A. Detailed mean field reinforcement learning algorithms

We published the code at `https://github.com/mlii/mfrl`.

Algorithm 1 Mean Field Q-learning (MF-Q)

```
Initialise $Q_{\phi^j}$, $Q_{\phi^j_-}$, and $\bar{a}^j$ for all $j \in \{1, \ldots, N\}$
while training not finished do
  for $m = 1, ..., M$ do
    For each agent $j$, sample action $a^j$ from $Q_{\phi^j}$ by Eq. (12), with the current mean action $\bar{a}^j$ and the exploration rate $\beta$
    For each agent $j$, compute the new mean action $\bar{a}^j$ by Eq. (11)
  Take the joint action $\boldsymbol{a} = [a^1, \ldots, a^N]$ and observe the reward $\boldsymbol{r} = [r^1, \ldots, r^N]$ and the next state $s'$
  Store $\langle s, \boldsymbol{a}, \boldsymbol{r}, s', \bar{\boldsymbol{a}} \rangle$ in replay buffer $\mathcal{D}$, where $\bar{\boldsymbol{a}} = [\bar{a}^1, \ldots, \bar{a}^N]$
  for $j = 1$ to $N$ do
    Sample a minibatch of $K$ experiences $\langle s, \boldsymbol{a}, \boldsymbol{r}, s', \bar{\boldsymbol{a}} \rangle$ from $\mathcal{D}$
    Sample action $a^j_-$ from $Q_{\phi^j_-}$ with $\bar{a}^j_- \leftarrow \bar{a}^j$
    Set $y^j = r^j + \gamma \, v^{\mathrm{MF}}_{\phi^j_-}(s')$ by Eq. (10)
    Update the $Q$-network by minimizing the loss $\mathcal{L}(\phi^j) = \frac{1}{K} \sum \big(y^j - Q_{\phi^j}(s^j, a^j, \bar{a}^j)\big)^2$
  Update the parameters of the target network for each agent $j$ with learning rate $\tau$:

      $$\phi^j_- \leftarrow \tau \phi^j + (1 - \tau) \phi^j_-$$
```

Algorithm 2 Mean Field Actor-Critic (MF-AC)

```
Initialize $Q_{\phi^j}$, $Q_{\phi^j_-}$, $\pi_{\theta^j}$, $\pi_{\theta^j_-}$, and $\bar{a}^j$ for all $j \in \{1, \ldots, N\}$
while training not finished do
  For each agent $j$, sample action $a^j = \pi_{\theta^j}(s)$; compute the new mean action $\bar{\boldsymbol{a}} = [\bar{a}^1, \ldots, \bar{a}^N]$
  Take the joint action $\boldsymbol{a} = [a^1, \ldots, a^N]$ and observe the reward $\boldsymbol{r} = [r^1, \ldots, r^N]$ and the next state $s'$
  Store $\langle s, \boldsymbol{a}, \boldsymbol{r}, s', \bar{\boldsymbol{a}} \rangle$ in replay buffer $\mathcal{D}$
  for $j = 1$ to $N$ do
    Sample a minibatch of $K$ experiences $\langle s, \boldsymbol{a}, \boldsymbol{r}, s', \bar{\boldsymbol{a}} \rangle$ from $\mathcal{D}$
    Set $y^j = r^j + \gamma \, v^{\mathrm{MF}}_{\phi^j_-}(s')$ by Eq. (10)
    Update the critic by minimizing the loss $\mathcal{L}(\phi^j) = \frac{1}{K} \sum \big(y^j - Q_{\phi^j}(s, a^j, \bar{a}^j)\big)^2$
    Update the actor using the sampled policy gradient:

      $$\nabla_{\theta^j} \mathcal{J}(\theta^j) \approx \frac{1}{K} \sum \nabla_{\theta^j} \log \pi_{\theta^j}(s') Q_{\phi^j_-}(s', a^j_-, \bar{a}^j_-) \Big|_{a^j_- = \pi_{\theta^j_-}(s')}$$

  Update the parameters of the target networks for each agent $j$ with learning rates $\tau_\phi$ and $\tau_\theta$:

      $$\phi^j_- \leftarrow \tau_\phi \phi^j + (1 - \tau_\phi) \phi^j_-$$
      $$\theta^j_- \leftarrow \tau_\theta \theta^j + (1 - \tau_\theta) \theta^j_-$$
```

