Adversarial Inverse Reinforcement Learning for Mean Field Games

Yang Chen, Libo Zhang, Jiamou Liu, Michael, Witbrock

University of Auckland

School of
Computer Science

Markov Games: Multi-agent Reinforcement Learning

A Markov game (stochastic game) is a tuple $(\{S_i\}_{i=1}^N, \{A_i\}_{i=1}^N, \{r_i\}_{i=1}^N, P, \gamma, \eta)$, where

- S_i is the local state set for the ith ($i = 1, 2, \ldots, N$) agent;
- A_i is the action set for the ith agent;
- $r_i(\boldsymbol{s}, \boldsymbol{a})$ is the reward function for the ith agent, where $\boldsymbol{s} = (s_1, \ldots, s_N)$ is the joint state and $\boldsymbol{a} = (a_1, \ldots, a_N)$ is the joint action;
- $P(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$ specifies the transition probabilities between joint states conditioned on joint actions;
- $\gamma \in (0, 1)$ is the discount factor;
- $\eta \in \Delta(S_1 \times \cdots \times S_N)$ specifies the initial distribution of joint states.

Nash equilibrium: A policy $\pi_i(a_i|\boldsymbol{s})$ guides an agent's selection of actions. A joint policy $\boldsymbol{\pi}^\star = (\pi_1^\star, \ldots, \pi_N^\star)$ is a Nash equilibrium if no agent can gain rewards by unilaterally deviating from it:

$$\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_i(\boldsymbol{s}, \boldsymbol{a}) \middle| \boldsymbol{\pi}^\star, P, \eta\right] \geq \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_i(\boldsymbol{s}, \boldsymbol{a}) \middle| \pi_i, \boldsymbol{\pi}_{-i}^\star, P, \eta\right], \forall i = 1, 2, \ldots, N, \text{ and valid } \pi_i$$


Human teams and


Markets and economies


Transportation networks

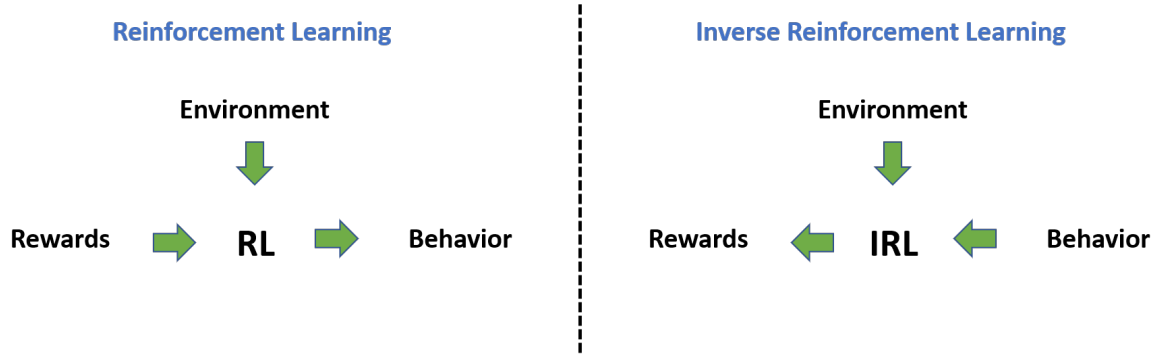

Distributed software


Communication networks


Robotic teams

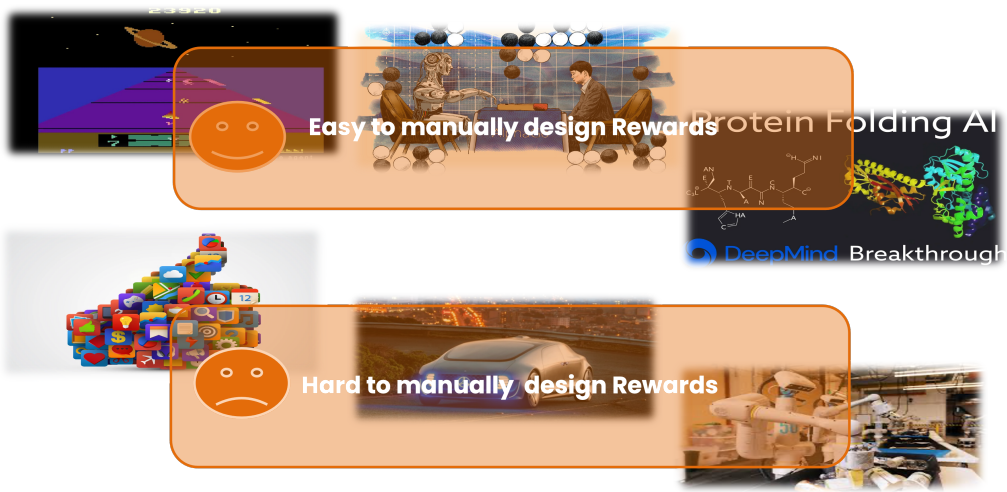
Multi-agent Inverse Reinforcement Learning

