Approximately Solving Mean Field Games via Entropy-Regularized Deep Reinforcement Learning Supplementary Materials

A Experimental Details

A.1 Algorithms

Algorithm 1 Exact fixed point iteration

Initialize $\mu^0 = \Psi(q)$ as the mean field induced by the uniformly random policy q.
for $k = 0, 1, \cdots$ **do**
  Compute the Q-function $Q^*(\mu^k, t, s, a)$ for fixed μ^k.
  Choose $\pi^k \in \Pi$ such that $\pi^k_t(a \mid s) \implies a \in \arg\max_{a \in \mathcal{A}} Q^k(\mu^k, t, s, a)$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ by putting all probability mass on the first optimal action, or evenly on all optimal actions.
  Optionally: Overwrite $\pi^k \leftarrow \frac{1}{k+1}\pi^k + \frac{k}{k+1}\pi^{k-1}$. (FP averaged policy)
  Compute the mean field $\mu^{k+1} = \Psi(\pi^k)$ induced by π^k.
  Optionally: Overwrite $\mu^{k+1} \leftarrow \frac{1}{k+1}\mu^{k+1} + \frac{k}{k+1}\mu^k$. (FP averaged mean field)
end for

Algorithm 2 Boltzmann / RelEnt iteration

Input: Temperature $\eta > 0$, prior policy $q \in \Pi$.
Initialize $\mu^0 = \Psi(q)$ as the mean field induced by q.
for $k = 0, 1, \cdots$ **do**
  Compute the Q-function (Boltzmann) or soft Q-function (RelEnt) $Q(\mu^k, t, s, a)$ for fixed μ^k.
  Define π^k by $\pi^k_t(a \mid s) = \frac{q_t(a|s) \exp\left(\frac{Q(\mu^k, t, s, a)}{\eta}\right)}{\sum_{a' \in \mathcal{A}} q_t(a'|s) \exp\left(\frac{Q(\mu^k, t, s, a')}{\eta}\right)}$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.
  Optionally: Overwrite $\pi^k \leftarrow \frac{1}{k+1}\pi^k + \frac{k}{k+1}\pi^{k-1}$. (FP averaged policy)
  Compute the mean field $\mu^{k+1} = \Psi(\pi^k)$ induced by π^k.
  Optionally: Overwrite $\mu^{k+1} \leftarrow \frac{1}{k+1}\mu^{k+1} + \frac{k}{k+1}\mu^k$. (FP averaged mean field)
end for

Algorithm 3 Boltzmann DQN iteration

Input: Temperature $\eta > 0$, prior policy $q \in \Pi$.
Input: Simulation parameters, DQN hyperparameters.
Initialize $\mu^0 \approx \Psi(q)$ as the mean field induced by q using Algorithm 5.
for $k = 0, 1, \cdots$ **do**
  Approximate the Q-function $Q^*(\mu^k, t, s, a)$ using Algorithm 4 on the MDP induced by μ^k.
  Define π^k by $\pi^k_t(a \mid s) = \frac{q_t(a|s) \exp\left(\frac{Q^*(\mu^k, t, s, a)}{\eta}\right)}{\sum_{a' \in \mathcal{A}} q_t(a'|s) \exp\left(\frac{Q^*(\mu^k, t, s, a')}{\eta}\right)}$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.
  Approximately simulate mean field $\mu^{k+1} \approx \Psi(\pi^k)$ induced by π^k using Algorithm 5.
end for

Algorithm 4 DQN

Input: Number of epochs L, mini-batch size N, target update frequency M, replay buffer size D.
Input: Probability of random action ϵ, Discount factor γ, ADAM and gradient clipping parameters.
Initialize network Q_θ, target network $Q_{\theta'} \leftarrow Q_\theta$ and replay buffer $\mathcal{D}$ of size D.
for L epochs **do**
  for $t = 1, \dots, \mathcal{T}$ **do**
    One environment step
      Let new action $a_t \leftarrow \arg\max_{a \in \mathcal{A}} Q_\theta(t, s, a)$, or with probability ϵ sample uniformly random instead.
      Sample new state $s_{t+1} \sim p(\cdot \mid s_t, a_t)$.
      Add transition tuple $(s_t, a_t, r(s_t, a_t), s_{t+1})$ to replay buffer $\mathcal{D}$.
    One mini-batch descent step
      Sample from the replay buffer: $\{(s_t^i, a_t^i, r_t^i, s_{t+1}^i)\}_{i=1,\dots,N} \sim \mathcal{D}$.
      Compute loss $J_Q = \sum_{i=1}^N \left(r_t^i + \gamma \max_{a' \in \mathcal{A}} Q(t+1, s_{t+1}^i, a') - Q(t, s_t^i, a_t^i)\right)^2$.
      Update θ according to $\nabla_\theta J_Q$ using ADAM with gradient norm clipping.
    if number of steps $\bmod M = 0$ **then**
      Update target network $\theta' \leftarrow \theta$.
    end if
  end for
end for

Algorithm 5 Stochastic mean field simulation

Input: Number of mean fields K, number of particles M, policy π.
for $k = 1, \dots, K$ **do**
  Initialize particles $x_m^0 \sim \mu_0$ for all $m = 1, \dots, M$.
  for $t \in \mathcal{T}$ **do**
    Define empirical measure $\mathbb{G}_t^k \leftarrow \sum_{m=1}^M \delta_{x_m^t}$.
    for $m = 1, \dots, M$ **do**
      Sample action $a \sim \pi_t(\cdot \mid x_m^t)$.
      Sample new particle state $x_m^{t+1} \sim p(\cdot \mid x_m^t, a, \mathbb{G}_t^k)$.
    end for
  end for
end for
return average empirical mean field $(\frac{1}{K} \sum_{k=1}^K \mathbb{G}_t^k)_{t \in \mathcal{T}}$

A.2 Implementation details

For all the DQN experiments, we use the configurations given in Table 1 and hyperparameters given in Table 2. Note that we add epsilon scheduling and a discount factor to DQN for stability reasons, i.e. the loss term has an additional factor smaller than one before the maximum operation, cf. Mnih et al. (2013). For the action-value network, we use a fully connected dueling architecture (Wang et al. (2016)) with one shared hidden layer of 256 neurons, and one separate hidden layer of 256 neurons for value and advantage stream each. As the activation function, we use ReLU. Further, we use gradient norm clipping and the ADAM optimizer. To allow for time-dependent policies, we append the current time to the observations.

We transform all discrete-valued observations except time to corresponding one-hot vectors, except in the intractably large Taxi environment where we simply observe one value in $\{0, 1\}$ for each tile's passenger status. For evaluation of exploitability, we compare the values of the optimal policy and the evaluated policy in the MDP induced by the mean field generated by the evaluated policy. In intractable cases, we use DQN to approximately obtain the optimal policy. In this case, we obtain the values by averaging over many episodes in the MDP induced by the mean field generated by the evaluated policy via Algorithm 5.

A.3 Problems

Summarizing properties of the considered problems are given in Table 3.

Algorithm 6 Prior descent

Input: Number of outer iterations I.
Input: Initial prior policy $q \in \Pi$.
for outer iteration $i = 1, \dots, I$ **do**
  Find η heuristically or minimally such that Algorithm 2 with temperature η and prior q converges.
  if no such η exists **then**
    return q
  end if
  $q \leftarrow$ solution of Algorithm 2 with temperature η and prior q.
end for

Table 1: Boltzmann DQN Iteration Parameters

Parameter	RPS	SIS	Taxi
Fixed point iteration count	1000	50	15
Number of particles for mean field	1000	1000	200
Number of mean fields	5	5	5
Number of episodes for evaluation	2000	2000	500

LR. Similar to the example mentioned in the main text, we let a large number of agents choose simultaneously between going left (L) or right (R). Afterwards, each agent shall be punished proportional to the number of agents that chose the same action, but more-so for choosing right than left.

More formally, let $\mathcal{S} = \{C, L, R\}$, $\mathcal{A} = \mathcal{S} \setminus \{C\}$, $\mu_0(C) = 1$, $r(s, a, \mu_t) = -\mathbf{1}_{\{L\}}(s) \cdot \mu_t(L) - 2 \cdot \mathbf{1}_{\{R\}}(s) \cdot \mu_t(R)$ and $\mathcal{T} = \{0, 1\}$. Note the difference to the toy example in the main text: right is punished more than left. The transition function allows picking the next state directly, i.e. for all $s, s' \in \mathcal{S}, a \in \mathcal{A}$,

$$\mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a) = \mathbf{1}_{\{s'\}}(a)\,.$$

For this example, we have $K_Q = 1$ since the return Q of the initial state changes linearly with μ_1 and lies between 0 and -2, while the distance between two mean fields is also bounded by 2. Analogously, $K_\Psi = 1$ since $(\Psi(\pi))_1$ similarly changes linearly with π_0, and both can change at most by 2. Thus, we obtain guaranteed convergence via Boltzmann iteration if $\eta > 1$. In numerical evaluations, we see convergence already for $\eta \geq 0.7$.

RPS. This game is inspired by Shapley (1964) and their generalized non-zero-sum version of Rock-Paper-Scissors, for which classical fictitious play would not converge. Each of the agents can choose between rock, paper and scissors, and obtains a reward proportional to double the number of beaten agents minus the number of agents beating the agent. We modify the proportionality factors such that a uniformly random prior policy does not constitute a mean field equilibrium.

Let $\mathcal{S} = \{0, R, P, S\}$, $\mathcal{A} = \mathcal{S} \setminus \{0\}$, $\mu_0(0) = 1$, $\mathcal{T} = \{0, 1\}$, and for any $a \in \mathcal{A}, \mu_t \in \mathcal{P}(\mathcal{S})$,

$$\begin{aligned} r(R, a, \mu_t) &= 2 \cdot \mu_t(S) - 1 \cdot \mu_t(P), \\ r(P, a, \mu_t) &= 4 \cdot \mu_t(R) - 2 \cdot \mu_t(S), \\ r(S, a, \mu_t) &= 6 \cdot \mu_t(P) - 3 \cdot \mu_t(R)\,. \end{aligned}$$

The transition function allows picking the next state directly, i.e. for all $s, s' \in \mathcal{S}, a \in \mathcal{A}$,

$$\mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a) = \mathbf{1}_{\{s'\}}(a)\,.$$

SIS. In this problem, a large number of agents can choose between social distancing (D) or going out (U). If a susceptible (S) agent chooses social distancing, they may not become infected (I). Otherwise, an agent may become infected with a probability proportional to the number of agents being infected. If infected, an agent will recover with a fixed chance every time step. Both social distancing and being infected have an associated cost.

Table 2: DQN Hyperparameters

Hyperparameter	Value
Replay buffer size	10000
ADAM Learning rate	0.0005
Discount factor	0.99
Target update frequency	500
Gradient clipping norm	40
Mini-batch size	128
Epsilon schedule	1 linearly down to 0.02 at 0.8 times maximum steps
Total epochs	1000

Table 3: Problem Properties

Problem	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{A}\rvert$
LR	2	3	2
RPS	2	4	3
SIS	50	2	2
Taxi	100	$\sim 2^{27}$	5

Let $\mathcal{S} = \{S, I\}$, $\mathcal{A} = \{U, D\}$, $\mu_0(I) = 0.6$, $r(s, a, \mu_t) = -\mathbf{1}_{\{I\}}(s) - 0.5 \cdot \mathbf{1}_{\{D\}}(s)$ and $\mathcal{T} = \{0, \dots, 50\}$. We find that similar parameters produce similar results, and set the transition probability mass functions as

$$\begin{aligned}\mathbb{P}(S_{t+1} = S \mid S_t = I) &= 0.3\\ \mathbb{P}(S_{t+1} = I \mid S_t = S, A_t = U) &= 0.9^2 \cdot \mu_t(I)\\ \mathbb{P}(S_{t+1} = I \mid S_t = S, A_t = D) &= 0\,.\end{aligned}$$

Taxi. In this problem, we consider a $K \times L$ grid. The state is described by a tuple (x, y, x', y', p, B) where (x, y) is the agent's position, (x', y') indicates the current desired destination of the passenger or is $(0, 0)$ otherwise, and $p \in \{0, 1\}$ indicates whether a passenger is in the taxi or not. Finally, B is a $K \times L$ matrix indicating whether a new passenger is available for the taxi on the corresponding tile. All taxis start on the same tile and have no passengers in the queue or on the map at the beginning. The problem runs for 100 time steps.

The taxi can choose between five actions W, U, D, L, R, where W (Wait) allows the taxi to pick up / deliver passengers, and U, D, L, R (Up, Down, Left, Right) allows it to move in all four directions. As there are many taxis, there is a chance of a jam on tile s given by $\min(0.7, 10 \cdot \mu_t(s))$, i.e. the taxi will not move with this probability. The taxi also cannot move into walls or back into the starting tile, in which case it will stay on its current tile. With a probability of 0.8, a new passenger spawns on one randomly chosen free tile of each region. On picking up a passenger, the destination is generated by randomly picking any free tile of the same region. Delivering passengers to a destination and picking them up gives a reward of 1 in region 1 and 1.2 in region 2.

For our experiments, we use the following small map, where S denotes the starting tile, 1 denotes a free tile from region 1, 2 denotes a free tile from region 2 and H denotes an impassable wall:

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ H & S & H \\ 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}$$

This produces a similar situation as in LR, where a fraction of taxis should choose each region so the values balance out, while also requiring solution of a problem that is intractable to solve exactly via dynamic programming.

A.4 Further experiments

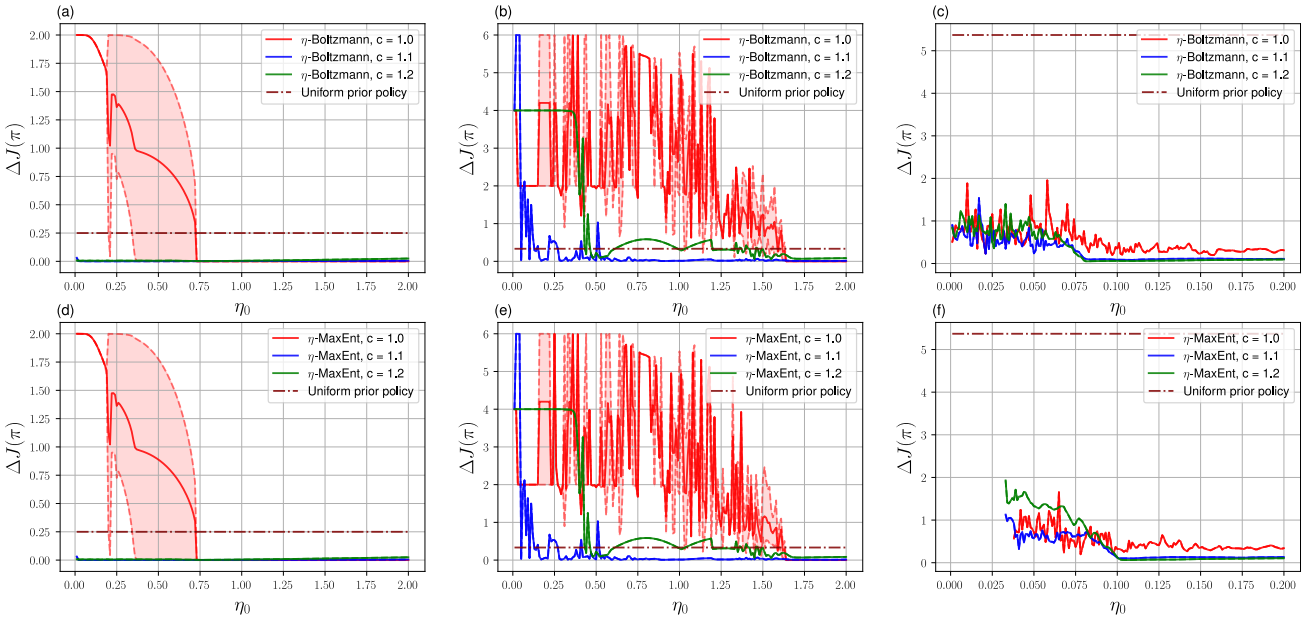

Figure 1: Mean exploitability (straight lines), maximum and minimum (dashed lines) over the final 10 iterations of the last outer iteration. 50 outer iterations and 100 inner iterations each; (a, d) LR; (b, e) RPS; (c, f) SIS. Maximum entropy (MaxEnt) results begin at higher temperatures due to limited floating point accuracy. The exploitability of the initial uniform prior policy is indicated by the dashed horizontal line.