B. Proof of the bound for the remainder term in Eq. 7

Recall Eq. (8) that we approximate the action a^k taken by the neighboring agent k with the mean action $\bar{a}$ calculated from the neighborhood $\mathcal{N}(j)$. The state s and the action a^j of the central agent j can be considered as fixed parameters; the indices j, k of agents are essentially irrelevant to the derivation. With those omitted for simplicity, We rewrite the expression of the pairwise Q-function as $Q(a) \triangleq Q^j(s, a^j, a^k)$.

Suppose that Q is M-smooth, where its gradient ∇Q is Lipschitz-continuous with constant M such that for all $a, \bar{a}$

$$\|\nabla Q(a) - \nabla Q(\bar{a})\|_2 \leq M\|a - \bar{a}\|_2, \tag{20}$$

where $\|\cdot\|_2$ indicates the ℓ_2-norm.

With the Lagrange's mean value theorem, we have

$$\nabla Q(a) - \nabla Q(\bar{a}) = \nabla Q(\bar{a} + 1 \cdot (a - \bar{a})) - \nabla Q(\bar{a}) = \nabla^2 Q(\bar{a} + \epsilon \cdot (a - \bar{a})) \cdot (a - \bar{a}), \quad \text{where } \epsilon \in [0, 1]. \tag{21}$$

Take the ℓ_2-norm on the both sides of the above equation, it follows from the smoothness condition that

$$\|\nabla Q(a) - \nabla Q(\bar{a})\|_2 = \|\nabla^2 Q(\bar{a} + \tau \cdot (a - \bar{a})) \cdot (a - \bar{a})\|_2 \leq M\|a - \bar{a}\|_2. \tag{22}$$

Define $\delta a \triangleq a - \bar{a}$ and the normalized vector $\delta\hat{a} \triangleq {}^{a-\bar{a}}/_{\|a-\bar{a}\|_2}$ with $\|\delta\hat{a}\|_2 = 1$, it follows from the above inequality

$$\|\nabla^2 Q(a + \tau \cdot \delta a) \cdot \delta\hat{a}\|_2 \leq M. \tag{23}$$

By arbitrary choice of (the unnormalized vector) δa such that the magnitude $\|\delta a\|_2 \to 0$, it follows from above that

$$\|\nabla^2 Q(a) \cdot \delta\hat{a}\|_2 \leq M. \tag{24}$$

By aligning (the normalized vector) $\delta\hat{a}$ in the direction of the eigenvectors of the Hessian matrix $\nabla^2 Q$, we can obtain for any eigenvalue λ of $\nabla^2 Q$ that

$$\|\nabla^2 Q(a) \cdot \delta\hat{a}\|_2 = \|\lambda \cdot \delta\hat{a}\|_2 = |\lambda| \cdot \|\delta\hat{a}\|_2 \leq M, \tag{25}$$

which indicates that all eigenvalues of $\nabla^2 Q$ can be bounded in the symmetric interval $[-M, M]$.

As the Hessian matrix $\nabla^2 Q$ is real symmetric and hence diagonalizable, there exist an orthogonal matrix U such that $U^\top[\nabla^2 Q]U = \Lambda \triangleq \mathrm{diag}[\lambda_1, \ldots, \lambda_D]$. It then follows that

$$\delta a \cdot \nabla^2 Q \cdot \delta a = [U\delta a]^\top \Lambda [U\delta a] = \sum_{i=1}^{D} \lambda_i [U\delta a]_i^2, \quad \text{with } -M\|U\delta a\|_2 \leq \sum_{i=1}^{D} \lambda_i [U\delta a]_i^2 \leq M\|U\delta a\|_2 \tag{26}$$

Recall the definition $\delta a = a - \bar{a}$ in Eq. (6), where a is the one-hot encoding for D actions, and $\bar{a}$ is a D-dimensional multinomial distribution. It can be shown that

$$\|U\delta a\|_2 = \|\delta a\|_2 = (a - \bar{a})^\top (a - \bar{a}) = a^\top a + \bar{a}^\top \bar{a} - \bar{a}^\top a - a^\top \bar{a} = 2(1 - \bar{a}_i) \leq 2, \tag{27}$$

where i represents the specific action a has represented such that $a_{i'} = 0$ for $i' \neq i$.