Problem Description: Infer the reward functions $r_1, \ldots, r_N$ from observed behaviour $(s_0, a_0, \ldots, s_T, a_T)$ sampled via $a_t \sim \pi^\star$, $s_0 \sim \eta$ and $s_t \sim P$.



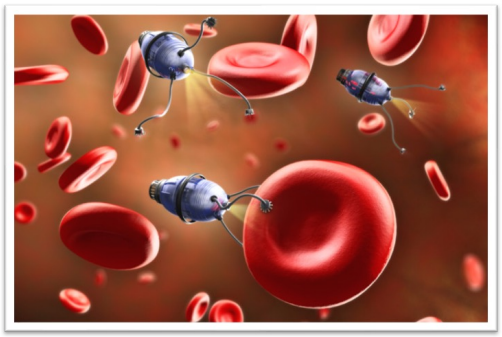
Motivations:

- Understand objectives of interacting agents
- Design environments for AI agents so that our expected behaviour emerge

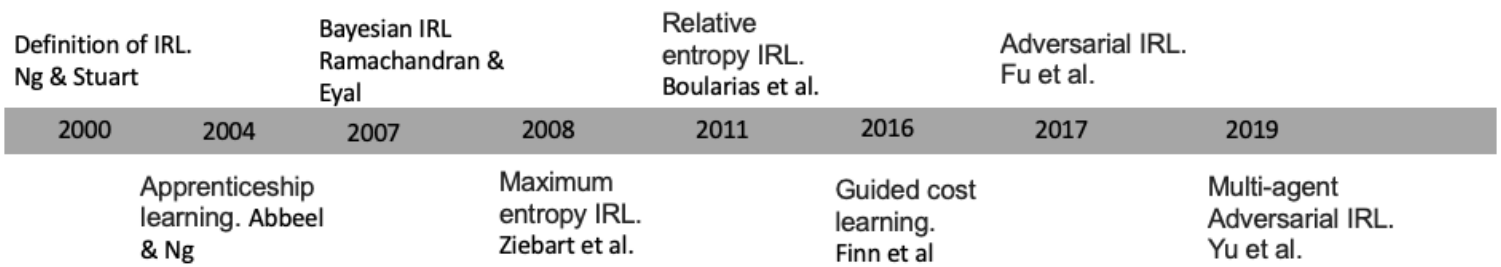


Curse of Dimensionality

Challenge. The joint state-action space grows exponentially as N goes large, making the inference of reward functions $r_i(s_1, \ldots, s_N, a_1, \ldots, a_N)$ intractable.



Historical Review of IRL.



Existing methods are not suitable for IRL with many agents!

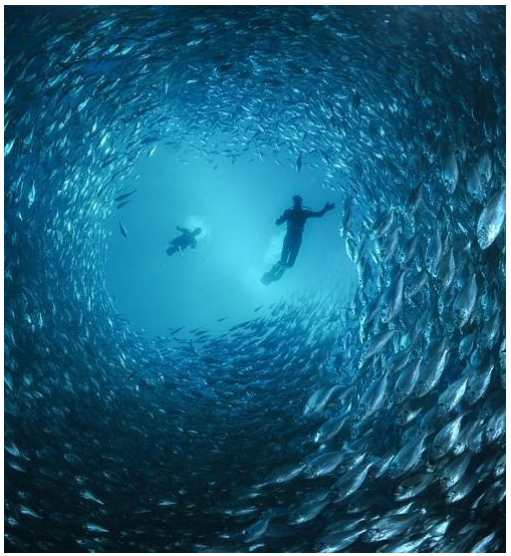
Mean Field Games

Mean Field Approximation.