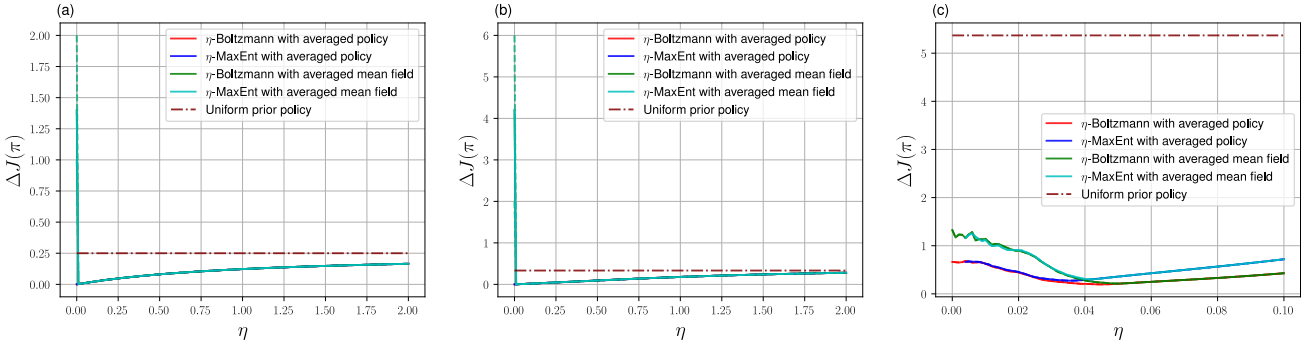

Figure 2: Mean exploitability over the final 10 iterations. Dashed lines represent maximum and minimum over the final 10 iterations. (a) LR, 10000 iterations; (b) RPS, 10000 iterations; (c) SIS, 1000 iterations. The exploitability of the uniform prior policy is indicated by the dashed horizontal line.

In Figure 1, we observe that prior descent for both Boltzmann and RelEnt MFE with the same uniform prior policy performs qualitatively similarly, and coincide in LR and SIS except for numerical inaccuracies. It can be seen that using a temperature sufficiently low to converge in LR and RPS allows prior descent to descend to the exact MFE iteratively. In SIS on the other hand, picking a fixed temperature that converges for the initial uniform prior policy does not guarantee monotonic improvement of exploitability afterwards. Instead, by applying the heuristic

$$\eta_{i+1} = \eta_i \cdot c$$

for each outer iteration i, where $c \geq 1$ adjusts the temperature after each outer iteration, we avoid scanning over all temperatures in each step and reach convergence to a good approximate mean field equilibrium for both Boltzmann and MaxEnt iteration.

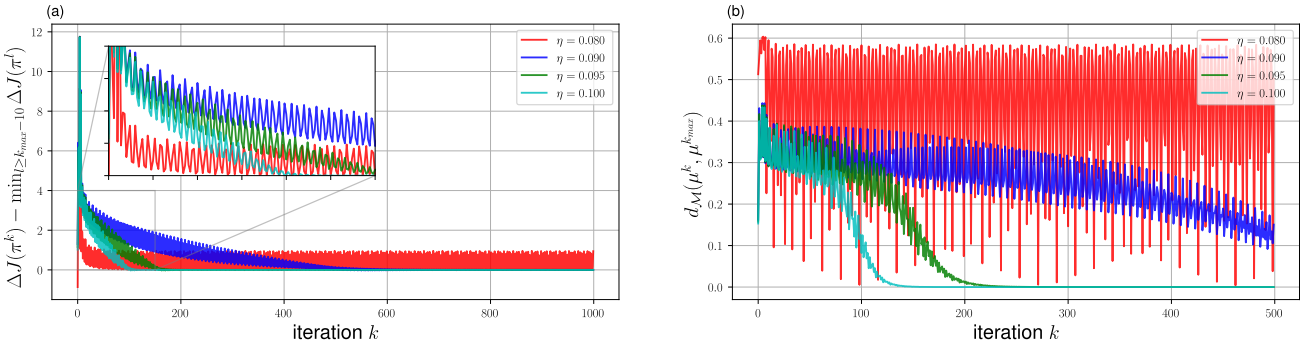

Figure 3: (a) Difference between current and final minimum exploitability over the last 10 iterations; (b) Distance between current and final mean field, cut off at 500 iterations for readability. Plotted for the η-RelEnt iterations in SIS for the indicated temperature settings and uniform prior policy.

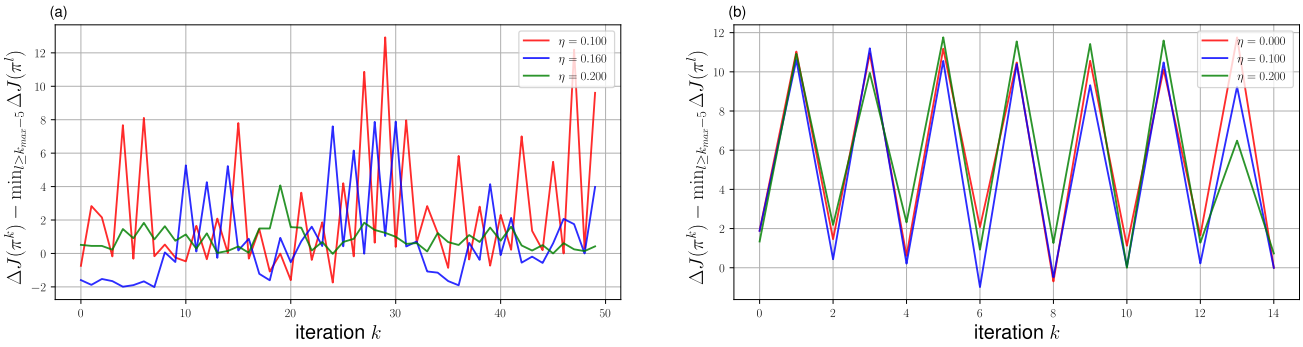

Figure 4: Difference between current and final estimated minimum exploitability over the last 5 iterations. (a) SIS, 50 iterations; (b) Taxi, 15 iterations. Plotted for the η-Boltzmann DQN iteration for the indicated temperature settings and uniform prior policy.

In Figure 2 empirical results are shown for fictitious play variants averaging only policy or mean field. In the simple one-step toy problems LR and RPS, averaging the policies appears to converge to the exact solution without regularization and to the regularized solution with regularization. Averaging the mean fields on the other hand fails, since this method can only produce deterministic policies. By applying any amount of regularization, averaging the mean fields is led to success in LR and SIS. Nonetheless, both methods fail to converge to the MFE in SIS and produce worse results than obtained by prior descent in Figure 1.

In Figure 3 we depict the convergence of exploitability and mean field of MaxEnt iteration in SIS. The results are qualitatively similar with Boltzmann iteration and, as in the main text, show the convergence behaviour near the critical temperature leading to convergence.

In Figure 4 we depict the convergence of exploitability for Boltzmann DQN iteration in SIS and Taxi during one of the runs. All 4 other runs show similar qualitative behaviour. As can be seen, the highest temperature of 0.2 shows less oscillatory behaviour, stabilizing Boltzmann DQN iteration. In Taxi, it can be seen that the used temperatures are insufficient to allow Boltzmann DQN iteration to converge. We believe that using prior descent could allow for better results. We could not verify this due to the high computational cost, as this includes repeatedly and sequentially solving an expensive reinforcement learning problem.

Finally, in Figure 5 we depict the resulting behavior in the SIS case. In the Boltzmann iteration result, at the beginning the number of infected is high enough to make social distancing the optimal action to take. As the number of infected falls, it reaches an equilibrium point where both social distancing or potentially getting infected are of equal value. Finally, as the game ends at time $t = T = 50$, there is no point in social distancing any more. Our approach yields intuitive results here, while exact fixed point iteration and FP fail to converge.

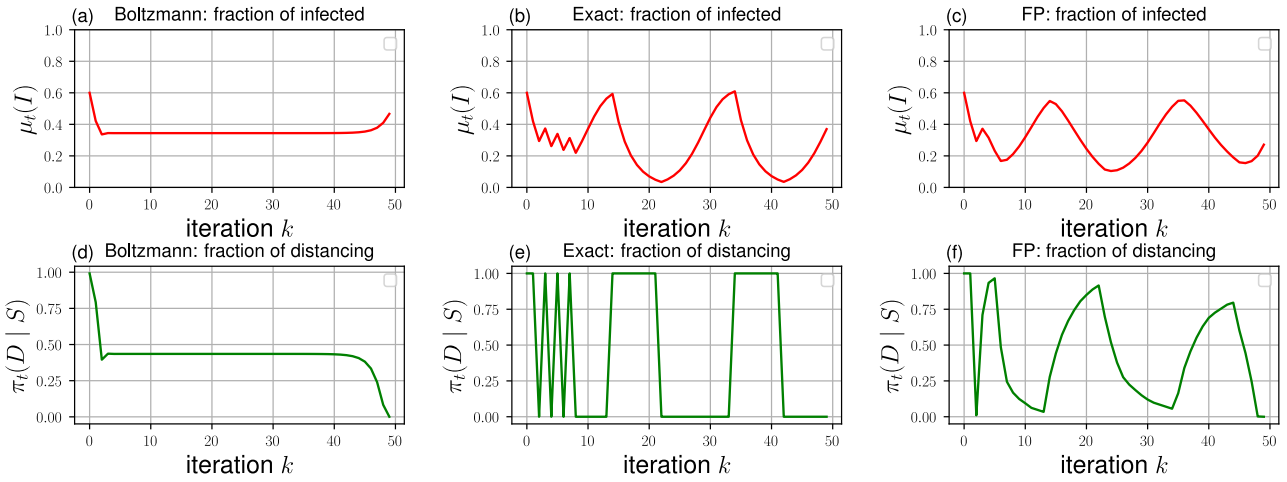

Figure 5: Fraction of infected agents and fraction of susceptible agents picking social distancing over time. (a, d): Boltzmann iteration ($\eta = 0.07$); (b, e): exact fixed point iteration; (c, f): fictitious play (averaging both policy and mean field) results in SIS after 500 iterations. More iterations and averaging only policy or mean field show same qualitative results.

B Proofs

B.1 Completeness of mean field and policy space

Lemma B.1.1. *The metric spaces* (Π, d_Π) *and* $(\mathcal{M}, d_\mathcal{M})$ *are complete metric spaces.*

Proof. The metric space $(\mathcal{M}, d_\mathcal{M})$ is a complete metric space. Let $(\mu^n)_{n\in\mathbb{N}} \in \mathcal{M}^\mathbb{N}$ be a Cauchy sequence of mean fields. Then by definition, for any $\varepsilon > 0$ there exists integer $N > 0$ such that for any $m, n > N$ we have

$$\begin{aligned} & d_\mathcal{M}(\mu^n, \mu^m) < 0.5\varepsilon \\ \implies \forall t \in \mathcal{T} : d_{TV}(\mu_t^n, \mu_t^m) &= \frac{1}{2}\sum_{s\in\mathcal{S}} |\mu_t^n(s) - \mu_t^m(s)| < 0.5\varepsilon \\ \implies \forall t \in \mathcal{T}, s \in \mathcal{S} &: |\mu_t^n(s) - \mu_t^m(s)| < \varepsilon\,. \end{aligned}$$

By completeness of $\mathbb{R}$ there exists the limit of $(\mu_t^n(s))_{n\in\mathbb{N}}$ for all $t \in \mathcal{T}, s \in \mathcal{S}$, suggestively denoted by $\mu_t(s)$. The mean field $\mu = \{\mu_t\}_{t\in\mathcal{T}}$ with the probabilities defined by the aforementioned limits fulfills $\mu^n \to \mu$ and is in $\mathcal{M}$, showing completeness of $\mathcal{M}$.

We do this analogously for (Π, d_Π). Thus, (Π, d_Π) and $(\mathcal{M}, d_\mathcal{M})$ are complete metric spaces. □

B.2 Lipschitz continuity

Lemma B.2.1. *Assume bounded and Lipschitz functions* $f : X \to \mathbb{R}$ *and* $g : X \to \mathbb{R}$ *mapping from a metric space* (X, d_X) *into* $\mathbb{R}$ *with Lipschitz constants* C_f, C_g *and bounds* $|f(x)| \le M_f$, $|g(x)| \le M_g$. *The sum of both functions* $f + g$, *the product of both functions* $f \cdot g$ *and the maximum of both functions* $\max(f, g)$ *are all Lipschitz and bounded with Lipschitz constants* $C_f + C_g$, $(M_f C_g + M_g C_f)$, $\max(C_f, C_g)$ *and bounds* $M_f + M_g$, $M_f M_g$, $\max(M_f, M_g)$.

Proof. Let $x, y \in X$ be arbitrary. By the triangle inequality, we obtain

$$|f(x) + g(x) - (f(y) + g(y))| \le |f(x) - f(y)| + |g(x) - g(y)| \le (C_f + C_g) d_X(x, y)\,.$$

Analogously, we obtain

$$|f(x)g(x) - f(y)g(y)| \le |f(x)g(x) - f(x)g(y)| + |f(x)g(y) - f(y)g(y)| \le (M_f C_g + M_g C_f) d_X(x, y)\,.$$

For the maximum of both functions, consider case by case. If $f(x) \geq g(x)$ and $f(y) \geq g(y)$ we obtain

$$|\max(f(x), g(x)) - \max(f(y), g(y))| = |f(x) - f(y)| \leq C_f d_X(x, y)$$

and analogously for $g(x) \geq f(x)$ and $g(y) \geq f(y)$

$$|\max(f(x), g(x)) - \max(f(y), g(y))| = |g(x) - g(y)| \leq C_g d_X(x, y)\,.$$

On the other hand, if $g(x) < f(x)$ and $g(y) \geq f(y)$, we have either $g(y) \geq f(x)$ and thus

$$|\max(f(x), g(x)) - \max(f(y), g(y))| = |f(x) - g(y)| = g(y) - f(x) < g(y) - g(x) \leq C_g d_X(x, y)$$

or $g(y) < f(x)$ and thus

$$|\max(f(x), g(x)) - \max(f(y), g(y))| = |f(x) - g(y)| = f(x) - g(y) \leq f(x) - f(y) \leq C_f d_X(x, y)\,.$$

The case for $f(x) < g(x)$ and $f(y) \geq g(y)$ as well as boundedness is analogous. □

B.3 Proof of Proposition 1

Proof. Since we work with finite $\mathcal{T}, \mathcal{S}, \mathcal{A}$, we identify the space of mean fields $\mathcal{M}$ with the $|\mathcal{T}|(|\mathcal{S}|-1)$-dimensional simplex $S_{|\mathcal{T}|(|\mathcal{S}|-1)} \subseteq \mathbb{R}^{|\mathcal{T}|(|\mathcal{S}|-1)}$ via the values of the probability mass functions at all times and states. Analogously the space of policies Π is identified with $S_{|\mathcal{T}||\mathcal{S}|(|\mathcal{A}|-1)} \subseteq \mathbb{R}^{|\mathcal{T}||\mathcal{S}|(|\mathcal{A}|-1)}$.

Define the set-valued map $\hat{\Gamma} : S_{|\mathcal{T}||\mathcal{S}|(|\mathcal{A}|-1)} \to 2^{S_{|\mathcal{T}||\mathcal{S}|(|\mathcal{A}|-1)}}$ mapping from a policy π represented by the input vector, to the set of vector representations of optimal policies in the MDP induced by $\Psi(\pi)$.

A policy π is optimal in the MDP induced by $\mu \in \mathcal{M}$ if and only if its value function defined by

$$V^{\pi}(\mu, t, s) = \sum_{a \in \mathcal{A}} \pi_t(a \mid s) \left(r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) V^{\pi}(\mu, t+1, s') \right),$$

is equal to the optimal action-value function defined by

$$V^{*}(\mu, t, s) = \max_{a \in \mathcal{A}} \left(r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) V^{*}(\mu, t+1, s') \right)$$

for every $t \in \mathcal{T}, s \in \mathcal{S}$, with terminal conditions $V^{*}(\mu, T, s) \equiv V^{\pi}(\mu, T, s) \equiv 0$. Moreover, an optimal policy always exists. For more details, see e.g. Puterman (2014). Define the optimal action-value function for every $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ via

$$Q^{*}(\mu, t, s, a) = r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) V^{*}(\mu, t+1, s')$$

with terminal condition $Q^{*}(\mu, T, s, a) \equiv 0$. Then, the following lemma characterizes optimality of policies.

Lemma B.3.1. *A policy π fulfills $\pi \in \hat{\Gamma}(\hat{\pi})$ if and only if*

$$\pi_t(a \mid s) > 0 \implies a \in \arg\max_{a' \in \mathcal{A}} Q^{*}(\Psi(\hat{\pi}), t, s, a')$$

for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.