With all elements assembled, we have proved that each single remainder term $R^j_{s,a^j}(a^k)$ in Eq. (8) is bounded in $[-2M, 2M]$.

C. Experiment details

C.1. Gaussian Squeeze

IL, FMQ, Rec-FMQ and MF-Q all use a three-layer MLP to approximate Q-value. All agents share the same Q-network for each experiment. The shared Q-network takes an agent embedding as input and computes Q-value for each candidate action. For MF-Q, we also feed in the action approximation $\bar{a}$. We use the Adam optimizer with a learning rate of 0.00001 and ϵ-greedy exploration unless otherwise specified. For FMQ, we set the exponential decay rate $s = 0.006$, start temperature *max_temp=1000* and FMQ heuristic $c = 5$. For Rec-FMQ, we set the frequency learning rate $\alpha_f = 0.01$.

MAAC and MF-AC use the Adam optimizer with a learning rate of 0.001 and 0.0001 for Critics and Actors respectively, and $\tau = 0.01$ for updating the target networks. We share the Critic among all agents in each experiment and feed in an agent embedding as extra input. Actors are kept separate. The discounted factor γ is set to be 0.95 and the mini-batch size is set to be 200. The size of the replay buffer is 10^6 and we update the network parameters after every 500 samples added to the replay buffer.

For all models, we use the performance of the joint-policy learned up to that point if learning and exploration were turned off (*i.e.*, take the greedy action w.r.t. the learned policy) to compare our method with the above baseline models.

C.2. Ising Model

An Ising model is defined as a stateless system with N homogeneous sites on a finite square lattice. Each site determines their individual spin a^j to interact with each other and aims to minimize the system energy for a more stable environment. The system energy is defined as

$$E(a,h) = -\sum_j \Big(h^j a^j + \frac{\lambda}{2} \sum_{k\in\mathcal{N}(j)} a^j a^k\Big) \tag{28}$$

where $\mathcal{N}(j)$ is the set of nearest neighbors of site j, $h^j \in \mathbb{R}$ is the external field affecting site j, and $\lambda \in \mathbb{R}$ is an interaction term determines how much the sites tend to align in the same direction. The system is said to reach an equilibrium point when the system energy is minimized, with the probability

$$P(a) = \frac{\exp\left(-E(a,h)/\tau\right)}{\sum_a \exp(-E(a,h)/\tau)}, \tag{29}$$

where τ is the system temperature. When the temperature rises beyond a certain point (the Curie temperature), the system can no longer keep a stable form and a phase transition happens. As the ground-truth is known, we would be able to evaluate the correctness of the Q-function learning when there is a large body of agents interacted.

The mean field theory provides an approximate solution to $\langle a^j\rangle = \sum_a a^j P(a)$ through a set of self-consistent mean field equations

$$\langle a^j\rangle = \frac{\exp\Big(-[h^j a^j + \lambda \sum_{k\in\mathcal{N}(j)} \langle a^k\rangle]/\tau\Big)}{1+\exp\Big(-[h^j a^j + \lambda \sum_{k\in\mathcal{N}(j)} \langle a^k\rangle]/\tau\Big)}. \tag{30}$$

which can be solved iteratively by

$$\langle a^j\rangle^{(t+1)} = \frac{\exp\Big(-[h^j a^j + \lambda \sum_{k\in\mathcal{N}(j)} \langle a^k\rangle^{(t)}]/\tau\Big)}{1+\exp\Big(-[h^j a^j + \lambda \sum_{k\in\mathcal{N}(j)} \langle a^k\rangle^{(t)}]/\tau\Big)}, \tag{31}$$

where t represents the number of iterations.