- Assumptions:
  1. The number of agents tends to infinity.
  2. Agents are identical.

- Collapse a joint state into a mean field (an empirical distribution):

$$(s_1, \ldots, s_N) \Rightarrow \mu(s) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{s_i = s}$$



Mean Field Games

A mean field game is a tuple $(S, A, r, P, \gamma, \mu_0)$:

- all agents have the same state set S and action set A;
- all agents have the same reward function $r(s, a, \mu)$;
- $P(s'|s, a, \mu)$ is the transition function;
- $\gamma \in (0, 1)$ is the discount factor;
- $\mu_0 \in \Delta(S)$ is the initial mean field.

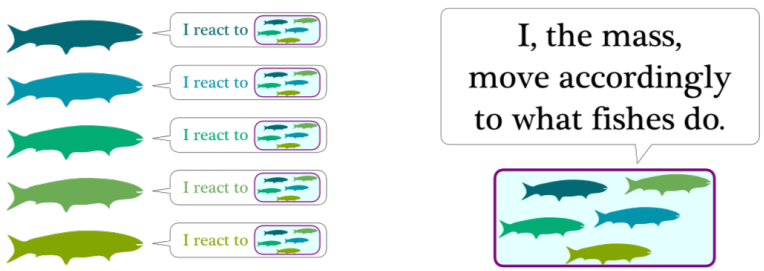


Figure: http://www.science4all.org/article/mean-field-games/

Mean Field Nash Equilibrium

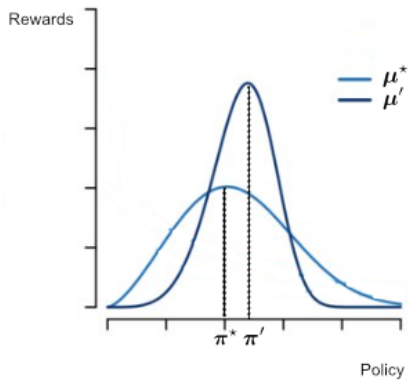
A mean field Nash equilibrium is a pair of policy $\boldsymbol{\pi}^\star = (\pi_0^\star, \pi_1^\star, \ldots, \pi_T^\star)$ and mean field flow $\boldsymbol{\mu}^\star = (\mu_0^\star, \mu_1^\star, \ldots, \mu_T^\star)$ that satisfies two conditions:

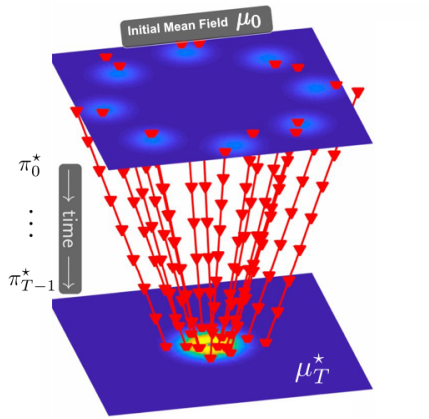
1. $\boldsymbol{\pi}$ is optimal w.r.t. $\boldsymbol{\mu}$:

$$\boldsymbol{\pi}^\star \in \arg\max_{\boldsymbol{\pi}} J(\boldsymbol{\mu}^\star, \boldsymbol{\pi}) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t^\star);$$

2. $\boldsymbol{\mu}$ is consistent with $\boldsymbol{\pi}$:

$$\mu_{t+1}(s') = \sum_{s \in S} \mu_t(s) \sum_{a \in A} \pi_t(a|s) P(s'|s, a, \mu_t).$$