Proof. To see the implication, consider $\pi \in \hat{\Gamma}(\hat{\pi})$. Then, if the right-hand side was false, there exists a maximal $t \in \mathcal{T}$ and $s \in \mathcal{S}, a \in \mathcal{A}$ such that $\pi_t(a \mid s) > 0$ but $a \notin \arg\max_{a' \in \mathcal{A}} Q^{*}(\Psi(\hat{\pi}), t, s, a')$. Since for any $t' > t$ we have optimality, $V^{\pi}(\mu, t+1, s') = V^{*}(\mu, t+1, s')$ by induction. However, $V^{\pi}(\mu, t, s) < V^{*}(\mu, t, s)$ since the suboptimal action is assigned positive probability, contradicting optimality of π. On the other hand, if the right-hand side is true, then $V^{\pi}(\mu, t, s) = V^{*}(\mu, t, s)$ by induction, which implies that π is optimal. ■

We will now check that the requirements of Kakutani's fixed point theorem hold for $\hat{\Gamma}$. The finite-dimensional simplices are convex, closed and bounded, hence compact. $\hat{\Gamma}$ maps to a non-empty set, as the induced mean field is uniquely defined and any finite MDP (induced by this mean field) has an optimal policy.

For any π, $\hat{\Gamma}(\pi)$ is convex, since the set of optimal policies is convex as shown in the following. Consider a convex combination $\tilde{\pi} = \lambda\pi + (1-\lambda)\pi'$ of optimal policies π, π' for $\lambda \in [0,1]$. Then, the resulting policy will be optimal, since we have

$$\tilde{\pi}_t(a \mid s) > 0 \implies \pi_t(a \mid s) > 0 \vee \pi'_t(a \mid s) > 0 \implies a \in \arg\max_{a \in \mathcal{A}} Q^*(\Psi(\hat{\pi}), t, s, a)$$

for any $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ and thus optimality by Lemma B.3.1.

Finally, we show that $\hat{\Gamma}$ has a closed graph. Consider arbitrary sequences $(\pi_n, \pi'_n) \to (\pi, \pi')$ with $\pi'_n \in \hat{\Gamma}(\pi_n)$. It is then sufficient to show that $\pi' \in \hat{\Gamma}(\pi)$. By the standing assumption, we have continuity of Ψ and $\mu \to Q^*(\mu, t, s, a)$ for any $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$, as sums, products and compositions of continuous functions remain continuous. Therefore, the composition $\pi \to Q^*(\Psi(\pi), t, s, a)$ is continuous. To show that $\pi' \in \hat{\Gamma}(\pi)$, assume that $\pi' \notin \hat{\Gamma}(\pi)$. By Lemma B.3.1 there exists $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ such that $\pi'_t(a \mid s) > 0$ and further there exists $a' \in \mathcal{A}$ such that $Q^*(\Psi(\pi), t, s, a') > Q^*(\Psi(\pi), t, s, a)$. Fix such an $a' \in \mathcal{A}$. Let $\delta \equiv Q^*(\Psi(\pi), t, s, a') - Q^*(\Psi(\pi), t, s, a)$, then by continuity there exists $\varepsilon > 0$ such that for all $\hat{\pi} \in \Pi$ we have

$$d_\Pi(\hat{\pi}, \pi) < \varepsilon \implies |Q^*(\Psi(\hat{\pi}), t, s, a) - Q^*(\Psi(\pi), t, s, a)| < \frac{\delta}{2}\,.$$

By convergence, there is an integer $N \in \mathbb{N}$ such that for all $n > N$ we have $d_\Pi(\pi_n, \pi) < \varepsilon$ and therefore

$$Q^*(\Psi(\pi_n), t, s, a') > Q^*(\Psi(\pi), t, s, a') - \frac{\delta}{2} = Q^*(\Psi(\pi), t, s, a) + \frac{\delta}{2} > Q^*(\Psi(\pi_n), t, s, a)\,.$$

Since $(\pi'_n)_t(a \mid s) \to \pi'_t(a \mid s) > 0$, there also exists $M \in \mathbb{N}$ such that for all $m > M$,

$$|(\pi'_m)_t(a \mid s) - \pi'_t(a \mid s)| < \pi'_t(a \mid s)\,.$$

Let $n > \max(N, M)$, then it follows that $(\pi'_n)_t(a \mid s) > 0$ but $a \notin \arg\max_{a' \in \mathcal{A}} Q^*(\Psi(\pi), t, s, a')$ since we have $Q^*(\Psi(\pi_n), t, s, a') > Q^*(\Psi(\pi_n), t, s, a)$, contradicting $\pi'_n \in \hat{\Gamma}(\pi_n)$ by Lemma B.3.1. Hence, $\hat{\Gamma}$ must have a closed graph.

By Kakutani's fixed point theorem, there exists a fixed point π^* that generates some mean field $\Psi(\pi^*)$. The associated pair $(\pi^*, \Psi(\pi^*))$ is an MFE by definition. □

B.4 Proof of Proposition 3

Proof. The space of mean fields $(\mathcal{M}, d_\mathcal{M})$ is equivalent to convex and compact finite-dimensional simplices. In this representation, each coordinate of the operators $\tilde{\Gamma}_\eta(\mu)$ and $\Gamma_\eta(\mu)$ consists of compositions, sums and products of continuous functions, since the functions $r(s, a, \mu_t)$ and $p(s' \mid s, a, \mu_t)$ are assumed to be continuous. Existence of a fixed point follows immediately by Brouwer's fixed point theorem. □

B.5 Proof of Theorem 1

Proof. The proof is a slightly simplified version of the one found in Saldi et al. (2018). Note that we require the results later, so for convenience we give the full details.

The empirical measure $\mathbb{G}^N_{S_t}$ is a random variable on $\mathcal{P}(\mathcal{S})$, i.e. its law $\mathcal{L}(\mathbb{G}^N_{S_t}) \in \mathcal{P}(\mathcal{P}(\mathcal{S}))$ is a distribution over probability measures. Since we want to show convergence of the empirical measure to the mean field, let us pick a metric on $\mathcal{P}(\mathcal{P}(\mathcal{S}))$. Remember that we metrized $\mathcal{P}(\mathcal{S})$ with the total variation distance. We metrize $\mathcal{P}(\mathcal{P}(\mathcal{S}))$ with the 1-Wasserstein metric defined for any $\Phi, \Psi \in \mathcal{P}(\mathcal{P}(\mathcal{S}))$ by the infimum over couplings

$$W_1(\Phi, \Psi) \equiv \inf_{\mathcal{L}(X_1)=\Phi, \mathcal{L}(X_2)=\Psi} \mathbb{E}\left[d_{TV}(X_1, X_2)\right]\,.$$

Lemma B.5.1. *Let $\{\Phi_n\}_{n\in\mathbb{N}}$ be a sequence of measures with $\Phi_n \in \mathcal{P}(\mathcal{P}(\mathcal{S}))$ for all $n \in \mathbb{N}$. Further, let $\mu \in \mathcal{P}(\mathcal{S})$ arbitrary. Then, the following are equivalent.*

(a) $W_1(\Phi_n, \delta_\mu) \to 0$ *as* $n \to \infty$

(b) $\mathbb{E}\left[|F(X_n) - F(X)|\right] \to 0$ *as* $n \to \infty$ *for any continuous, bounded* $F : \mathcal{P}(\mathcal{S}) \to \mathbb{R}$*, any sequence* $\{X_n\}_{n\in\mathbb{N}}$ *of* $\mathcal{P}(\mathcal{S})$*-valued random variables and any* $\mathcal{P}(\mathcal{S})$*-valued random variable* X *with* $\mathcal{L}(X_n) = \Phi_n$ *and* $\mathcal{L}(X) = \delta_\mu$*.*

(c) $\mathbb{E}\left[|X_n(f) - X(f)|\right] \to 0$ *as* $n \to \infty$ *for any* $f : \mathcal{S} \to \mathbb{R}$*, any sequence* $\{X_n\}_{n\in\mathbb{N}}$ *of* $\mathcal{P}(\mathcal{S})$*-valued random variables and any* $\mathcal{P}(\mathcal{S})$*-valued random variable* X *with* $\mathcal{L}(X_n) = \Phi_n$ *and* $\mathcal{L}(X) = \delta_\mu$*.*

Proof. Define the only possible coupling $\Delta_n \equiv \Phi_n \times \delta_\mu$.

(b), (c) $\implies$ (a):

Define $F_s(x) \equiv x(s)$ and $f_s(s') \equiv \mathbf{1}_{\{s\}}(s')$ for all $s \in \mathcal{S}$, where F_s is continuous. By assumption,

$$\begin{aligned} W_1(\Phi_n, \delta_\mu) &= \inf_{\mathcal{L}(X_n)=\Phi_n, \mathcal{L}(X)=\delta_\mu} \mathbb{E}\left[d_{TV}(X_n, X)\right] \\ &= \frac{1}{2}\int_{\mathcal{P}(\mathcal{S})\times\mathcal{P}(\mathcal{S})} \sum_{s\in\mathcal{S}} |X_n(s) - X(s)|\, d\Delta_n \\ &= \frac{1}{2}\sum_{s\in\mathcal{S}} \mathbb{E}\left[|X_n(s) - X(s)|\right] \to 0 \end{aligned}$$

since for any $s \in \mathcal{S}$, we have

$$\mathbb{E}\left[|X_n(s) - X(s)|\right] = \mathbb{E}\left[|F_s(X_n) - F_s(X)|\right] = \mathbb{E}\left[|X_n(f_s) - X(f_s)|\right] .$$

(a) $\implies$ (b), (c):

We have

$$\begin{aligned} \mathbb{E}\left[|F(X_n) - F(X)|\right] &= \int_{\mathcal{P}(\mathcal{S})\times\mathcal{P}(\mathcal{S})} |F(\nu) - F(\nu')|\, \Delta_n(d\nu, d\nu') \\ &= \int_{\mathcal{P}(\mathcal{S})} |F(\nu) - F(\mu)|\, \Phi_n(d\nu) \\ &\to \int_{\mathcal{P}(\mathcal{S})} |F(\nu) - F(\mu)|\, \delta_\mu(d\nu) = 0 \end{aligned}$$

by continuity and boundedness of $|F(\nu) - F(\mu)|$, and convergence in W_1 implying weak convergence. Analogously,

$$\mathbb{E}\left[|X_n(f) - X(f)|\right] = \int_{\mathcal{P}(\mathcal{S})} |\nu(f) - \mu(f)|\, \Phi_n(d\nu) \to \int_{\mathcal{P}(\mathcal{S})} |\nu(f) - \mu(f)|\, \delta_\mu(d\nu) = 0$$

since f and thus $|\nu(f) - \mu(f)|$ is automatically bounded from finiteness of $\mathcal{S}$, and $\nu(f) = \sum_{s\in\mathcal{S}} \nu(s)f(s) \to \sum_{s\in\mathcal{S}} \mu(s)f(s)$ as $\nu \to \mu$ in total variation distance implies continuity of $|\nu(f) - \mu(f)|$. ∎

First, it is shown that when all other agents follow the same policy π, then the empirical distribution is essentially the deterministic mean field as $N \to \infty$, i.e. $\mathcal{L}(\mathbb{G}^N_{S_t}) \to \mathcal{L}(\mu_t) \equiv \delta_{\mu_t}$ with $\mu = \Psi(\pi)$

Lemma B.5.2. *Consider a set of policies* $(\tilde{\pi}, \pi, \ldots, \pi) \in \Pi^N$ *for all agents. Under this set of policies, the law of the empirical distribution* $\mathcal{L}(\mathbb{G}^N_{S_t}) \in \mathcal{P}(\mathcal{M})$ *converges to* δ_{μ_t} *where* $\mu = \Psi(\pi)$ *as* $N \to \infty$ *in 1-Wasserstein distance.*

Proof. Define the Markov kernel $P^\pi_{t,\nu}$ such that its probability mass function fulfills

$$P^\pi_{t,\nu}(s' \mid s) \equiv \sum_{a\in\mathcal{A}} \pi_t(a \mid s) p(s' \mid s, a, \nu)$$

for any $t \in \mathcal{T}, s \in \mathcal{S}, \nu \in \mathcal{P}(\mathcal{S}), \pi \in \Pi$ and analogously

$$\tilde{\nu}P^\pi_{t,\nu}(s') \equiv \sum_{s\in\mathcal{S}} \tilde{\nu}(s) \sum_{a\in\mathcal{A}} \pi_t(a \mid s) p(s' \mid s, a, \nu)$$

for any $\tilde{\nu} \in \mathcal{P}(\mathcal{S})$. Note that $\mu_{t+1} = \mu_t P^{\pi}_{t,\mu_t}(g)$ for mean fields $\mu = \Psi(\pi)$ induced by π.

We show that $\mathbb{E}\left[\left|\mathbb{G}^N_{S_t}(f) - \mu_t(f)\right|\right] \to 0$ as $N \to \infty$ for any function $f : \mathcal{S} \to \mathbb{R}$ and any time $t \in \mathcal{T}$. From this, the desired result follows by Lemma B.5.1. Since $\mathbb{G}^N_{S_t}(\cdot) \equiv \frac{1}{N}\sum_{i=1}^N \delta_{S^i_t}(\cdot)$ and $S^i_0 \sim \mu_0$ we have at time $t = 0$ that

$$\lim_{N\to\infty} \mathbb{E}\left[\left|\mathbb{G}^N_{S_0}(f) - \mu_0(f)\right|\right] = \lim_{N\to\infty} \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N f(S^i_0) - \mathbb{E}\left[f(S^i_0)\right]\right|\right] = 0$$

by the strong law of large numbers and the dominated convergence theorem.

Assuming this holds for t, then for $t+1$ we have

$$\begin{aligned}
\mathbb{E}\left[\left|\mathbb{G}^N_{S_{t+1}}(f) - \mu_{t+1}(f)\right|\right] \leq\ & \mathbb{E}\left[\left|\mathbb{G}^N_{S_{t+1}}(f) - \mathbb{G}^{N-1}_{S_{t+1}}(f)\right|\right] \\
&+ \mathbb{E}\left[\left|\mathbb{G}^{N-1}_{S_{t+1}}(f) - \mathbb{G}^{N-1}_{S_t} P^{\pi}_{t,\mathbb{G}^N_{S_t}}(f)\right|\right] \\
&+ \mathbb{E}\left[\left|\mathbb{G}^{N-1}_{S_t} P^{\pi}_{t,\mathbb{G}^N_{S_t}}(f) - \mathbb{G}^N_{S_t} P^{\pi}_{t,\mathbb{G}^N_{S_t}}(f)\right|\right] \\
&+ \mathbb{E}\left[\left|\mathbb{G}^N_{S_t} P^{\pi}_{t,\mathbb{G}^N_{S_t}}(f) - \mu_t P^{\pi}_{t,\mu_t}(f)\right|\right]
\end{aligned}$$

where we defined $\mathbb{G}^{N-1}_{S_t}(\cdot) \equiv \frac{1}{N-1}\sum_{i=2}^N \delta_{S^i_t}(\cdot)$.