To learn an optimal joint policy $\boldsymbol{\pi}^*$ for Ising model, we use the stateless Q-learning with mean field approximation (MF-Q), defined as

$$Q^j(a^j,\bar{a}^j) \leftarrow Q^j(a^j,\bar{a}^j) + \alpha[r^j - Q^j(a^j,\bar{a}^j)], \tag{32}$$

Algorithm 3 MCMC in Ising Model

```
initialize spin state $\boldsymbol{a} \in \{-1, 1\}^N$ for $N$ sites
while training not finished do
  randomly choose site $j \in \mathcal{N}(j)$
  flip the spin state for site $j$: $a^j_- \leftarrow -a^j$
  compute neighbor energy $E(a, h) = -\sum_j (h^j a^j + \frac{\lambda}{2} \sum_{k \in \mathcal{N}(j)} a^j a^k)$ for $a^j$ and $a^j_-$
  randomly choose $\epsilon \sim U(0, 1)$
  if $\exp((E(a^j, h) - E(a^j_-, h))/\tau) > \epsilon$ then
    $a^j \leftarrow a^j_-$
```

where the mean $\bar{a}^j$ is given as the mean $\langle a^j \rangle$ from the last time step, and the individual reward is

$$r^j = h^j a^j + \frac{\lambda}{2} \sum_{k \in \mathcal{N}(j)} a^j a^k. \tag{33}$$

To balance the trade-off between exploration and exploitation under low temperature settings, we use a policy with Boltzmann exploration and a decayed exploring temperature. The temperature for Boltzmann exploration of MF-Q is multiplied by a decay factor exponentially through out the training process.

Without lost of generality, we assume $\lambda > 0$, thus neighboring sites with the same action result in lower energy (observe higher reward) and are more stable. Each site should also align with the sign of external field h^j to reduce the system energy. For simplification, we eliminate the effect of external fields and assume the model to be discrete, *i.e.*, $\forall j \in N, h^j = 0, a^j \in \{-1, 1\}$.

We simulate the Ising model using Metropolis Monte Carlo methods (MCMC). After initialization, we randomly change a site's spin state and calculate the energy change, select a random number between 0 and 1, and accept the state change only if the number is less than $e^{\frac{(E^j - E^j_-)}{\tau}}$. This is called the Metropolis technique, which saves computation time by selecting the more probable spin states.

C.3. Battle Game

IL and MF-Q have almost the same hyper-parameters settings. The learning rate is $\alpha = 10^{-4}$, and with a dynamic exploration rate linearly decays from $\gamma = 1.0$ to $\gamma = 0.05$ during the 2000 rounds training. The discounted factor γ is set to be 0.95 and the mini-batch size is 128. The size of replay buffer is 5×10^5.

AC and MF-AC also have almost the same hyper-parameters settings. The learning rate is $\alpha = 10^{-4}$, the temperature of soft-max layer in *actor* is $\tau = 0.1$. And the coefficient of entropy in the total loss is 0.08, the coefficient of value in the total loss is 0.1.

D. Further details towards the theoretical guarantee of MF-Q

Proposition 1. *Let the metric space be $\mathbb{R}^N$ and the metric be $d(\boldsymbol{a},\boldsymbol{b}) = \sum_j |a^j - b^j|$, for $\boldsymbol{a} = [a^j]_1^N, \boldsymbol{b} = [b^j]_1^N$. If the Q-function is K-Lipschitz continuous* w.r.t. *a^j, then the operator $\mathscr{B}(a^j) \triangleq \pi^j(a^j|s,\bar{a}^j)$ in Eq.* (12) *forms a contraction mapping under sufficiently low temperature β.*

Proof. Following the contraction mapping theorem (Kreyszig, 1978), in order to be a contraction, the operator has to satisfy:

$$d(\mathscr{B}(\boldsymbol{a}),\mathscr{B}(\boldsymbol{b})) \leq \alpha d(\boldsymbol{a},\boldsymbol{b}), \quad \forall \boldsymbol{a},\boldsymbol{b}$$

where $0 \leq \alpha < 1$ and $\mathscr{B}(\boldsymbol{a}) \triangleq [\mathscr{B}(a^1),\ldots,\mathscr{B}(a^N)]$.