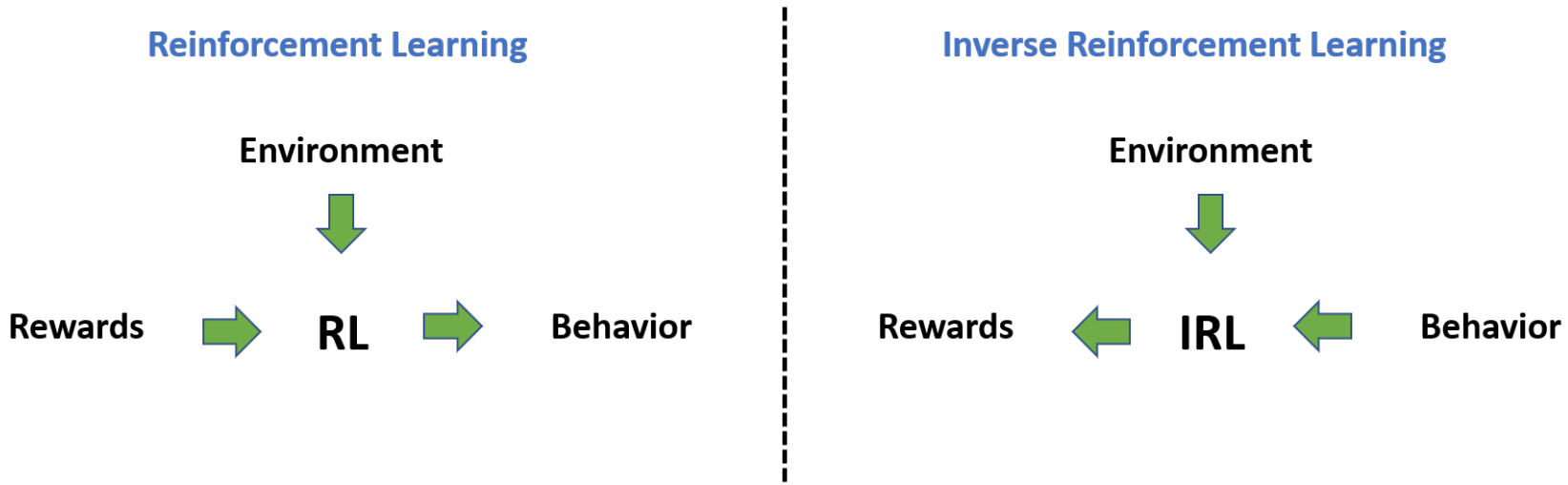


Figures: Ruthotto, L., Osher, S. J., Li, W., Nurbekyan, L., & Fung, S. W. (2020). A machine learning framework for solving high-dimensional mean field game and mean field control problems. Proceedings of the National Academy

IRL for MFGs

Input: A collection of observed state-action trajectories $\{\tau = s_0, a_0, \ldots, s_T, a_T\}$ generated by a policy $\boldsymbol{\pi}^E$ and a mean field flow $\boldsymbol{\mu}^E$ via $s_0 \sim \mu_0, a_t \sim \pi_t(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t, \mu_t)$.

Output: A reward function $r(s, a, \mu)$ under which $(\boldsymbol{\mu}^E, \boldsymbol{\pi}^E)$ forms a mean field Nash equilibrium.



The Reduction from MFGs to MDPs

Basic idea: View an MFG as a single-player decision-making process of the population.

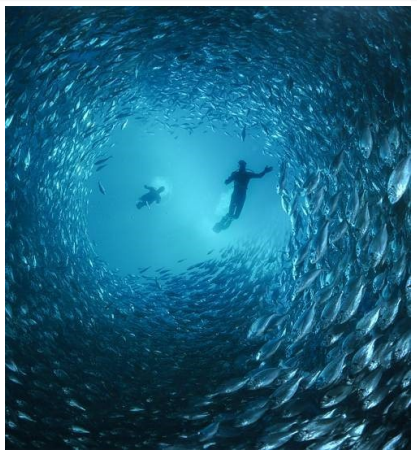
Given an MFG $(S, A, r, P, \gamma, \mu_0)$, an Markov decision process can be constructed as follows:

- State: μ_t, i.e., the state at step t is the mean field μ_t.
- Action: π_t, i.e., the action is a per-step policy in MFG.
- Reward: $\bar{r}(\mu, \pi) = \sum_{s \in S} \mu(s) \sum_{a \in A} \pi(a|s) r(s, a, \mu)$, i.e., population's averaged rewards.
- Transition: $\mu_{t+1}(s') = \sum_{s \in S} \mu_t(s) \sum_{a \in A} \pi_t(a|s) P(s'|s, a, \mu_t)$.
- Policy: $\pi_{MDP} : \mu \mapsto \pi$

Population-Level IRL for MFGs:

$$\max_{\bar{r}} \left[J(\boldsymbol{\mu}^E, \boldsymbol{\pi}^E) - \max_{\boldsymbol{\mu}, \boldsymbol{\pi}} J(\boldsymbol{\mu}, \boldsymbol{\pi}) \right]$$

s.t. $\mu_{t+1}(s') = \sum_{s \in S} \mu_t(s) \sum_{a \in A} \pi_t(a|s) P(s'|s, a, \mu_t)$.