For the first term, we have as $N \to \infty$

$$\begin{aligned}
\mathbb{E}\left[\left|\mathbb{G}^N_{S_{t+1}}(f) - \mathbb{G}^{N-1}_{S_{t+1}}(f)\right|\right] &= \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N f(S^i_{t+1}) - \frac{1}{N-1}\sum_{i=2}^N f(S^i_{t+1})\right|\right] \\
&\leq \frac{1}{N}\mathbb{E}\left[\left|f(S^1_{t+1})\right|\right] + \left|\frac{1}{N} - \frac{1}{N-1}\right| \sum_{i=2}^N \mathbb{E}\left[\left|f(S^i_{t+1})\right|\right] \\
&\leq \left(\frac{1}{N} + \frac{N-1}{N(N-1)}\right) \max_{s\in\mathcal{S}} |f(s)| \to 0\,.
\end{aligned}$$

For the second term, as $N \to \infty$ we have by Jensen's inequality and bounds $|f| \leq M_f$ (by finiteness of $\mathcal{S}$)

$$\begin{aligned}
\mathbb{E}\left[\left|\mathbb{G}^{N-1}_{S_{t+1}}(f) - \mathbb{G}^{N-1}_{S_t} P^{\pi}_{t,\mathbb{G}^{N-1}_{S_t}}(f)\right|\right]^2 &= \mathbb{E}\left[\mathbb{E}\left[\left|\mathbb{G}^{N-1}_{S_{t+1}}(f) - \mathbb{G}^{N-1}_{S_t} P^{\pi}_{t,\mathbb{G}^{N-1}_{S_t}}(f)\right| \mid S_t\right]\right]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[\left|\frac{1}{N-1}\sum_{i=2}^N \left(f(S^i_{t+1}) - \mathbb{E}\left[f(S^i_{t+1})\right]\right)\right| \mid S_t\right]\right]^2 \\
&\leq \frac{1}{(N-1)^2}\sum_{i=2}^N \mathbb{E}\left[\mathbb{E}\left[\left(f(S^i_{t+1}) - \mathbb{E}\left[f(S^i_{t+1})\right]\right)^2 \mid S_t\right]\right] \\
&\leq \frac{1}{N-1}\cdot 4M_f^2 \to 0\,.
\end{aligned}$$

For the third term, we again have as $N \to \infty$

$$\begin{aligned}
\mathbb{E}\left[\left|\mathbb{G}^{N-1}_{S_t} P^{\pi}_{t,\mathbb{G}^N_{S_t}}(f) - \mathbb{G}^N_{S_t} P^{\pi}_{t,\mathbb{G}^N_{S_t}}(f)\right|\right] &= \mathbb{E}\left[\left|\sum_{s\in\mathcal{S}} \left(\mathbb{G}^{N-1}_{S_t}(s) - \mathbb{G}^N_{S_t}(s)\right) \sum_{a\in\mathcal{A}} \pi_t(a \mid s) \sum_{s'\in\mathcal{S}} p(s' \mid s, a, \mathbb{G}^N_{S_t}) f(s')\right|\right] \\
&\leq \mathbb{E}\left[\left|\left(\frac{1}{N-1} - \frac{1}{N}\right)\sum_{i=2}^N \sum_{a\in\mathcal{A}} \pi_t(a \mid S^i_t) \sum_{s'\in\mathcal{S}} p(s' \mid S^i_t, a, \mathbb{G}^N_{S_t}) f(s')\right|\right] \\
&\quad + \mathbb{E}\left[\left|\frac{1}{N}\sum_{a\in\mathcal{A}} \pi_t(a \mid S^1_t) \sum_{s'\in\mathcal{S}} p(s' \mid S^1_t, a, \mathbb{G}^N_{S_t}) f(s')\right|\right]
\end{aligned}$$

$$\leq \left(\frac{N-1}{N(N-1)} + \frac{1}{N}\right) \max_{s\in\mathcal{S}} |f(s)| \to 0\,.$$

For the fourth term, define $F : \mathcal{P}(\mathcal{S}) \to \mathbb{R}$, $F(\nu) = \nu P^{\pi}_{t,\nu}(f)$ and observe that F is continuous, since $\nu \to \nu'$ if and only if $\nu(s) \to \nu'(s)$ for all $s \in \mathcal{S}$, and therefore (as p is assumed continuous by Assumption 1)

$$F(\nu) = \nu P^{\pi}_{t,\nu}(f) = \sum_{s\in\mathcal{S}} \nu(s) \sum_{a\in\mathcal{A}} \pi_t(a \mid s) \sum_{s'\in\mathcal{S}} p(s' \mid s, a, \nu) f(s')$$

is continuous for any $s' \in \mathcal{S}$. By Lemma B.5.1, we have from the induction hypothesis $\mathbb{G}^N_{S_t} \to \mu_t$ that

$$\mathbb{E}\left[\left|\mathbb{G}^N_{S_t} P^{\pi}_{t,\mathbb{G}^N_{S_t}}(f) - \mu_t P^{\pi}_{t,\mu_t}(f)\right|\right] \to 0\ .$$

Therefore, $\mathbb{E}\left[\left|\mathbb{G}^N_{S_{t+1}}(f) - \mu_{t+1}(f)\right|\right] \to 0$ which implies the desired result by induction. ■

Consider the case where all agents follow a set of policies $(\pi^N, \pi, \ldots, \pi) \in \Pi^N$ for each $N \in \mathbb{N}$. Define new single-agent random variables S^{μ}_t and A^{μ}_t with $S^{\mu}_0 \sim \mu_0$ and

$$\mathbb{P}(A^{\mu}_t = a \mid S^{\mu}_t = s) = \pi^N_t(a \mid s),$$
$$\mathbb{P}(S^{\mu}_{t+1} = s' \mid S^{\mu}_t = s, A^{\mu}_t = a) = p(s' \mid s, a, \mu_t)\,,$$

where the deterministic mean field μ is used instead of the empirical distribution.

Lemma B.5.3. *Consider an equicontinuous, uniformly bounded family of functions $\mathcal{F}$ on $\mathcal{P}(\mathcal{S})$ and define*

$$F_t(\nu) \equiv \sup_{f\in\mathcal{F}} |f(\nu) - f(\mu_t)|$$

for any $t \in \mathcal{T}$. Then, F_t is continuous and bounded and by Lemma B.5.1 we have

$$\lim_{N\to\infty} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|f(\mathbb{G}^N_{S_t}) - f(\mu)\right|\right] = 0$$

Proof. F_t is continuous, since for $\nu_n \to \nu$

$$|F_t(\nu_n) - F_t(\nu)| = \left|\sup_{f\in\mathcal{F}} |f(\nu) - f(\mu_t)| - \sup_{f\in\mathcal{F}} |f(\nu') - f(\mu_t)|\right| \leq \sup_{f\in\mathcal{F}} |f(\nu) - f(\nu')| \to 0$$

by equicontinuity. Further, F_t is bounded since $|F_t(\nu)| \leq \sup_{f\in\mathcal{F}} |f(\nu)| + |f(\mu_t)|$ is uniformly bounded. By Lemma B.5.2, we have $W_1(\mathbb{G}^N_{S_t}, \delta_{\mu_t}) \to 0$ as $N \to \infty$, therefore Lemma B.5.1 applies. ■

Lemma B.5.4. *Suppose that at some time $t \in \mathcal{T}$, it holds that*

$$\lim_{N\to\infty} \left|\mathcal{L}(S^1_t)(g_N) - \mathcal{L}(S^{\mu}_t)(g_N)\right| = 0$$

for any sequence of functions $\{g_N\}_{N\in\mathbb{N}}$ from $\mathcal{S}$ to $\mathbb{R}$ that is uniformly bounded. Then, we have

$$\lim_{N\to\infty} \left|\mathcal{L}(S^1_t, \mathbb{G}^N_{S_t})(T_N) - \mathcal{L}(S^{\mu}_t, \mu_t)(T_N)\right| = 0$$

for any sequence of functions $\{T_N\}_{N\in\mathbb{N}}$ from $\mathcal{S} \times \mathcal{P}(\mathcal{S})$ to $\mathbb{R}$ that is equicontinuous and uniformly bounded.

Proof. We have

$$\left|\mathcal{L}(S^1_t, \mathbb{G}^N_{S_t})(T_N) - \mathcal{L}(S^{\mu}_t, \mu_t)(T_N)\right| \leq \left|\mathcal{L}(S^1_t, \mathbb{G}^N_{S_t})(T_N) - \mathcal{L}(S^1_t, \mu_t)(T_N)\right| + \left|\mathcal{L}(S^1_t, \mu_t)(T_N) - \mathcal{L}(S^{\mu}_t, \mu_t)(T_N)\right|$$

The first term becomes

$$\begin{aligned}\left|\mathcal{L}(S_t^1, \mathbb{G}_{S_t}^N)(T_N) - \mathcal{L}(S_t^1, \mu_t)(T_N)\right| &= \left|\int T_N(x,\nu)\mathcal{L}(S_t^1, \mathbb{G}_{S_t}^N)(dx, d\nu) - \int T_N(x,\nu)\mathcal{L}(S_t^1, \mu_t)(dx, d\nu)\right| \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\left|T_N(x, G_{S_t}^N) - T_N(x, \mu_t)\right| S_t^1\right]\right] \\ &\leq \mathbb{E}\left[\sup_{f \in \{T_N(\cdot,\nu)\}_{\nu \in \mathcal{P}(\mathcal{S}), N \in \mathbb{N}}} \left|f(G_{S_t}^N) - f(\mu_t)\right|\right] \to 0\end{aligned}$$

by Lemma B.5.3, since $\{T_N\}_{N\in\mathbb{N}}$ is equicontinuous and uniformly bounded. Similarly for the second term,

$$\begin{aligned}\left|\mathcal{L}(S_t^1, \mu_t)(T_N) - \mathcal{L}(S_t^\mu, \mu_t)(T_N)\right| &= \left|\int T_N(x,\nu)\mathcal{L}(S_t^1, \mu_t)(dx, d\nu) - \int T_N(x,\nu)\mathcal{L}(S_t^\mu, \mu_t)(dx, d\nu)\right| \\ &\leq \mathbb{E}\left[\left|T_N(S_t^1, \mu_t) - T_N(S_t^\mu, \mu_t)\right|\right] \to 0\end{aligned}$$

by the assumption, since T_N fulfills the condition of being uniformly bounded. ■

Lemma B.5.5. *For any sequence $\{g_N\}_{N\in\mathbb{N}}$ of functions from $\mathcal{S}$ to $\mathbb{R}$ that is uniformly bounded, we have*

$$\lim_{N\to\infty} \left|\mathcal{L}(S_t^1)(g_N) - \mathcal{L}(S_t^\mu)(g_N)\right| = 0$$

for all times $t \in \mathcal{T}$.

Proof. Define $l_{N,t}$ as

$$l_{N,t}(s,\nu) \equiv \sum_{a\in\mathcal{A}} \pi_t^N(a \mid s) \sum_{s'\in\mathcal{S}} p(s' \mid s, a, \nu) g_N(s')\,.$$

$\{l_{N,t}(s,\cdot)\}_{s\in\mathcal{S}, N\in\mathbb{N}}$ is equicontinuous, since for any $\nu, \nu' \in \mathcal{M}$ with $d_{TV}(\nu,\nu') \to 0$,

$$\begin{aligned}\sup_{s\in\mathcal{S}, N\in\mathbb{N}} |l_{N,t}(s,\nu) - l_{N,t}(s,\nu')| &\leq M_g \sup_{s\in\mathcal{S}, N\in\mathbb{N}} \left|\sum_{a\in\mathcal{A}} \pi_t^N(a \mid s) \sum_{s'\in\mathcal{S}} \left(p(s' \mid s, a, \nu) - p(s' \mid s, a, \nu')\right)\right| \\ &\leq M_g |\mathcal{S}| \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \max_{s'\in\mathcal{S}} |p(s' \mid s, a, \nu) - p(s' \mid s, a, \nu')| \to 0\end{aligned}$$

since $|g_N| < M_g$ is uniformly bounded and p is continuous by assumption. Furthermore, $l_{N,t}(s,\nu)$ is always uniformly bounded by M_g. Now the result can be shown by induction.

For $t = 0$, $\mathcal{L}(S_0^\mu) = \mathcal{L}(S_0^1)$ fulfills the hypothesis. Assume this holds for t, then

$$\left|\mathcal{L}(S_{t+1}^1)(g_N) - \mathcal{L}(S_{t+1}^\mu)(g_N)\right| = \left|\mathcal{L}(S_t^1, \mathbb{G}_{S_t}^N)(l_{N,t}) - \mathcal{L}(S_t^\mu, \mu_t)(l_{N,t})\right| \to 0$$

as $N \to \infty$ by Lemma B.5.4. ■

Thus, for any sequence of policies $\{\pi^N\}_{N\in\mathbb{N}}$ with $\pi^N \in \Pi$ for all $N \in \mathbb{N}$, the achieved return of the N-agent game converges to the return of the mean field game under the mean field generated by the other agent's policy π as $N \to \infty$.

Lemma B.5.6. *Let $\{\pi^N\}_{N\in\mathbb{N}}$ with $\pi^N \in \Pi$ for all $N \in \mathbb{N}$ be an arbitrary sequence of policies and $\pi \in \Pi$ an arbitrary policy. Further, let the mean field $\mu = \Psi(\pi)$ be generated by π. Then, under the joint policy $(\pi^N, \pi, \ldots, \pi)$, we have as $N \to \infty$ that*

$$\left|J_1^N(\pi^N, \pi, \ldots, \pi) - J^\mu(\pi^N)\right| \to 0\,.$$

Proof. Define for any $t \in \mathcal{T}$, $N \in \mathbb{N}$

$$r_{\pi_t^N}(s,\nu) \equiv \sum_{a\in\mathcal{A}} r(s,a,\nu)\pi_t^N(a \mid s)$$

such that the family $\{r_{\pi_t^N}(s,\cdot)\}_{s\in\mathcal{S},N\in\mathbb{N}}$ is equicontinuous, since for any $\nu,\nu'\in\mathcal{M}$ as $d_{\mathcal{M}}(\nu,\nu')\to 0$,

$$\max_{s\in\mathcal{S}}\max_{N\in\mathbb{N}}\left|r_{\pi_t^N}(s,\nu)-r_{\pi_t^N}(s,\nu')\right|\to 0$$

by continuity of r. The function $r_{\pi_t^N}$ is uniformly bounded for all $N\in\mathbb{N}$ by assumption of uniformly bounded r. By Lemma B.5.4 and Lemma B.5.5,

$$\begin{aligned}&\lim_{N\to\infty}\left|\mathbb{E}\left[r(S_t^1,A_t^1,\mathbb{G}_{S_t}^N)\right]-\mathbb{E}\left[r(S_t^\mu,A_t^\mu,\mu_t)\right]\right|\\&=\lim_{N\to\infty}\left|\mathbb{E}\left[r_{\pi_t^N}(S_t^1,\mathbb{G}_{S_t}^N)\right]-\mathbb{E}\left[r_{\pi_t^N}(S_t^\mu,\mu_t)\right]\right|=0\,.\end{aligned}$$

such that we have

$$\lim_{N\to\infty}\left|J_1^N(\pi^N,\pi,\dots,\pi)-J^\mu(\pi^N)\right|\le\sum_{t\in\mathcal{T}}\lim_{N\to\infty}\left|\mathbb{E}\left[r(S_t^1,A_t^1,\mathbb{G}_{S_t}^N)\right]-\mathbb{E}\left[r(S_t^\mu,A_t^\mu,\mu_t)\right]\right|=0\,.$$

which is the desired result. ■

From Lemma B.5.6, it follows that for any sequence of optimal exploiting policies $\{\pi^N\}_{N\in\mathbb{N}}$ with $\pi^N\in\Pi$ for all $N\in\mathbb{N}$ and

$$\pi^N\in\arg\max_{\pi\in\Pi}J_1^N(\pi,\pi^*,\dots,\pi^*)$$

for all $N\in\mathbb{N}$, it holds that for any MFE $(\pi^*,\mu^*)\in\Pi\times\mathcal{M}$,

$$\begin{aligned}\lim_{N\to\infty}J_1^N(\pi^N,\pi^*,\dots,\pi^*)&\le\max_{\pi\in\Pi}J^{\mu^*}(\pi)\\&=J^{\mu^*}(\pi^*)\\&=\lim_{N\to\infty}J_1^N(\pi^*,\dots,\pi^*)\end{aligned}$$

and by instantiating for arbitrary $\epsilon>0$, for sufficiently large N we obtain

$$\begin{aligned}J_1^N(\pi^N,\pi^*,\dots,\pi^*)-\epsilon&=\max_{\pi\in\Pi}J_1^N(\pi,\pi^*,\dots,\pi^*)-\epsilon\\&\le\max_{\pi\in\Pi}J^{\mu^*}(\pi)-\frac{\epsilon}{2}\\&=J^{\mu^*}(\pi^*)-\frac{\epsilon}{2}\\&=J_1^N(\pi^*,\pi^*,\dots,\pi^*)\end{aligned}$$

which is the desired approximate Nash property that applies to all agents by symmetry. □

B.6 Proof of Theorem 2

Proof. If Φ or Ψ is constant, or if the restriction $\Psi\restriction_{\Pi_\Phi}$ of Ψ to Π_Φ is constant, then $\Gamma=\Psi\circ\Phi$ is constant. Assume that this is not the case.

Then there exist distinct $\pi,\pi'\in\Pi_\Phi$ such that $\Psi(\pi)\neq\Psi(\pi')$. By definition of Π_Φ there also exist distinct $\mu,\mu'\in\mathcal{M}$ such that $\Phi(\mu)=\pi$ and $\Phi(\mu')=\pi'$. Note that for any $\nu,\nu'\in\mathcal{M}$ with $\Gamma(\nu)\neq\Gamma(\nu')$,

$$d_{\mathcal{M}}(\Gamma(\nu),\Gamma(\nu'))\ge\min_{\pi,\pi'\in\Pi_\Phi,\pi\neq\pi'}d_{\mathcal{M}}(\Psi(\pi),\Psi(\pi'))$$

where the right-hand side is greater zero by finiteness of Π_Φ. This holds for μ,μ'.