Here we start from binomial case and then adapt to the multinomial case in general. We first rewrite $\mathscr{B}(a^j)$ as

$$\begin{aligned}\mathscr{B}(a^j) = \pi^j(a^j|s,\bar{a}^j) =& \frac{\exp\big(-\beta Q_t^j(s,a^j,\bar{a}^j)\big)}{\exp\big(-\beta Q_t^j(s,a^j,\bar{a}^j)\big) + \exp\big(-\beta Q_t^j(s,\neg a^j,\bar{a}^j)\big)} \\ =& \frac{1}{1+\exp\left(-\beta \cdot \varDelta Q(s,a^j,\bar{a})\right)}\,,\end{aligned} \tag{34}$$

where $\varDelta Q(s,a^j,\bar{a}) = Q(s,a^{\neg j},\bar{a}) - Q(s,a^j,\bar{a})$.

Then we have

$$\begin{aligned}|\mathscr{B}(a^j) - \mathscr{B}(b^j)| =& \left|\frac{1}{1+e^{-\beta\cdot\varDelta Q(s,a^j,\bar{a})}} - \frac{1}{1+e^{-\beta\cdot\varDelta Q(s,b^j,\bar{a})}}\right| \\ =& \left|\frac{\beta e^{-\beta\varDelta Q_0}}{(1+e^{-\beta\varDelta Q_0})^2}\right| \left|\varDelta Q(s,a^j,\bar{a}) - \varDelta Q(s,b^j,\bar{a})\right| \\ \leq& \frac{1}{4T}\cdot\left|Q(s,a^{\neg j},\bar{a}) - Q(s,b^{\neg j},\bar{a}) + Q(s,b^j,\bar{a}) - Q(s,a^j,\bar{a})\right| \\ \leq& \frac{1}{4T}\cdot\left(K\cdot\left|1-a^j-(1-b^j)\right| + K\cdot\left|a^j-b^j\right|\right) \\ \leq& \frac{1}{4T}\cdot 2K\cdot\sum_j\left|a^j-b^j\right|\ .\end{aligned} \tag{35}$$

In the second equation, we apply the mean value theorem in calculus: $\exists x_0 \in [x_1,x_2]$, *s.t.*, $f(x_1) - f(x_2) = f'(x_0)(x_1-x_2)$. In the third equation we use the maximum value for $e^{-\beta\varDelta Q_0}/(1+e^{-\beta\varDelta Q_0})^2 = 1/4$ when $Q_0 = 0$. In the last equation we apply the Lipschitz constraint in the assumption where constant $K \geq 0$. Finally, we have:

$$\begin{aligned}d(\mathscr{B}(\boldsymbol{a}),\mathscr{B}(\boldsymbol{b})) \leq& \frac{1}{4T}\cdot 2K\cdot\sum_j\left|a^j-b^j\right| \\ =& \frac{K}{2T}d(\boldsymbol{a},\boldsymbol{b})\end{aligned} \tag{36}$$

In order for the contraction to hold, $T > \frac{K}{2}$. In other words, when the action space is binary for each agent, and the temperature is sufficiently large, the mean field procedure converges.

This proposition can be easily extended to multinomial case by replacing binary variable a^j by a multi-dimensional binary indicator vector $\boldsymbol{a}^j$, on each dimension, the rest of the derivations would remain essentially the same. □

D.1. Discussion on Rationality

In line with (Bowling & Veloso, 2001; 2002), we argue that to better evaluate a multi-agent learning algorithm, on top of the convergence guarantee, discussion on property of Rationality is also needed.

Property 1. *(also see (Bowling & Veloso, 2001; 2002)) In an N-agent stochastic game defined in this paper, given all agents converge to stationary policies, if the learning algorithm converges to a policy that is a **best response** to the other agents' policies, then the algorithm is Rationale.*

Our mean field Q-learning is rational in that Eq. (5) converts many agents interactions into two-body interactions between a single agent and the distribution of other agents actions. When all agents follow stationary policies, their policy distribution would be stationary too. As such the two-body stochastic game becomes an MDP, and the agent would choose a policy (based on Assumption 2) which is the best response to the distribution of other stationary policies. As agents are symmetric in our case, they all show the best response to the distributions, and are therefore rational.