Yang, J., Ye, X., Trivedi, R., Xu, H., & Zha, H. (2018, February). Learning Deep Mean Field Games for Modeling Large Population Behavior. InÂ International Conference on Learning Representations.

Individual-Level IRL for MFGs

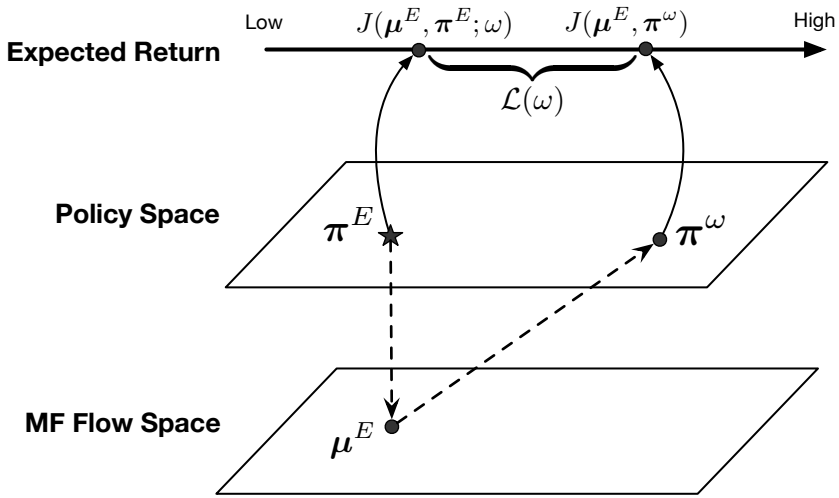
Theorem 1 (Informal).

The reduction from MFGs to MDPs holds only for the fully cooperative setting, i.e., all agents aim to optimise the population's average rewards.

Individual-Level IRL for MFGs

Theorem 1 (Informal).

The reduction from MFGs to MDPs holds only for the fully cooperative setting, i.e., all agents aim to optimise the population's average rewards.



A practical algorithm using bi-level optimisation:

$$\max_{\omega} \mathcal{L}(\omega) = \mathbb{E}_{\tau \sim D}\left[\sum_{t=0}^{T} r_\omega(s_t, a_t, \hat{\mu}_t^E)\right] - J(\hat{\mu}_t^E, \pi^\omega)$$

Update policy

Update reward

Theorem 2 (Informal).

With probability 1 as the number of demonstrations tends to infinity, the optimal solution tends to be the optimal parameter.

Individual-Level Inverse Reinforcement Learning for Mean Field Games. Yang Chen, Libo Zhang, Jiamou Liu, Shuyue Hu. The 21st International Conference on Autonomous Agents and Multi-agent Systems. AAMAS 2022.

Imperfect Observed Behaviour

- What if the demonstrations are imperfect?
- How can we model and quantify such imperfection?
- Can we design a many-agent IRL method that is able to handle imperfect demonstrated behaviours?