To show that Γ cannot be Lipschitz continuous, assume that Γ has a Lipschitz constant $C>0$. We can find an integer N such that

$$d_{\mathcal{M}}(\mu^i,\mu^{i+1})=\frac{d_{\mathcal{M}}(\mu,\mu')}{N-1}<\frac{\min_{\pi,\pi'\in\Pi_\Phi,\pi\neq\pi'}d_{\mathcal{M}}(\Psi(\pi),\Psi(\pi'))}{C}$$

for all $i \in \{0, \dots, N-1\}$ by defining

$$\mu^i = \frac{i}{N}\mu + \frac{N-i}{N}\mu'$$

for all $i \in \{0, \dots, N\}$, and $\mu^i \in \mathcal{M}$ holds. By the triangle inequality

$$d_{\mathcal{M}}(\Gamma(\mu), \Gamma(\mu')) \leq d_{\mathcal{M}}(\Gamma(\mu^0), \Gamma(\mu^1)) + \ldots + d_{\mathcal{M}}(\Gamma(\mu^{N-1}), \Gamma(\mu^N))$$

there exists a pair (μ^i, μ^{i+1}) with $\Gamma(\mu^i) \neq \Gamma(\mu^{i+1})$. For this pair, we have

$$d_{\mathcal{M}}(\Gamma(\mu^i), \Gamma(\mu^{i+1})) \geq d_{\mathcal{M}}(\Gamma(\mu), \Gamma(\mu')) \geq \min_{\pi, \pi' \in \Pi_\Phi, \pi \neq \pi'} d_{\mathcal{M}}(\Psi(\pi), \Psi(\pi'))\,.$$

On the other hand, since Γ is Lipschitz with constant C, we have

$$d_{\mathcal{M}}(\Gamma(\mu^i), \Gamma(\mu^{i+1})) \leq C \cdot d_{\mathcal{M}}(\mu^i, \mu^{i+1}) < \min_{\pi, \pi' \in \Pi_\Phi, \pi \neq \pi'} d_{\mathcal{M}}(\Psi(\pi), \Psi(\pi'))$$

which is a contradiction. Thus, Γ cannot be Lipschitz continuous and by extension cannot be contractive. □

B.7 Proof of Theorem 3

Proof. For all $\eta > 0, \mu \in \mathcal{M}, t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$, the soft action-value function of the MDP induced by $\mu \in \mathcal{M}$ is given by

$$\tilde{Q}_\eta(\mu, t, s, a) = r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t)\eta \log \sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, s', a')}{\eta}\right)$$

and terminal condition $\tilde{Q}_\eta(\mu, T-1, s, a) \equiv r(s, a, \mu_{T-1})$. Analogously, the action-value function of the MDP induced by $\mu \in \mathcal{M}$ is given by

$$Q^*(\mu, t, s, a) = r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \max_{a' \in \mathcal{A}} Q^*(\mu, t+1, s', a')$$

and the similarly defined policy action-value function for $\pi \in \Pi$ is given by

$$Q^\pi(\mu, t, s, a) = r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \sum_{a' \in \mathcal{A}} \pi_{t+1}(a' \mid s') Q^\pi(\mu, t+1, s', a')\,,$$

with terminal conditions $Q^*(\mu, T-1, s, a) \equiv Q^\pi(\mu, T-1, s, a) \equiv r(s, a, \mu_{T-1})$.

We will show that we can find a Lipschitz constant $K_{\tilde{Q}_\eta}$ of $\tilde{Q}_\eta$ that is independent of η if η is not arbitrarily small. To show this, we will explicitly compute such a Lipschitz constant. Note first that $\tilde{Q}_\eta$, Q^* and Q^π are all uniformly bounded by $M_Q \equiv |\mathcal{T}| M_r$ by assumption, where M_r is the uniform bound of r.

Lemma B.7.1. *The functions $\tilde{Q}_\eta(\mu, t, s, a)$, $Q^*(\mu, t, s, a)$ and $Q^\pi(\mu, t, s, a)$ are uniformly bounded for all $\eta > 0, \mu \in \mathcal{M}, t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ by*

$$\left|\tilde{Q}_\eta(\mu, t, s, a)\right| \leq (T-t)M_r \leq TM_r =: M_Q$$

where M_r is the uniform bound of $|r(s, a, \mu_t)| \leq M_r$, and $T = |\mathcal{T}|$.

Proof. Make the induction hypothesis for all $t \in \mathcal{T}$ that

$$\left|\tilde{Q}_\eta(\mu, t, s, a)\right| \leq (T-t)M_r$$

for all $\eta > 0, \mu \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}$ and note that this holds for $t = T-1$, as by assumption

$$\left|\tilde{Q}_\eta(\mu, T-1, s, a)\right| = |r(s, a, \mu_t)| \leq M_r\,.$$

The induction step from $t+1$ to t holds by

$$\begin{aligned}\left|\tilde{Q}_\eta(\mu,t,s,a)\right| &= \left|r(s,a,\mu_t) + \sum_{s'\in\mathcal{S}} p(s' \mid s,a,\mu_t)\eta \log \sum_{a'\in\mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu,t+1,s',a')}{\eta}\right)\right| \\ &\leq |r(s,a,\mu_t)| + \eta \max_{s'\in\mathcal{S}} \left|\log \sum_{a'\in\mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu,t+1,s',a')}{\eta}\right)\right| \\ &\leq M_r + \eta \left|\log\left(\exp\left(\frac{(T-t-1)M_r}{\eta}\right)\right)\right| \\ &= M_r + (T-t-1)M_r = (T-t)M_r\,.\end{aligned}$$

By maximizing over all $t \in \mathcal{T}$, we obtain the uniform bound. The other cases are analogous. ■

Now we can find a Lipschitz constant of $\tilde{Q}_\eta(\mu,t,s,a)$ that is independent of η.

Lemma B.7.2. *Let C_r be a Lipschitz constant of $\mu \to r(s,a,\mu_t)$ and C_p a Lipschitz constant of $\mu \to p(s' \mid s,a,\mu_t)$. Further, let $\eta_{\min} > 0$. Then, for all $\eta > \eta_{\min}, t \in \mathcal{T}$, the map $\mu \mapsto \tilde{Q}_\eta(\mu,t,s,a)$ is Lipschitz for all $s \in \mathcal{S}, a \in \mathcal{A}$ with a Lipschitz constant $K^t_{\tilde{Q}_\eta}$ independent of η. Therefore, by picking $K_{\tilde{Q}_\eta} \equiv \max_{t\in\mathcal{T}} K^t_{\tilde{Q}_\eta}$, we have one single Lipschitz constant for all $\eta > \eta_{\min}, t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.*

Proof. We show by induction that for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$, we can find Lipschitz constants such that $\tilde{Q}_\eta(\mu,t,s,a)$ is Lipschitz in μ with a Lipschitz constant that does not depend on η.

To see this, note that this is true for $t = T-1$ and any $s \in \mathcal{S}, a \in \mathcal{A}$, as for any μ, μ' we have

$$\left|\tilde{Q}_\eta(\mu,T-1,s,a) - \tilde{Q}_\eta(\mu',T-1,s,a)\right| = \left|r(s,a,\mu_{T-1}) - r(s,a,\mu'_{T-1})\right| \leq C_r d_{\mathcal{M}}(\mu,\mu')\,.$$

The induction step from $t+1$ to t is

$$\begin{aligned}&\left|\tilde{Q}_\eta(\mu,t,s,a) - \tilde{Q}_\eta(\mu,t,s,a)\right| \\ &\leq |r(s,a,\mu_t) - r(s,a,\mu'_t)| + \sum_{s'\in\mathcal{S}} \left| p(s' \mid s,a,\mu_t)\eta \log \sum_{a'\in\mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu,t+1,s',a')}{\eta}\right)\right. \\ &\qquad \left. - p(s' \mid s,a,\mu'_t)\eta \log \sum_{a'\in\mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu',t+1,s',a')}{\eta}\right)\right| \\ &\leq C_r d_{\mathcal{M}}(\mu,\mu') + \eta|\mathcal{S}| \max_{s'\in\mathcal{S}} 1 \cdot \left|\log \sum_{a'\in\mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu,t+1,s',a')}{\eta}\right)\right. \\ &\qquad \left. - \log \sum_{a'\in\mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu',t+1,s',a')}{\eta}\right)\right| \\ &\quad + \eta|\mathcal{S}| \max_{s'\in\mathcal{S}} \frac{M_Q}{\eta} \cdot |p(s' \mid s,a,\mu_t) - p(s' \mid s,a,\mu'_t)| \\ &\leq C_r d_{\mathcal{M}}(\mu,\mu') + \eta|\mathcal{S}| \max_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} \left|\frac{\frac{1}{\eta} q_{t+1}(a' \mid s') \exp\left(\frac{\xi_{a'}}{\eta}\right)}{\sum_{a''\in\mathcal{A}} q_{t+1}(a'' \mid s') \exp\left(\frac{\xi_{a''}}{\eta}\right)}\right| \left|\tilde{Q}_\eta(\mu,t+1,s',a') - \tilde{Q}_\eta(\mu',t+1,s',a')\right| \\ &\quad + |\mathcal{S}| M_Q \cdot C_p d_{\mathcal{M}}(\mu,\mu') \\ &\leq C_r d_{\mathcal{M}}(\mu,\mu') + \frac{|\mathcal{A}| q_{\max}}{|\mathcal{A}| q_{\min}} \exp\left(2 \cdot \frac{M_Q}{\eta}\right) K^{t+1}_{\tilde{Q}_\eta} d_{\mathcal{M}}(\mu,\mu') + |\mathcal{S}| M_Q C_p d_{\mathcal{M}}(\mu,\mu') \\ &< \left(C_r + \frac{q_{\max}}{q_{\min}} \exp\left(\frac{2M_Q}{\eta_{\min}}\right) K^{t+1}_{\tilde{Q}_\eta} + |\mathcal{S}| M_Q C_p\right) d_{\mathcal{M}}(\mu,\mu')\end{aligned}$$

where we use the mean value theorem to obtain some $\xi_a \in [-M_Q, M_Q]$ for all $a \in \mathcal{A}$ bounded by Lemma B.7.1, Lemma B.2.1 for the second inequality, and defined $q_{\max} = \max_{t\in\mathcal{T}, s\in\mathcal{S}, a\in\mathcal{A}} q_t(a \mid s)$, $q_{\min} = \min_{t\in\mathcal{T}, s\in\mathcal{S}, a\in\mathcal{A}} q_t(a \mid s)$. Since $s \in \mathcal{S}, a \in \mathcal{A}$ were arbitrary, this holds for all $s \in \mathcal{S}, a \in \mathcal{A}$.

Thus, as long as $\eta > \eta_{\min}$, we have the Lipschitz constant $K^t_{\tilde{Q}_\eta} \equiv \left(C_r + \frac{q_{\max}}{q_{\min}} \exp\left(\frac{2M_Q}{\eta_{\min}}\right) K^{t+1}_{\tilde{Q}_\eta} + |\mathcal{S}| M_Q C_p\right)$ independent of η, since by induction assumption $K^{t+1}_{\tilde{Q}_\eta}$ is independent of η. ∎

The optimal action-value function and the policy action-value function for any fixed policy are Lipschitz in μ.

Lemma B.7.3. *The functions $\mu \mapsto Q^*(\mu, t, s, a)$ and $\mu \mapsto Q^\pi(\mu, t, s, a)$ for any fixed $\pi \in \Pi, t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ are Lipschitz continuous. Therefore, for any fixed $\pi \in \Pi$ we can choose a Lipschitz constant K_Q for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ by taking the maximum over all Lipschitz constants.*

Proof. The action-value function is given by the recursion

$$Q^*(\mu, t, s, a) = r(s, a, \mu_t) + \sum_{s'\in\mathcal{S}} p(s' \mid s, a, \mu_t) \max_{a'\in\mathcal{A}} Q^*(\mu, t+1, s', a')$$

with terminal condition $Q^*(\mu, T-1, s, a) \equiv r(s, a, \mu_{T-1})$. The functions $r(s, a, \mu_t)$ and $p(s' \mid s, a, \mu_t)$ are Lipschitz continuous by Assumption 2. Note that for any $\mu, \mu' \in \mathcal{M}$ and any $t \in \mathcal{T}$, $d_{TV}(\mu_t, \mu'_t) \leq d_M(\mu, \mu')$. Therefore, the terminal condition and all terms in the above recursion are Lipschitz. Further, $Q^*(\mu, t, s, a)$ is uniformly bounded, since r is assumed uniformly bounded.

Since a finite maximum, product and sum of Lipschitz and bounded functions is again Lipschitz and bounded by Lemma B.2.1, we obtain Lipschitz constants $K_{Q,t,s,a}$ of the maps $\mu \to Q^*(\mu, t, s, a)$ for any $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ and define $K_Q \equiv \max_{t\in\mathcal{T}, s\in\mathcal{S}, a\in\mathcal{A}} K_{Q,t,s,a}$. The case for Q^π with fixed $\pi \in \Pi$ is analogous. ∎

The same holds for $\Psi(\pi)$ mapping from policy π to its induced mean field.

Lemma B.7.4. *The function $\Psi(\pi)$ is Lipschitz with some Lipschitz constant K_Ψ.*

Proof. Recall that $\Psi(\pi)$ maps to the mean field μ starting with μ_0 and obtained by the recursion

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

We proceed analogously to Lemma B.7.3. μ is uniformly bounded by normalization. The constant function $\pi \mapsto \mu_0(s)$ is Lipschitz and bounded for any $s \in \mathcal{S}$. The functions $r(s, a, \mu_t)$ and $p(s' \mid s, a, \mu_t)$ are Lipschitz continuous by Assumption 2. Since a finite sum, product and composition of Lipschitz and bounded functions is again Lipschitz and bounded by Lemma B.2.1, we obtain Lipschitz constants $K_{\Psi,t,s}$ of the maps $\pi \to \mu_t(s)$ for any $t \in \mathcal{T}, s \in \mathcal{S}$ and define $K_\Psi \equiv \max_{t\in\mathcal{T}, s\in\mathcal{S}} K_{\Psi,t,s}$, which is the desired Lipschitz constant of Ψ. ∎

Finally, the map from an energy function to its associated Boltzmann distribution is Lipschitz for any $\eta > 0$ with a Lipschitz constant explicitly depending on η.

Lemma B.7.5. *Let $\eta > 0$ arbitrary and $f_a : \mathcal{M} \to \mathbb{R}$ be a Lipschitz continuous function with Lipschitz constant K_f for any $a \in \mathcal{A}$. Further, let $g : \mathcal{A} \to \mathbb{R}$ be bounded by $g_{\max} > g(a) > g_{\min} > 0$ for any $a \in \mathcal{A}$. The function*

$$\mu \mapsto \frac{g(a) \exp\left(\frac{f_a(\mu)}{\eta}\right)}{\sum_{a'\in\mathcal{A}} g(a') \exp\left(\frac{f_{a'}(\mu)}{\eta}\right)}$$

is Lipschitz with Lipschitz constant $K = \frac{(|\mathcal{A}|-1) K_f g^2_{\max}}{2\eta g^2_{\min}}$ for any $a \in \mathcal{A}$.

