Mean field game model

- Agent set: \( N := \{1, \ldots, n\} \) agents;
- State and action spaces for each agent: \( X, A \) (finite and identical for all agents);
- Dynamic state evolution for each agent \( i \in N \):
  \[ P[X_{t+1} = x' | X_t, A_t] = P[X_{t+1} = x' | X_t, A_t] \];
- Empirical mean field (or population average): \( L_t \in \Delta(X) \), given by:
  \[ L_t(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[X_i = x], \quad \forall x \in X. \]
- Per-step payoff to agent \( i: u(X_i; A_t, L_t) \)

Stationary mean field equilibrium (SMFE)

SMFE is a pair of a belief \( \pi \in \Delta(X) \) and a policy \( \mu: \mathcal{X} \to \Delta(A) \), which satisfies the following two properties:

1. Sequential Rationality: For any other policy \( \hat{\mu}: \mathcal{X} \to \Delta(A) \),
   \[ V_{\hat{\mu}}(x) \geq V_{\pi}(x), \quad \forall x \in X. \]
2. Consistency: The belief \( \pi \) is stationary under policy \( \mu \), i.e., \( \pi = \text{StatDist}(\pi, \mu) \).

Policy gradient based algorithm: Main proposition

- If \( \delta_{k+1} = [\delta_t + \alpha_k G_{k0}]_{t=0}^{\infty} \) converges to a limit \( \delta^* \) along any sample path, then \((\delta^*, \pi^*)\) is a \( \nabla \)-SMFE.
- Likelihood ratio based gradient estimate:
  \[
  G_0 = \mathbb{E}_k \left[ \nabla V_{\pi}(X_t) \right], \quad \text{where} \quad \nabla V_{\pi}(x) = \mathbb{E}_{\pi \sim \pi_0} \left[ \sum_{i=0}^{\infty} \lambda_i^t \nabla V_{\pi}(X_t) \big| X_0 = x \right].
  \]
- Simultaneous perturbation based gradient estimate:
  \[
  G_0 = \eta [I_{n+2p+1} - I_{n+2p+1}]/2\beta.
  \]

Policy improvement

**input**: \( \theta_0 \): Initial parameter; K: \# iterations; \( L_0 \): initial mean field distribution; \( B \): burn-in period; \( n_p \): \# particles

**for iterations** \( k = 1 : K \) do
  \[
  \sigma_k = \text{StatDist}(\theta_k, \mu_0, B, n_p).
  \]
  \[
  G_k = \text{PolicyGradient}(\theta_k, L_0, \sigma_k).
  \]
  \[
  \theta_{k+1} \leftarrow (\theta_k + n_p G_k)_{t=0}^{\infty}
  \]
  return \( \theta_{K+1} \)

Key assumptions

1. An agent uses only its current state to pick actions: \( \mu_i: \mathcal{X} \to \Delta(A) \) and \( A_t = \mu_i(X_t) \).
2. \( \mu_i \) does not depend on time.
3. All agents play identical policies. Thus \( \mu = [\mu_1, \ldots, \mu_n] \).
4. Each agent assumes that the population average is stationary. Thus agent i’s assessment of its payoff is:
   \[ V_{\pi}(x) = \mathbb{E}_{\pi \sim \pi_0} [\sum_{i=0}^{\infty} y^i u(X_i; A_i, \pi) | X_0 = x]. \]
5. We consider parametrized policies \( \mu_\theta \), where \( \theta \in \Theta \) (a closed, convex space).

Solution concept

- Mean field equilibrium and its refinements are standard solution concepts in mean field games.
- **Our contribution**
  - Definition of an equilibrium for stationary mean field games based on bounded rationality.
  - This equilibrium is a generalization of Nash equilibrium and mean field equilibrium.
  - Development of a policy gradient based algorithm to predict this equilibrium.

Conclusions

- In this work an RL algorithm is used for planning. This implies that the iterates in our algorithm are not representative of the learning dynamics of individual agents.
- For this to be an RL algorithm, each agent would have to make an assumption on all other agents’ behaviour in the learning phase.
- This coordination in learning is not easily justified in a competitive game with strategic agents, where the agents can try and influence their opponents during learning.
- Although we presented only policy based algorithms, bounded rationality can also be modelled using a critic only variant with function approximation.