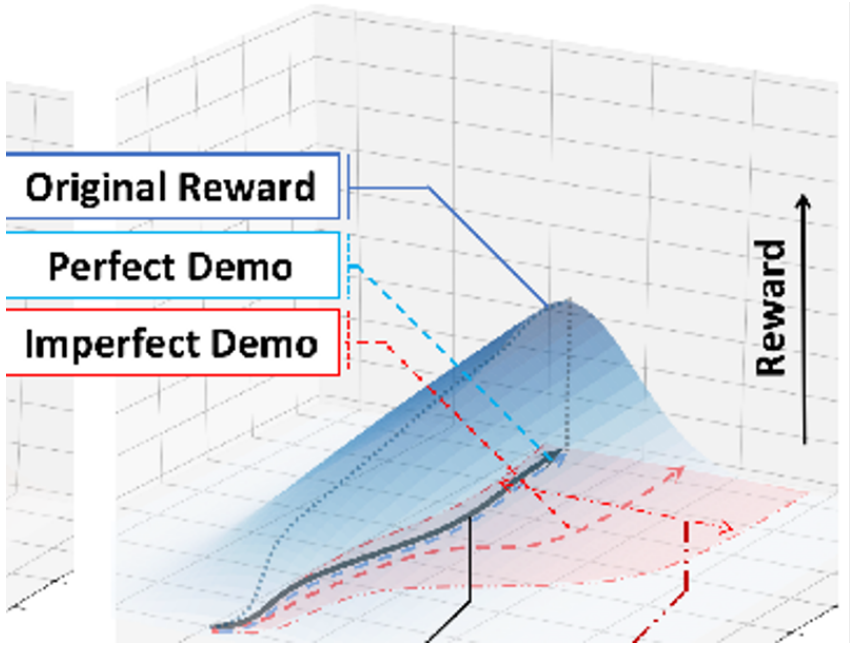


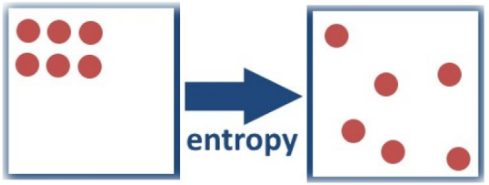
Figure: Jing, M., Ma, X., Huang, W., Sun, F., Yang, C., Fang, B., & Liu, H. Reinforcement learning from imperfect demonstrations under soft expert guidance. In Proceedings of the AAAI conference on artificial intelligence (pp. 5109–5116).

Use Entropy to Reason about Uncertainties

Augment rewards with Policy Entropy:

$$\max_{\boldsymbol{\pi}} \tilde{J}(\boldsymbol{\mu}, \boldsymbol{\pi}) \triangleq \mathbb{E}_{a_t \sim \pi_t} \left[\sum_{t=0}^{T} r(s_t, a_t, \mu_t) + \beta \mathcal{H}(\pi_t(\cdot|s)) \right]$$

$$\beta > 0, \mathcal{H}(\pi_t(\cdot|s)) = -\sum a \in A \pi_t(a|s) \log \pi_t)(a|s)$$

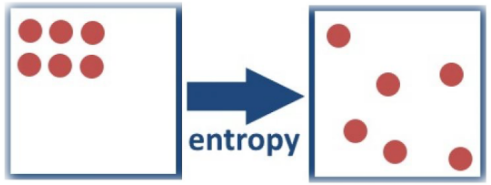


Use Entropy to Reason about Uncertainties

Augment rewards with Policy Entropy:

$$\max_{\pi} \tilde{J}(\mu, \pi) \triangleq \mathbb{E}_{a_t \sim \pi_t}\left[\sum_{t=0}^{T} r(s_t, a_t, \mu_t) + \beta \mathcal{H}(\pi_t(\cdot|s))\right]$$

$$\beta > 0, \mathcal{H}(\pi_t(\cdot|s)) = -\sum a \in A \pi_t(a|s) \log \pi_t)(a|s)$$



Entropy-regularised MFNE

A pair of MF flow and policy $(\tilde{\mu}^\star, \tilde{\pi}^\star)$ is called an entropy-regularised MFNE (ERMFNE) if it satisfies:

1. Maximum entropy-regularised rewards:

$$\tilde{\pi}^\star \in \arg\max_{\pi} \tilde{J}(\pi, \tilde{\mu}^\star)$$

2. Population consistency

$$\tilde{\mu}^\star_{t+1}(s') = \sum_{s \in S} \tilde{\mu}^\star_t(s) \sum_{a \in A} \pi_t(a|s) P(s'|s, a, \tilde{\mu}^\star_t)$$

Theorem 3.

An ERMFNE is an ε-MFNE in the sense that

$$J(\tilde{\mu}^\star, \tilde{\pi}^\star) \geq J(\tilde{\mu}^\star, \pi) - \varepsilon$$

for any valid π, where

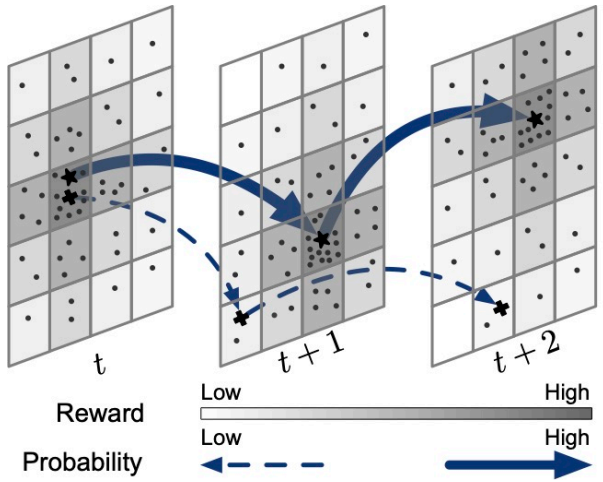
$$\varepsilon < (\beta + |A|) \cdot T \log |A|$$

Probabilistic IRL for MFGs

Theorem 4.