Proof. Let $\mu, \mu' \in \mathcal{M}$ be arbitrary and define

$$\Delta_a f_{a'}(\mu) \equiv f_{a'}(\mu) - f_a(\mu)$$

for any $a' \in \mathcal{A}$, which is Lipschitz with constant $2K_f$. Then, we have

$$\begin{aligned}
&\left| \frac{g(a)\exp\left(\frac{f_a(\mu)}{\eta}\right)}{\sum_{a'\in\mathcal{A}} g(a')\exp\left(\frac{f_{a'}(\mu)}{\eta}\right)} - \frac{g(a)\exp\left(\frac{f_a(\mu')}{\eta}\right)}{\sum_{a'\in\mathcal{A}} g(a')\exp\left(\frac{f_{a'}(\mu')}{\eta}\right)} \right| \\
&= \left| \frac{1}{1+\sum_{a'\neq a}\frac{g(a')}{g(a)}\exp\left(\frac{\Delta_a f_{a'}(\mu)}{\eta}\right)} - \frac{1}{1+\sum_{a'\neq a}\frac{g(a')}{g(a)}\exp\left(\frac{\Delta_a f_{a'}(\mu')}{\eta}\right)} \right| \\
&\leq \left| \sum_{a'\neq a} \frac{\frac{g(a')}{g(a)}\cdot\frac{1}{\eta}\exp\left(\frac{\xi_{a'}}{\eta}\right)}{\left(1+\sum_{a''\neq a}\frac{g(a'')}{g(a)}\exp\left(\frac{\xi_{a''}}{\eta}\right)\right)^2} \cdot (\Delta_a f_{a'}(\mu) - \Delta_a f_{a'}(\mu')) \right| \\
&\leq \sum_{a'\neq a} \left| \frac{\frac{g_{\max}}{g_{\min}}\cdot\frac{1}{\eta}\exp\left(\frac{\xi_{a'}}{\eta}\right)}{\left(1+\frac{g_{\min}}{g_{\max}}\exp\left(\frac{\xi_{a'}}{\eta}\right)\right)^2} \right| \cdot |\Delta_a f_{a'}(\mu) - \Delta_a f_{a'}(\mu')| \\
&\leq \frac{g_{\max}^2}{4\eta g_{\min}^2} \cdot \sum_{a'\neq a} 2K_f d_{\mathcal{M}}(\mu,\mu') = \frac{(|\mathcal{A}|-1)K_f g_{\max}^2}{2\eta g_{\min}^2} \cdot d_{\mathcal{M}}(\mu,\mu')
\end{aligned}$$

where we applied the mean value theorem to obtain some $\xi_{a'} \in \mathbb{R}$ for all $a' \in \mathcal{A}$ and used the maximum $\frac{1}{4c}$ of the function $\tilde{f}(x) = \frac{\exp(x/\eta)}{(1+c\cdot\exp(x/\eta))^2}$ at $x = 0$. ■

For RelEnt MFE, by Lemma B.7.2 we obtain a Lipschitz constant $K_{\tilde{Q}_\eta}$ of $\mu \to \tilde{Q}_\eta(\mu,t,s,a)$ as long as $\eta > \eta_{\min}$ for some $\eta_{\min} > 0$. Furthermore, note that for $\tilde{\pi}^{\mu,\eta} \equiv \tilde{\Phi}_\eta(\mu)$, we have

$$\left|\tilde{\pi}_t^{\mu,\eta}(a \mid s) - \tilde{\pi}_t^{\mu',\eta}(a \mid s))\right| = \left| \frac{q_t(a \mid s)\exp\left(\frac{\tilde{Q}_\eta(\mu,t,s,a)}{\eta}\right)}{\sum_{a'\in\mathcal{A}} q_t(a' \mid s)\exp\left(\frac{\tilde{Q}_\eta(\mu,t,s,a')}{\eta}\right)} - \frac{q_t(a \mid s)\exp\left(\frac{\tilde{Q}_\eta(\mu',t,s,a)}{\eta}\right)}{\sum_{a'\in\mathcal{A}} q_t(a' \mid s)\exp\left(\frac{\tilde{Q}_\eta(\mu',t,s,a')}{\eta}\right)} \right|.$$

We obtain the Lipschitz constant of $\tilde{\Phi}_\eta$ by applying Lemma B.7.5 to each of the maps given by

$$\mu \mapsto \frac{q_t(a \mid s)\exp\left(\frac{\tilde{Q}_\eta(\mu,t,s,a)}{\eta}\right)}{\sum_{a'\in\mathcal{A}} q_t(a' \mid s)\exp\left(\frac{\tilde{Q}_\eta(\mu,t,s,a')}{\eta}\right)}$$

for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$, resulting in the Lipschitz property

$$\begin{aligned}
d_\Pi(\tilde{\Phi}_\eta(\mu), \tilde{\Phi}_\eta(\mu')) &= \max_{s\in\mathcal{S}} \max_{t\in\mathcal{T}} \sum_{a\in\mathcal{A}} \left|\tilde{\pi}_t^{\mu,\eta}(a \mid s) - \tilde{\pi}_t^{\mu',\eta}(a \mid s))\right| \\
&\leq \sum_{a\in\mathcal{A}} \frac{(|\mathcal{A}|-1)K_{\tilde{Q}_\eta} q_{\max}^2}{2\eta q_{\min}^2} \cdot d_{\mathcal{M}}(\mu,\mu') = \frac{|\mathcal{A}|\,(|\mathcal{A}|-1)K_{\tilde{Q}_\eta} q_{\max}^2}{2\eta q_{\min}^2} \cdot d_{\mathcal{M}}(\mu,\mu'),
\end{aligned}$$

where we define $q_{\max} = \max_{t\in\mathcal{T},s\in\mathcal{S},a\in\mathcal{A}} q_t(a \mid s)$ and analogously $q_{\min} = \min_{t\in\mathcal{T},s\in\mathcal{S},a\in\mathcal{A}} q_t(a \mid s)$.

By Lemma B.7.4, $\Psi(\pi)$ is Lipschitz with some Lipschitz constant K_Ψ. Therefore, the resulting Lipschitz constant of the composition $\tilde{\Gamma}_\eta = \Psi \circ \tilde{\Phi}_\eta$ is $\frac{|\mathcal{A}|(|\mathcal{A}|-1)K_{\tilde{Q}_\eta} K_\Psi q_{\max}^2}{2\eta q_{\min}^2}$ and leads to a contraction for any

$$\eta > \max\left(\eta_{\min}, \frac{|\mathcal{A}|\,(|\mathcal{A}|-1)K_{\tilde{Q}_\eta} K_\Psi q_{\max}^2}{2q_{\min}^2}\right).$$

Analogously for Boltzmann MFE, by Lemma B.7.3 the mapping $\mu \to Q^*(\mu,t,s,a)$ is Lipschitz with some Lipschitz constant K_{Q^*} for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$. For $\pi^{\mu,\eta} \equiv \Phi_\eta(\mu)$, we have

$$\left|\pi_t^{\mu,\eta}(a \mid s) - \pi_t^{\mu',\eta}(a \mid s))\right| = \left| \frac{q_t(a \mid s)\exp\left(\frac{Q^*(\mu,t,s,a)}{\eta}\right)}{\sum_{a'\in\mathcal{A}} q_t(a' \mid s)\exp\left(\frac{Q^*(\mu,t,s,a')}{\eta}\right)} - \frac{q_t(a \mid s)\exp\left(\frac{Q^*(\mu',t,s,a)}{\eta}\right)}{\sum_{a'\in\mathcal{A}} q_t(a' \mid s)\exp\left(\frac{Q^*(\mu',t,s,a')}{\eta}\right)} \right|.$$

We obtain the Lipschitz constant of Φ_η by applying Lemma B.7.5 to each of the maps given by

$$\mu \mapsto \frac{q_t(a \mid s) \exp\left(\frac{Q^*(\mu,t,s,a)}{\eta}\right)}{\sum_{a' \in \mathcal{A}} q_t(a' \mid s) \exp\left(\frac{Q^*(\mu,t,s,a')}{\eta}\right)}$$

for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$, resulting in the Lipschitz property

$$\begin{aligned} d_\Pi(\Phi_\eta(\mu), \Phi_\eta(\mu')) &= \max_{s \in \mathcal{S}} \max_{t \in \mathcal{T}} \sum_{a \in \mathcal{A}} \left| \pi_t^{\mu,\eta}(a \mid s) - \pi_t^{\mu',\eta}(a \mid s)) \right| \\ &\leq \sum_{a \in \mathcal{A}} \frac{(|\mathcal{A}|-1) K_{Q^*} q_{\max}^2}{2\eta q_{\min}^2} \cdot d_{\mathcal{M}}(\mu, \mu') = \frac{|\mathcal{A}|\,(|\mathcal{A}|-1) K_{Q^*} q_{\max}^2}{2\eta q_{\min}^2} \cdot d_{\mathcal{M}}(\mu, \mu')\,. \end{aligned}$$

By Lemma B.7.4, $\Psi(\pi)$ is Lipschitz with some Lipschitz constant K_Ψ. The resulting Lipschitz constant of the composition $\Gamma_\eta = \Psi \circ \Phi_\eta$ is $\frac{|\mathcal{A}|(|\mathcal{A}|-1) K_{Q^*} K_\Psi q_{\max}^2}{2\eta q_{\min}^2}$ and leads to a contraction for any

$$\eta > \frac{|\mathcal{A}|\,(|\mathcal{A}|-1) K_{Q^*} K_\Psi q_{\max}^2}{2 q_{\min}^2}$$

where for the uniform prior policy, $q_{\max} = q_{\min}$. If required, the Lipschitz constants can be computed recursively according to Lemma B.2.1. □

B.8 Proof of Theorem 4

Proof. Consider any sequence $(\pi_n^*, \mu_n^*)_{n \in \mathbb{N}}$ of η_n-Boltzmann or η_n-RelEnt MFE with $\eta_n \to 0^+$ as $n \to \infty$. Note that a pair (π_n^*, μ_n^*) is completely specified by μ_n^*, since $\pi_n^* = \Phi_{\eta_n}(\mu_n^*)$ or $\pi_n^* = \tilde{\Phi}_{\eta_n}(\mu_n^*)$ uniquely. Therefore, it suffices to show that the associated functions $(\mu \mapsto Q^{\Phi_{\eta_n}(\mu)}(\mu,t,s,a))_{n \in \mathbb{N}}$ and $(\mu \mapsto Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu,t,s,a))_{n \in \mathbb{N}}$ converge uniformly to $\mu \mapsto Q^*(\mu,t,s,a)$, from which the desired result will follow. For definitions of the different action-value functions, see Appendix B.7.

Note that pointwise convergence is insufficient, since there is no guarantee that μ_n^* itself will converge as $n \to \infty$. However, we can obtain uniform convergence by pointwise convergence and equicontinuity. For RelEnt MFE, we will additionally require uniform convergence of the sequence $(\mu \mapsto \tilde{Q}_{\eta_n}(\mu,t,s,a))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$. We begin with pointwise convergence of $(\mu \mapsto Q^{\Phi_{\eta_n}(\mu)}(\mu,t,s,a))_{n \in \mathbb{N}}$ to the optimal action-value function $\mu \mapsto Q^*(\mu,t,s,a)$.

Lemma B.8.1. *Any sequence of functions $(\mu \mapsto Q^{\Phi_{\eta_n}(\mu)}(\mu,t,s,a))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ converges pointwise to $\mu \mapsto Q^*(\mu,t,s,a)$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.*

Proof. Fix $\mu \in \mathcal{M}$. We make the induction hypothesis for arbitrary $t \in \mathcal{T}$ that for all $s \in \mathcal{S}, a \in \mathcal{A}, \varepsilon > 0$, there exists $n' \in \mathbb{N}$ such that for any $n > n'$ we have

$$\left| Q^{\Phi_{\eta_n}(\mu)}(\mu,t,s,a) - Q^*(\mu,t,s,a) \right| < \varepsilon\,.$$

The induction hypothesis is fulfilled for $t = T-1$, as by definition

$$\left| Q^{\Phi_{\eta_n}(\mu)}(\mu,t,s,a) - Q^*(\mu,t,s,a) \right| = |r(s,a,\mu_t) - r(s,a,\mu_t)| = 0\,.$$

Assume that the induction hypothesis is fulfilled for $t+1$, then at time t let $s \in \mathcal{S}, a \in \mathcal{A}, \varepsilon > 0$ arbitrary. Furthermore, let $s' \in \mathcal{S}$ arbitrary. Collect all optimal actions into a set $\mathcal{A}_{\text{opt}}^{s'} \subseteq \mathcal{A}$, i.e. for $a' \in \mathcal{A}_{\text{opt}}^{s'}$ we have

$$Q^*(\mu,t,s',a_{\text{opt}}) = \max_{a \in \mathcal{A}} Q^*(\mu,t,s',a)\,.$$

We define the minimal action gap

$$\Delta Q_{\min}^{s',\mu} \equiv \min_{a_{\text{opt}} \in \mathcal{A}_{\text{opt}}^{s'}, a_{\text{sub}} \in \mathcal{A} \setminus \mathcal{A}_{\text{opt}}^{s'}} (Q^*(\mu,t,s',a_{\text{opt}}) - Q^*(\mu,t,s',a_{\text{sub}})) > 0$$

such that for arbitrary suboptimal actions $a_{\text{sub}} \in \mathcal{A} \setminus \mathcal{A}_{\text{opt}}^{s'}$ and optimal actions $a_{\text{opt}} \in \mathcal{A}_{\text{opt}}^{s'}$,

$$Q^*(\mu, t, s', a_{\text{opt}}) - Q^*(\mu, t, s', a_{\text{sub}}) \geq \Delta Q_{\min}^{s',\mu}.$$

This is well defined if there are suboptimal actions, since there is always at least one optimal action. If all actions are optimal, we can skip bounding the probability of taking suboptimal actions and the result will hold trivially. Thus, we assume henceforth that there exists a suboptimal action.

It follows that the probability of taking suboptimal actions $a_{\text{sub}} \in \mathcal{A} \setminus \mathcal{A}_{\text{opt}}^{s'}$ disappears, since

$$\begin{aligned}
(\Phi_{\eta_n}(\mu))_t(a_{\text{sub}} \mid s') &= \frac{q_t(a_{\text{sub}} \mid s)}{\sum_{a' \in \mathcal{A}} q_t(a' \mid s) \exp\left(\frac{Q^*(\mu,t,s,a') - Q^*(\mu,t,s,a_{\text{sub}})}{\eta}\right)} \\
&\leq \frac{1}{1 + \sum_{a' \in \mathcal{A}} \frac{q_t(a'|s)}{q_t(a_{\text{sub}}|s)} \exp\left(\frac{Q^*(\mu,t,s,a') - Q^*(\mu,t,s,a_{\text{sub}})}{\eta}\right)} \\
&\leq \frac{1 \mid s)}{1 + \frac{q_t(a_{\text{opt}}|s)}{q_t(a_{\text{sub}}|s)} \exp\left(\frac{Q^*(\mu,t,s,a_{\text{opt}}) - Q^*(\mu,t,s,a_{\text{sub}})}{\eta}\right)} \\
&\leq \frac{1 \mid s)}{1 + \frac{q_t(a_{\text{opt}}|s)}{q_t(a_{\text{sub}}|s)} \exp\left(\frac{\Delta Q_{\min}^{s',\mu}}{\eta}\right)} \to 0
\end{aligned}$$

as $\eta \to 0^+$ for some arbitrary optimal action $a_{\text{opt}} \in \mathcal{A}_{\text{opt}}^{s'}$. Since $s' \in \mathcal{S}$ was arbitrary, this holds for all $s' \in \mathcal{S}$. Therefore, by finiteness of $\mathcal{S}$ and $\mathcal{A}$ we can choose $n_1 \in \mathbb{N}$ such that for all $n > n_1$ and for all $a_{\text{sub}} \in \mathcal{A} \setminus \mathcal{A}_{\text{opt}}^{s'}$ we have η_n sufficiently small such that

$$(\Phi_{\eta_n}(\mu))_t(a_{\text{sub}} \mid s') < \frac{\varepsilon}{2|\mathcal{A}|M_Q}$$

where M_Q is the uniform bound of $Q^{\Phi_{\eta_n}(\mu)}$.

Further, by induction assumption, we can choose $n_{s',a'}$ for any $s' \in \mathcal{S}, a' \in \mathcal{A}$ such that for all $n > n_{s',a'}$ we have

$$\left|Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, s', a') - Q^*(\mu, t+1, s', a')\right| < \frac{\varepsilon}{3}$$

Therefore, as long as $n > n' \equiv \max(n_1, \max_{s' \in \mathcal{S}, a' \in \mathcal{A}} n_{s',a'})$, we have

$$\begin{aligned}
&\left|Q^{\Phi_{\eta_n}(\mu)}(\mu, t, s, a) - Q^*(\mu, t, s, a)\right| \\
&= \left|\sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \left(\sum_{a' \in \mathcal{A}} (\Phi_{\eta_n}(\mu))_t(a' \mid s') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, s', a') - \max_{a'' \in \mathcal{A}} Q^*(\mu, t+1, s', a'')\right)\right| \\
&\leq \max_{s' \in \mathcal{S}} \left|\sum_{a' \in \mathcal{A}} (\Phi_{\eta_n}(\mu))_t(a' \mid s') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, s', a') - \max_{a'' \in \mathcal{A}} Q^*(\mu, t+1, s', a'')\right| \\
&\leq \max_{s' \in \mathcal{S}} \left|\sum_{a' \in \mathcal{A}_{\text{opt}}^{s'}} (\Phi_{\eta_n}(\mu))_t(a' \mid s') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, s', a') - \max_{a'' \in \mathcal{A}} Q^*(\mu, t+1, s', a'')\right| \\
&\quad + \max_{s' \in \mathcal{S}} \left|\sum_{a' \in \mathcal{A} \setminus \mathcal{A}_{\text{opt}}^{s'}} (\Phi_{\eta_n}(\mu))_t(a' \mid s') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, s', a')\right| \\
&\leq \max_{s' \in \mathcal{S}} \left|\sum_{a' \in \mathcal{A}_{\text{opt}}^{s'}} (\Phi_{\eta_n}(\mu))_t(a' \mid s') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, s', a') - \sum_{a' \in \mathcal{A}_{\text{opt}}^{s'}} (\Phi_{\eta_n}(\mu))_t(a' \mid s') \max_{a'' \in \mathcal{A}} Q^*(\mu, t+1, s', a'')\right|
\end{aligned}$$

$$
\begin{aligned}
&+ \max_{s' \in \mathcal{S}} \left| \sum_{a' \in \mathcal{A}^{s'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(a' \mid s') \max_{a'' \in \mathcal{A}} Q^*(\mu, t+1, s', a'') - \max_{a'' \in \mathcal{A}} Q^*(\mu, t+1, s', a'') \right| \\
&+ \max_{s' \in \mathcal{S}} \left| \sum_{a' \in \mathcal{A} \setminus \mathcal{A}^{s'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(a' \mid s') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, s', a') \right| \\
&\leq \max_{s' \in \mathcal{S}} \max_{a' \in \mathcal{A}^{s'}_{\mathrm{opt}}} \left| Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, s', a') - \max_{a'' \in \mathcal{A}} Q^*(\mu, t+1, s', a'') \right| \\
&\quad + \max_{s' \in \mathcal{S}} M_Q \left| - \sum_{a' \in \mathcal{A} \setminus \mathcal{A}^{s'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(a' \mid s') \right| + \max_{s' \in \mathcal{S}} M_Q \left| \sum_{a' \in \mathcal{A} \setminus \mathcal{A}^{s'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(a' \mid s') \right| \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3|\mathcal{A}|M_Q} \cdot |\mathcal{A}| M_Q + \frac{\varepsilon}{3|\mathcal{A}|M_Q} \cdot |\mathcal{A}| M_Q = \varepsilon \,.
\end{aligned}
$$

Since $s \in \mathcal{S}, a \in \mathcal{A}, \varepsilon > 0$ were arbitrary, the desired result follows immediately by induction. ■

As we have no control over μ_n^* and the sequence $(\pi_n^*, \mu_n^*)_{n \in \mathbb{N}}$ may not even converge, pointwise convergence is insufficient. To obtain uniform convergence, we shall use compactness of $\mathcal{M}$ and equicontinuity.

