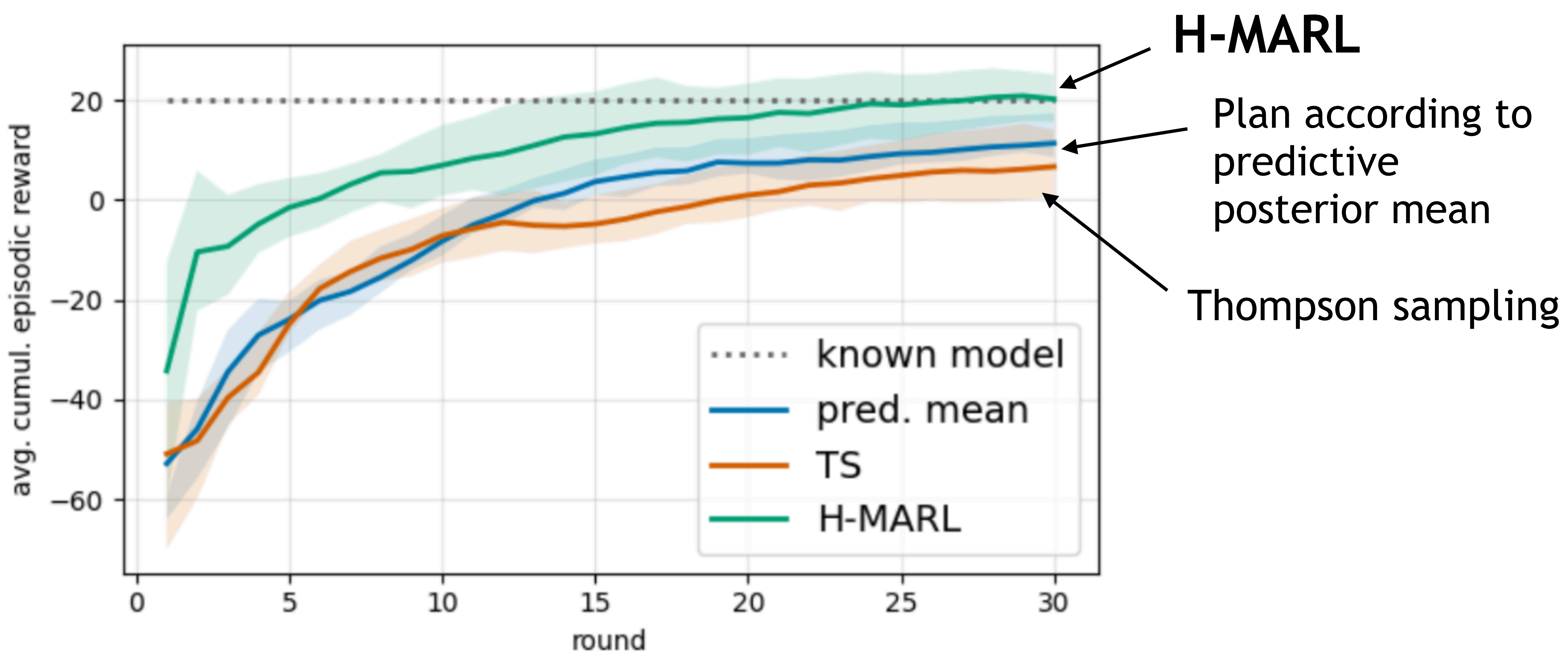
Sample-complexity of the transition fcn.
(can be bounded for most kernels)

Experiments

- SMARTS [Zhou et al. 2020] environment:



—> Human driving behaviour is a-priori unknown and can only be inferred by sequential interaction rounds



	Avg. completion rate during learning	Avg. completion time during learning
pred. mean	72.0 %	8.90 s
TS	69.9 %	8.87 s
H-MARL	**80.9** %	**8.66** s

- **H-MARL** displays faster learning than considered baselines. Higher completion rates and lower completion times.