Under the ERMFNE with $\beta = 1$, a state-action trajectory $\tau = (s_0, a_0, \ldots, s_T, a_T)$ can be characterised with an energy-based model:

$$\Pr(\tau) \propto \mu_0(s_0) \cdot \prod_{t=0}^{T} P(s_{t+1}|s_t, a_t, \mu_t^\star) \cdot \exp\left(\sum_{t=0}^{T} r(s_t, a_t, \mu_t^\star)\right)$$



Ziebart, B. D., Bagnell, J. A., & Dey, A. K. (2010, January). Modeling interaction via the principle of maximum causal entropy. In ICML.

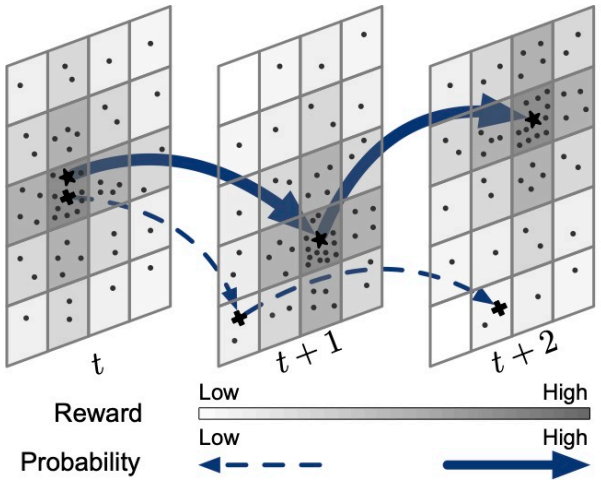
Haarnoja, T., Tang, H., Abbeel, P., & Levine, S. (2017, July). Reinforcement learning with deep energy-based policies. In ICML (pp. 1352-1361). PMLR.

Probabilistic IRL for MFGs

Theorem 4.

Under the ERMFNE with $\beta = 1$, a state-action trajectory $\tau = (s_0, a_0, \ldots, s_T, a_T)$ can be characterised with an energy-based model:

$$\Pr(\tau) \propto \mu_0(s_0) \cdot \prod_{t=0}^{T} P(s_{t+1}|s_t, a_t, \mu_t^\star) \cdot \exp\left(\sum_{t=0}^{T} r(s_t, a_t, \mu_t^\star)\right)$$



Ziebart, B. D., Bagnell, J. A., & Dey, A. K. (2010, January). Modeling interaction via the principle of maximum causal entropy. In ICML.

Haarnoja, T., Tang, H., Abbeel, P., & Levine, S. (2017, July). Reinforcement learning with deep energy-based policies. In ICML (pp. 1352-1361). PMLR.

Tune a reward function by maximum likelihood estimation:

$$\max_{\omega} L(\omega) = \mathbb{E}_{\tau\sim(\mu^E,\pi^E)}\left[\sum_{t=0}^{T} r_\omega(s, a, \mu_t^\omega) + \sum_{t=0}^{T} \log P(s_{t+1}|s_t, a_t, \mu_t^\omega)\right] - \log Z_\omega$$

$$Z_\omega = \sum_{\tau\sim(\mu^E,\pi^E)} \mu_0(s_0) \cdot \prod_{t=0}^{T} P(s_{t+1}|s_t, a_t, \mu_t^\star) \cdot \exp\left(\sum_{t=0}^{T} r(s_t, a_t, \mu_t^\star)\right)$$

A New Form of Maximum Likelihood Estimation

$$\max_{\omega} L(\omega) = \mathbb{E}_{\tau \sim (\mu^E, \pi^E)} \left[\sum_{t=0}^{T} r_\omega(s, a, \mu_t^\omega) + \sum_{t=0}^{T} \log P(s_{t+1} | s_t, a_t, \mu_t^\omega) \right] - \log Z_\omega$$

- **Challenge:** Reward and mean field are coupled through the parameter ω. We cannot directly optimise the likelihood objective.