Lemma B.8.2. *The family of functions $\mathcal{F} \equiv \{\mu \mapsto Q^{\Phi_\eta(\mu)}(\mu, t, s, a)\}_{\eta > 0, t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}}$ is equicontinuous, i.e. for any $\varepsilon > 0$ and any $\mu \in \mathcal{M}$, we can choose a $\delta > 0$ such that for all $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu, \mu') < \delta$ and any $f \in \mathcal{F}$ we have*

$$
|f(\mu) - f(\mu')| < \varepsilon \,.
$$

Proof. Fix an arbitrary $\mu \in \mathcal{M}$. We make the (backwards in time) induction hypothesis for all $t \in \mathcal{T}$ that for any $s \in \mathcal{S}, a \in \mathcal{A}, \varepsilon_{t,s,a} > 0$, there exists $\delta_{t,s,a} > 0$ such that for any $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu, \mu') < \delta_{t,s,a}$ and any $f \in \mathcal{F}$ we have

$$
\left| Q^{\Phi_\eta(\mu)}(\mu, t, s, a) - Q^{\Phi_\eta(\mu')}(\mu', t, s, a) \right| < \varepsilon_{t,s,a} \,.
$$

The induction hypothesis is fulfilled for $t = T - 1$, as by assumption, $\nu \to r(s, a, \nu_t)$ is Lipschitz with constant $C_r > 0$. Therefore, for all $s \in \mathcal{S}, a \in \mathcal{A}$ we can choose $\delta_{T-1,s,a} = \frac{\varepsilon_{t,s,a}}{C_r}$ such that for any μ, μ' with $d_{\mathcal{M}}(\mu, \mu') < \delta'$ we have

$$
\left| Q^{\Phi_\eta(\mu)}(\mu, t, s, a) - Q^{\Phi_\eta(\mu')}(\mu', t, s, a) \right| = |r(s, a, \mu_t) - r(s, a, \mu'_t)| \leq C_r d_{\mathcal{M}}(\mu, \mu') < \varepsilon_{t,s,a} \,.
$$

Assume that the induction hypothesis holds for $t+1$, then at time t let $\varepsilon_{t,s,a} > 0, s \in \mathcal{S}, a \in \mathcal{A}$ arbitrary. By definition, we have

$$
\begin{aligned}
&\left| Q^{\Phi_\eta(\mu)}(\mu, t, s, a) - Q^{\Phi_\eta(\mu')}(\mu', t, s, a) \right| \\
&= \Bigg| r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \sum_{a' \in \mathcal{A}} (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu, t+1, s', a') \\
&\quad - r(s, a, \mu'_t) - \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu'_t) \sum_{a' \in \mathcal{A}} (\Phi_\eta(\mu'))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu', t+1, s', a') \Bigg| \\
&\leq |r(s, a, \mu_t) - r(s, a, \mu'_t)| \\
&\quad + \sum_{s' \in \mathcal{S}} \left| (p(s' \mid s, a, \mu_t) - p(s' \mid s, a, \mu'_t)) \sum_{a' \in \mathcal{A}} (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu, t+1, s', a') \right| \\
&\quad + \sum_{s' \in \mathcal{S}} \left| p(s' \mid s, a, \mu'_t) \sum_{a' \in \mathcal{A}} \Big((\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu, t+1, s', a') - (\Phi_\eta(\mu'))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu', t+1, s', a') \Big) \right|
\end{aligned}
$$

$$
\begin{aligned}
&\leq |r(s,a,\mu_t) - r(s,a,\mu_t')| \\
&\quad + \sum_{s'\in\mathcal{S}} \left| (p(s' \mid s,a,\mu_t) - p(s' \mid s,a,\mu_t')) \sum_{a'\in\mathcal{A}} (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu, t+1, s', a') \right| \\
&\quad + \max_{s'\in\mathcal{S}} \left| \sum_{a'\in\mathcal{A}^{s'}_{\text{opt}}} \Big((\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu, t+1, s', a') - (\Phi_\eta(\mu'))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu', t+1, s', a') \Big) \right| \\
&\quad + \max_{s'\in\mathcal{S}} \left| \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^{s'}_{\text{opt}}} \Big((\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu, t+1, s', a') - (\Phi_\eta(\mu'))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu', t+1, s', a') \Big) \right|
\end{aligned}
$$

where we define $\mathcal{A}^{s'}_{\text{opt}} \subseteq \mathcal{A}$ for any $s' \in \mathcal{S}$ to include all optimal actions $a_{\text{opt}} \in \mathcal{A}^{s'}_{\text{opt}}$ such that

$$
Q^*(\mu, t, s', a_{\text{opt}}) = \max_{a\in\mathcal{A}} Q^*(\mu, t, s', a)\,.
$$

We bound each of the four terms separately.

For the first term, we choose $\delta^1_{t,s,a} = \frac{\varepsilon_{t,s,a}}{4C_r}$ by Lipschitz continuity such that

$$
|r(s,a,\mu_t) - r(s,a,\mu_t')| < \frac{\varepsilon_{t,s,a}}{4}
$$

for all μ' with $d_{\mathcal{M}}(\mu,\mu') < \delta^1_{t,s,a}$.

For the second term, we choose $\delta^2_{t,s,a} = \frac{1}{4|\mathcal{S}|M_Q C_p}$ such that for any $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu,\mu') < \delta^2_{t,s,a}$ we have

$$
\begin{aligned}
&\sum_{s'\in\mathcal{S}} \left| (p(s' \mid s,a,\mu_t) - p(s' \mid s,a,\mu_t')) \sum_{a'\in\mathcal{A}} (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu, t+1, s', a') \right| \\
&\leq |\mathcal{S}| C_p d_{\mathcal{M}}(\mu,\mu') M_Q < \frac{\varepsilon_{t,s,a}}{4}
\end{aligned}
$$

where M_Q denotes the uniform bound of Q and C_p is the Lipschitz constant of $\nu \mapsto p(s' \mid s,a,\nu_t)$.

For the third and fourth term, we first fix $s' \in \mathcal{S}$ and define the minimal action gap as

$$
\Delta Q^{s',\mu}_{\min} \equiv \min_{a_{\text{opt}}\in\mathcal{A}^{s'}_{\text{opt}}, a_{\text{sub}}\in\mathcal{A}\setminus\mathcal{A}^{s'}_{\text{opt}}} \left(Q^*(\mu,t,s',a_{\text{opt}}) - Q^*(\mu,t,s',a_{\text{sub}}) \right)\,.
$$

This is well defined if there are suboptimal actions, since there is always at least one optimal action. If all actions are optimal, we can skip bounding the probability of taking suboptimal actions and the result will still hold. Henceforth, we assume that there exists a suboptimal action.

By Lipschitz continuity of $\mu \mapsto Q^*(\mu,t,s,a)$ from Lemma B.7.3 implying uniform continuity, there exists some $\delta^{3,s'}_{t,s,a} > 0$ such that

$$
|Q^*(\mu',t,s',a) - Q^*(\mu,t,s',a)| < \frac{\Delta Q^{s',\mu}_{\min}}{4}
$$

for all $\mu' \in \mathcal{M}, a \in \mathcal{A}$ where $d_{\mathcal{M}}(\mu,\mu') < \delta^{3,s'}_{t,s,a}$, and thus

$$
\Delta Q^{s',\mu'}_{\min} = \min_{a_{\text{opt}}\in\mathcal{A}^{s'}_{\text{opt}}, a_{\text{sub}}\in\mathcal{A}\setminus\mathcal{A}^{s'}_{\text{opt}}} \left(Q^*(\mu',t,s',a_{\text{opt}}) - Q^*(\mu',t,s',a_{\text{sub}}) \right) > \frac{\Delta Q^{s',\mu}_{\min}}{2}\,.
$$

Under this condition, we can now show that the probability of any suboptimal action can be controlled. Define $R^{\min}_q \equiv \min_{t\in\mathcal{T}, s\in\mathcal{S}, a\in\mathcal{A}, a'\in\mathcal{A}} \frac{q_t(a'|s)}{q_t(a|s)} > 0$ and $R^{\max}_q \equiv \max_{t\in\mathcal{T}, s\in\mathcal{S}, a\in\mathcal{A}, a'\in\mathcal{A}} \frac{q_t(a'|s)}{q_t(a|s)} > 0$. Let $a_{\text{sub}} \in \mathcal{A} \setminus \mathcal{A}^{s'}_{\text{opt}}$, then we either have

$$
|(\Phi_\eta(\mu))_{t+1}(a_{\text{sub}} \mid s') - (\Phi_\eta(\mu'))_{t+1}(a_{\text{sub}} \mid s')|
$$

$$
\begin{aligned}
&= \left| \frac{1}{1 + \sum_{a' \neq a_{\text{sub}}} \frac{q_t(a'|s')}{q_t(a_{\text{sub}}|s')} \exp\left(\frac{Q^*(\mu,t,s',a') - Q^*(\mu,t,s',a_{\text{sub}})}{\eta}\right)} \right. \\
&\quad \left. - \frac{1}{1 + \sum_{a' \neq a_{\text{sub}}} \frac{q_t(a'|s')}{q_t(a_{\text{sub}}|s')} \exp\left(\frac{Q^*(\mu',t,s',a') - Q^*(\mu',t,s',a_{\text{sub}})}{\eta}\right)} \right| \\
&\leq \frac{1}{1 + \max_{a' \neq a_{\text{sub}}} R_q^{\min} \exp\left(\frac{Q^*(\mu,t,s',a') - Q^*(\mu,t,s',a_{\text{sub}})}{\eta}\right)} \\
&\quad + \frac{1}{1 + \max_{a' \neq a_{\text{sub}}} R_q^{\min} \exp\left(\frac{Q^*(\mu',t,s',a') - Q^*(\mu',t,s',a_{\text{sub}})}{\eta}\right)} \\
&< \frac{1}{1 + R_q^{\min} \exp\left(\frac{\Delta Q_{\min}^{s',\mu}}{\eta}\right)} + \frac{1}{1 + R_q^{\min} \exp\left(\frac{\Delta Q_{\min}^{s',\mu}}{2\eta}\right)} \\
&\leq \frac{2}{1 + R_q^{\min} \exp\left(\frac{\Delta Q_{\min}^{s',\mu}}{2\eta}\right)} < \frac{\varepsilon_{t,s,a}}{8M_Q|\mathcal{A}|}
\end{aligned}
$$

if $\varepsilon_{t,s,a} > 16M_Q|\mathcal{A}|$ trivially, or otherwise if $\eta < \eta_{\min}^{s'}$ with

$$
\eta_{\min}^{s'} \equiv \frac{\Delta Q_{\min}^{s',\mu}}{2\log\left(\frac{16M_Q|\mathcal{A}|}{\varepsilon_{t,s,a} R_q^{\min}} - \frac{1}{R_q^{\min}}\right)},
$$

in which case we arbitrarily define $\delta_{t,s,a}^{4,s'} = 1$, or if neither apply, then $\eta \geq \eta_{\min}^{s'}$ and thus

$$
\begin{aligned}
&|(\Phi_\eta(\mu))_{t+1}(a_{\text{sub}} \mid s') - (\Phi_\eta(\mu'))_{t+1}(a_{\text{sub}} \mid s')| \\
&= \left| \frac{1}{1 + \sum_{a' \neq a_{\text{sub}}} \frac{q_t(a'|s')}{q_t(a_{\text{sub}}|s')} \exp\left(\frac{Q^*(\mu,t,s',a') - Q^*(\mu,t,s',a_{\text{sub}})}{\eta}\right)} \right. \\
&\quad \left. - \frac{1}{1 + \sum_{a' \neq a_{\text{sub}}} \frac{q_t(a'|s')}{q_t(a_{\text{sub}}|s')} \exp\left(\frac{Q^*(\mu',t,s',a') - Q^*(\mu',t,s',a_{\text{sub}})}{\eta}\right)} \right| \\
&= \left| \frac{\sum_{a' \neq a_{\text{sub}}} \frac{q_t(a'|s)}{q_t(a_{\text{sub}}|s')} \left(\exp\left(\frac{Q^*(\mu',t,s',a') - Q^*(\mu',t,s',a_{\text{sub}})}{\eta}\right) - \exp\left(\frac{Q^*(\mu,t,s',a') - Q^*(\mu,t,s',a_{\text{sub}})}{\eta}\right)\right)}{(1 + \cdots) \cdot (1 + \cdots)} \right| \\
&\leq R_q^{\max} \sum_{a' \neq a_{\text{sub}}} \left| \exp\left(\frac{Q^*(\mu',t,s',a') - Q^*(\mu',t,s',a_{\text{sub}})}{\eta}\right) - \exp\left(\frac{Q^*(\mu,t,s',a') - Q^*(\mu,t,s',a_{\text{sub}})}{\eta}\right) \right| \\
&\leq R_q^{\max} \sum_{a' \neq a_{\text{sub}}} \left| \frac{1}{\eta} \exp\left(\frac{\xi_{a'}}{\eta}\right) \right| |(Q^*(\mu',t,s',a') - Q^*(\mu',t,s',a_{\text{sub}})) - (Q^*(\mu,t,s',a') - Q^*(\mu,t,s',a_{\text{sub}}))| \\
&\leq R_q^{\max}|\mathcal{A}| \cdot \frac{1}{\eta_{\min}^{s'}} \exp\left(\frac{2M_Q}{\eta_{\min}^{s'}}\right) (|Q^*(\mu',t,s',a') - Q^*(\mu,t,s',a')| + |Q^*(\mu,t,s',a_{\text{sub}}) - Q^*(\mu',t,s',a_{\text{sub}})|) \\
&\leq R_q^{\max}|\mathcal{A}| \cdot \frac{1}{\eta_{\min}^{s'}} \exp\left(\frac{2M_Q}{\eta_{\min}^{s'}}\right) \cdot 2K_Q d_{\mathcal{M}}(\mu,\mu') < \frac{\varepsilon_{t,s,a}}{8M_Q|\mathcal{A}|}
\end{aligned}
$$

by the mean value theorem with some $\xi_{a'} \in [-2M_Q, 2M_Q]$ for all $a' \in \mathcal{A}$, where we abbreviated the denominator $(1 + \cdots) \cdot (1 + \cdots) \geq 1$, as long as we choose

$$
\delta_{t,s,a}^{4,s'} = \frac{\varepsilon_{t,s,a} \eta_{\min}^{s'}}{8M_Q|\mathcal{A}|^2 R_q^{\max} \cdot \exp\left(\frac{2M_Q}{\eta_{\min}^{s'}}\right) \cdot 2K_Q}
$$

and $d_{\mathcal{M}}(\mu,\mu') < \delta^{4,s'}_{t,s,a}$, where K_Q is the Lipschitz constant of $\mu \mapsto Q^*(\mu,t,s,a)$ given by Lemma B.7.3.