A New Form of Maximum Likelihood Estimation

$$\max_{\omega} L(\omega) = \mathbb{E}_{\tau\sim(\mu^E,\pi^E)}\left[\sum_{t=0}^{T} r_\omega(s,a,\mu_t^\omega) + \sum_{t=0}^{T} \log P(s_{t+1}|s_t,a_t,\mu_t^\omega)\right] - \log Z_\omega$$

- **Challenge:** Reward and mean field are coupled through the parameter ω. We cannot directly optimise the likelihood objective.
- **Solution:** Decouple reward and mean field: Replace μ_t^ω with the empirically optimal estimation:

$$\hat{\mu}_t^E \triangleq \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_{\{s_{j,t}=s\}}, \quad D = \{\tau_j\}_{j=1}^{M}$$

- **Empirical likelihood objective:**

$$\max_{\omega} \hat{L}(\omega;\hat{\mu}^E) \triangleq \mathbb{E}_{\tau\sim D}\left[\sum_{t=0}^{T} r_\omega(s,a,\hat{\mu}^E)\right] - \log \hat{Z}_\omega$$

Asymptotic Optimality

- Original likelihood objective

$$\max_{\omega} L(\omega) = \mathbb{E}_{\tau\sim(\mu^E,\pi^E)}\left[\sum_{t=0}^{T} r_\omega(s,a,\mu_t^\omega) + \sum_{t=0}^{T} \log P(s_{t+1}|s_t,a_t,\mu_t^\omega)\right] - \log Z_\omega$$

- Empirical likelihood objective

$$\max_{\omega} \hat{L}(\omega;\hat{\mu}^E) \triangleq \mathbb{E}_{\tau\sim D}\left[\sum_{t=0}^{T} r_\omega(s,a,\hat{\mu}^E)\right] - \log \hat{Z}_\omega$$

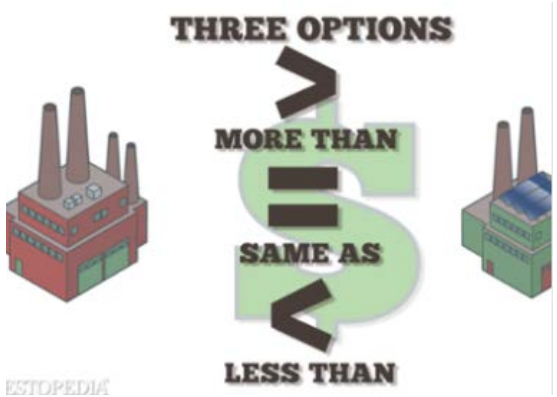
$\hat{\mu}_t^E \to \mu_t^E$ as $M \to \infty$ due to law of large numbers

Theorem 5.

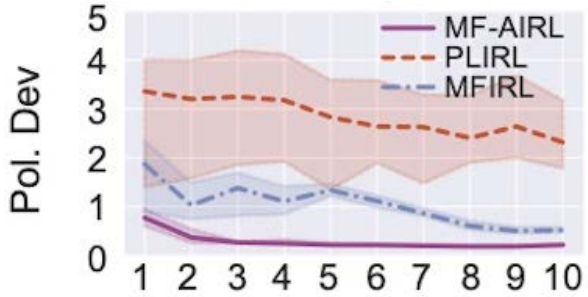
With probability 1 as the number of demonstrations M tends to infinity, there exists an optimal solution of the empirical likelihood objective which is also a maximiser of the original likelihood objective.

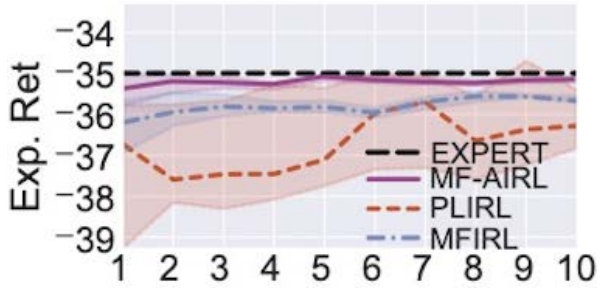
Case Study: Pricing Strategy in Large-scale Market

- **Goal:** Recover the underlying factors that affect the price of companies sharing a common market



- **Results:**





Future Work

1. IRL for MFGs with heterogeneous agents.
2. IRL for MFGs with networked structure.