Since $s' \in \mathcal{S}$ was arbitrary, we now define $\delta^3_{t,s,a} \equiv \min_{s'\in\mathcal{S}} \delta^{3,s'}_{t,s,a}$, $\delta^4_{t,s,a} \equiv \min_{s'\in\mathcal{S}} \delta^{4,s'}_{t,s,a}$ and let $d_{\mathcal{M}}(\mu,\mu') < \min(\delta^3_{t,s,a}, \delta^4_{t,s,a})$. Under these assumptions, for the third term we have approximate optimality for all optimal actions in $\mathcal{A}^{s'}_{\text{opt}}$, since by induction assumption we can choose $\delta_{t+1,s',a'}$ for all $s' \in \mathcal{S}, a' \in \mathcal{A}$ such that for all $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu,\mu') < \delta_{t+1,s',a'}$ it holds that

$$\left|Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') - Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a')\right| < \frac{\varepsilon_{t,s,a}}{16|\mathcal{A}|+8}\,.$$

and therefore for all $\mu' \in \mathcal{M}$, as long as $d_{\mathcal{M}}(\mu,\mu') < \min_{s'\in\mathcal{S},a'\in\mathcal{A}} \delta_{t+1,s',a'}$, we have

$$\begin{aligned}
&\max_{s'\in\mathcal{S}} \left| \sum_{a'\in\mathcal{A}^{s'}_{\text{opt}}} (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') - \sum_{a'\in\mathcal{A}^{s'}_{\text{opt}}} (\Phi_\eta(\mu'))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a') \right| \\
&\leq \max_{s'\in\mathcal{S}} \left| \sum_{a'\in\mathcal{A}^{s'}_{\text{opt}}} (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') - \sum_{a'\in\mathcal{A}^{s'}_{\text{opt}}} (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a') \right| \\
&\quad + \max_{s'\in\mathcal{S}} \left| \sum_{a'\in\mathcal{A}^{s'}_{\text{opt}}} (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a') - \sum_{a'\in\mathcal{A}^{s'}_{\text{opt}}} (\Phi_\eta(\mu'))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a') \right| \\
&\leq \max_{s'\in\mathcal{S}} \max_{a'\in\mathcal{A}} \left| Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') - Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a') \right| \\
&\quad + \max_{s'\in\mathcal{S}} \left| \sum_{a'\in\mathcal{A}^{s'}_{\text{opt}}} ((\Phi_\eta(\mu))_{t+1}(a' \mid s') - (\Phi_\eta(\mu'))_{t+1}(a' \mid s')) \left(Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a') - Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') \right) \right| \\
&\quad + \max_{s'\in\mathcal{S}} \left| \sum_{a'\in\mathcal{A}^{s'}_{\text{opt}}} ((\Phi_\eta(\mu))_{t+1}(a' \mid s') - (\Phi_\eta(\mu'))_{t+1}(a' \mid s')) Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') \right| \\
&\leq \max_{s'\in\mathcal{S}} \max_{a'\in\mathcal{A}} \left| Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') - Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a') \right| \\
&\quad + \max_{s'\in\mathcal{S}} \max_{a'\in\mathcal{A}} 2|\mathcal{A}| \left| Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a') - Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') \right| \\
&\quad + \max_{s'\in\mathcal{S}} \max_{a''\in\mathcal{A}} \left| Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a'') \right| \cdot \left| \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^{s'}_{\text{opt}}} ((\Phi_\eta(\mu'))_{t+1}(a' \mid s') - (\Phi_\eta(\mu))_{t+1}(a' \mid s')) \right| \\
&< (1+2|\mathcal{A}|) \cdot \frac{\varepsilon_{t,s,a}}{16|\mathcal{A}|+8} + M_Q|\mathcal{A}| \cdot \frac{\varepsilon_{t,s,a}}{8M_Q|\mathcal{A}|} < \frac{\varepsilon_{t,s,a}}{4}
\end{aligned}$$

where we use that for any $a' \in \mathcal{A}^{s'}_{\text{opt}}$ we have

$$Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') = \max_{a''\in\mathcal{A}} Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a'')\,.$$

Analogously, for the fourth term we have

$$\begin{aligned}
&\max_{s'\in\mathcal{S}} \left| \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^{s'}_{\text{opt}}} ((\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') - (\Phi_\eta(\mu'))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a')) \right| \\
&\leq \max_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^{s'}_{\text{opt}}} \left| (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu)}(\mu,t+1,s',a') - (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu',t+1,s',a') \right|
\end{aligned}$$

$$
\begin{aligned}
&+ \max_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A} \setminus \mathcal{A}^{s'}_{\text{opt}}} \left| (\Phi_\eta(\mu))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu', t+1, s', a') - (\Phi_\eta(\mu'))_{t+1}(a' \mid s') Q^{\Phi_\eta(\mu')}(\mu', t+1, s', a') \right| \\
&\leq \max_{s' \in \mathcal{S}} \max_{a' \in \mathcal{A}} \left| Q^{\Phi_\eta(\mu)}(\mu, t+1, s', a') - Q^{\Phi_\eta(\mu')}(\mu', t+1, s', a') \right| \\
&\quad + \max_{s' \in \mathcal{S}} M_Q \sum_{a' \in \mathcal{A} \setminus \mathcal{A}^{s'}_{\text{opt}}} |(\Phi_\eta(\mu))_{t+1}(a' \mid s') - (\Phi_\eta(\mu'))_{t+1}(a' \mid s')| \\
&< \frac{\varepsilon_{t,s,a}}{8} + M_Q |\mathcal{A}| \cdot \frac{\varepsilon_{t,s,a}}{8 M_Q |\mathcal{A}|} = \frac{\varepsilon_{t,s,a}}{4}
\end{aligned}
$$

under the previous conditions, since as long as we have $d_{\mathcal{M}}(\mu, \mu') < \delta_{t+1,s',a'}$ for all $s' \in \mathcal{S}, a' \in \mathcal{A}$ from before, we have

$$
\left| Q^{\Phi_\eta(\mu)}(\mu, t+1, s', a') - Q^{\Phi_\eta(\mu')}(\mu', t+1, s', a') \right| < \frac{\varepsilon_{t,s,a}}{16|\mathcal{A}| + 8} < \frac{\varepsilon_{t,s,a}}{8} \,.
$$

Finally, by choosing $\delta_{t,s,a}$ such that all conditions are fulfilled, i.e.

$$
\delta_{t,s,a} \equiv \min \left(\delta^1_{t,s,a}, \delta^2_{t,s,a}, \delta^3_{t,s,a}, \delta^4_{t,s,a}, \min_{s' \in \mathcal{S}, a' \in \mathcal{A}} \delta_{t+1,s',a'} \right) > 0 \,,
$$

the induction hypothesis is fulfilled, since then for any μ' with $d_{\mathcal{M}}(\mu, \mu') < \delta_{t,s,a}$ we have

$$
\left| Q^{\Phi_\eta(\mu)}(\mu, t, s, a) - Q^{\Phi_\eta(\mu')}(\mu', t, s, a) \right| < \varepsilon_{t,s,a} \,.
$$

Since $\eta > 0$ is arbitrary, the desired result follows immediately, as we can set $\varepsilon_{t,s,a} = \varepsilon$ for each $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ and obtain $\delta \equiv \max_{t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}} \delta_{t,s,a}$, fulfilling the required equicontinuity property at μ. ■

From equicontinuity, we get the desired uniform convergence via compactness.

Lemma B.8.3. *If $(f_n)_{n \in \mathbb{N}}$ with $f_n : \mathcal{M} \to \mathbb{R}$ is an equicontinuous sequence of functions and for all $\mu \in \mathcal{M}$ we have $f_n(\mu) \to f(\mu)$ pointwise, then $f_n(\mu) \to f(\mu)$ uniformly.*

Proof. Let $\varepsilon > 0$ arbitrary, then there exists by equicontinuity for any point $\mu \in \mathcal{M}$ a $\delta(\mu)$ such that for all $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu, \mu') < \delta(\mu)$ we have for all $n \in \mathbb{N}$

$$
|f_n(\mu) - f_n(\mu')| < \frac{\varepsilon}{3}
$$

which via pointwise convergence implies

$$
|f(\mu) - f(\mu')| \leq \frac{\varepsilon}{3} \,.
$$

Since $\mathcal{M}$ is compact, it is separable, i.e. there exists a countable dense subset $(\mu_j)_{j \in \mathbb{N}}$ of $\mathcal{M}$. Let $\delta(\mu)$ be as defined above and cover $\mathcal{M}$ by the open balls $(B_{\delta(\mu_j)}(\mu_j))_{j \in \mathbb{N}}$. By the compactness of $\mathcal{M}$, finitely many of these balls $B_{\delta(\mu_{n_1})}(\mu_{n_1}), \ldots, B_{\delta(\mu_{n_k})}(\mu_{n_k})$ cover $\mathcal{M}$. By pointwise convergence, for any $i = 1, \ldots, k$ we can find an integer n_i such that for all $n > n_i$ we have

$$
|f_n(\mu_{n_i}) - f(\mu_{n_i})| < \frac{\varepsilon}{3} \,.
$$

Taken together, we find that for $n > \max_{i=1,\ldots,k} n_i$ and arbitrary $\mu \in \mathcal{M}$, we have

$$
|f_n(\mu) - f(\mu)| < |f_n(\mu) - f_n(\mu_{n_i})| + |f_n(\mu_{n_i}) - f(\mu_{n_i})| + |f(\mu_{n_i}) - f(\mu)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} < \varepsilon
$$

for some center point μ_{n_i} of a ball containing μ from the finite cover. ■

Therefore, a sequence of Boltzmann MFE with vanishing η is approximately optimal in the MFG.

Lemma B.8.4. *For any sequence $(\pi_n^*, \mu_n^*)_{n\in\mathbb{N}}$ of η_n-Boltzmann MFE with $\eta_n \to 0^+$ and for any $\varepsilon > 0$ there exists integer $N \in \mathbb{N}$ such that for all integers $n > N$ we have*

$$J^{\mu_n^*}(\pi_n^*) \geq \max_{\pi} J^{\mu_n^*}(\pi) - \varepsilon \ .$$

Proof. By Lemma B.8.2, $\mathcal{F} \equiv (\mu \mapsto Q^{\Phi_\eta(\mu)}(\mu, t, s, a))_{\eta>0, t\in\mathcal{T}, s\in\mathcal{S}, a\in\mathcal{A}}$ is equicontinuous. Therefore, any sequence $(\mu \mapsto Q^{\Phi_{\eta_n}(\mu)}(\mu, t, s, a))_{n\in\mathbb{N}}$ with $\eta_n \to 0^+$ is also equicontinuous for any $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.

Furthermore, by Lemma B.8.1, the sequence $(\mu \mapsto Q^{\Phi_{\eta_n}(\mu)}(\mu, t, s, a))_{n\in\mathbb{N}}$ converges pointwise to $\mu \to Q^*(\mu, t, s, a)$ for any $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.

By Lemma B.8.3, we thus have $\left|Q^{\Phi_{\eta_n}(\mu)}(\mu, t, s, a) - Q^*(\mu, t, s, a)\right| \to 0$ uniformly. Therefore, for any $\varepsilon > 0$, there exists an integer N by uniform convergence such that for all integers $n > N$ we have

$$Q^{\pi_n^*}(\mu_n^*, t, s, a) \geq Q^*(\mu_n^*, t, s, a) - \varepsilon = \max_{\pi\in\Pi} Q^{\pi}(\mu_n^*, t, s, a) - \varepsilon \,,$$

and since by Lemma B.3.1 we have

$$J^{\mu_n^*}(\pi_n^*) = \sum_{s\in\mathcal{S}} \mu_0(s) \cdot \sum_{a\in\mathcal{A}} Q^{\pi_n^*}(\mu_n^*, t, s, a) \geq \sum_{s\in\mathcal{S}} \mu_0(s) \cdot \max_{\pi\in\Pi} \sum_{a\in\mathcal{A}} Q^{\pi}(\mu_n^*, t, s, a) - \varepsilon = \max_{\pi\in\Pi} J^{\mu_n^*}(\pi) - \varepsilon \,,$$

the desired result follows immediately. ∎

Finally, we show approximate optimality in the actual N-agent game as long as a pair $(\pi^*, \mu^*) \in \Pi \times \mathcal{M}$ with $\mu^* = \Psi(\pi^*)$ has vanishing exploitability in the MFG. By Lemma B.8.4, for any sequence $(\pi_n^*, \mu_n^*)_{n\in\mathbb{N}}$ of η_n-Boltzmann MFE with $\eta_n \to 0^+$ and for any $\varepsilon > 0$ there exists an integer $n' \in \mathbb{N}$ such that for all integers $n > n'$ we have

$$J^{\mu_n^*}(\pi_n^*) \geq \max_{\pi} J^{\mu_n^*}(\pi) - \varepsilon \ .$$

Let $\varepsilon' > 0$ be arbitrary and choose a sequence of optimal policies $\{\pi^N\}_{N\in\mathbb{N}}$ such that for all $N \in \mathbb{N}$ we have

$$\pi^N \in \arg\max_{\pi\in\Pi} J_1^N(\pi, \pi_n^*, \ldots, \pi_n^*) \,.$$

By Lemma B.5.6 there exists $N' \in \mathbb{N}$ such that for all $N > N'$ and all $n > n'$, we have

$$\begin{aligned}
\max_{\pi\in\Pi} J_1^N(\pi, \pi_n^*, \ldots, \pi_n^*) - \varepsilon - \varepsilon' &\leq \max_{\pi\in\Pi} J^{\mu_n^*}(\pi) - \varepsilon - \frac{\varepsilon'}{2} \\
&\leq J^{\mu_n^*}(\pi_n^*) - \frac{\varepsilon'}{2} \\
&\leq J_1^N(\pi_n^*, \pi_n^*, \ldots, \pi_n^*)
\end{aligned}$$

which is the desired approximate Nash equilibrium property since $\varepsilon, \varepsilon'$ are arbitrary. This applies by symmetry to all agents.

For RelEnt MFE, the same can be done by first showing the uniform convergence of the soft action-value function to the usual action-value function. For this, note that the smooth maximum Bellman recursion converges to the hard maximum Bellman recursion for any fixed μ.

Lemma B.8.5. *For any $f : \mathcal{A} \to \mathbb{R}$ and any $g : \mathcal{A} \to \mathbb{R}$ with $g(a) > 0$ for all $a \in \mathcal{A}$, we have*

$$\lim_{\eta\to 0^+} \eta \log \sum_{a\in\mathcal{A}} g(a) \exp \frac{f(a)}{\eta} = \max_{a\in\mathcal{A}} f(a) \,.$$

Proof. Let $\delta = \frac{1}{\eta} \to +\infty$. Then, by L'Hospital's rule we have

$$\lim_{\delta\to+\infty} \frac{\log \sum_{a\in\mathcal{A}} g(a) \exp\left(\delta f(a)\right)}{\delta} = \lim_{\delta\to+\infty} \frac{\sum_{a\in\mathcal{A}} g(a) \exp\left(\delta f(a)\right) f(a)}{\sum_{a\in\mathcal{A}} g(a) \exp\left(\delta f(a)\right)}$$

$$= \lim_{\delta \to +\infty} \frac{\sum_{a \in \mathcal{A}} g(a) \exp\left(\delta(f(a) - \max_{a \in \mathcal{A}} f(a))\right) f(a)}{\sum_{a \in \mathcal{A}} g(a) \exp\left(\delta(f(a) - \max_{a \in \mathcal{A}} f(a))\right)}$$
$$= \frac{|\mathcal{A}_{\max}| \max_{a \in \mathcal{A}} f(a)}{|\mathcal{A}_{\max}|} = \max_{a \in \mathcal{A}} f(a)$$

where $|\mathcal{A}_{\max}|$ is the number of elements in $\mathcal{A}$ that maximize f. ■

Using this result, we can show pointwise convergence of the soft action-value function to the action-value function.

Lemma B.8.6. *Any sequence of functions $(\mu \mapsto \tilde{Q}_{\eta_n}(\mu, t, s, a))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ converges pointwise to $\mu \mapsto Q^*(\mu, t, s, a)$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.*

Proof. Fix $\mu \in \mathcal{M}$. We show by induction that for any $\varepsilon > 0$, there exists $\eta_t > 0$ such that for all $\eta < \eta_t$ we have $\left|\tilde{Q}_\eta(\mu, t, s, a) - Q^*(\mu, t, s, a)\right| < \varepsilon$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$. This holds for $t = T - 1$ and arbitrary $s \in \mathcal{S}, a \in \mathcal{A}$ by Lemma B.8.5, since $r(s, a, \mu_{T-1})$ is independent of η. Assume this holds for $t+1$ and consider t. Then, by the induction assumption we can choose $\eta_{t+1} > 0$ such that for $\eta < \eta_{t+1}$, as $\eta \to 0^+$ we have

$$\begin{aligned}
\tilde{Q}_\eta(\mu, t, s, a) &= r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \eta \log \sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, s', a')}{\eta}\right) \\
&\leq r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \eta \log \sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{Q^*(\mu, t+1, s', a') + \frac{\varepsilon}{2}}{\eta}\right) \\
&\to r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \max_{a' \in \mathcal{A}} Q^*(\mu, t+1, s', a') + \frac{\varepsilon}{2}
\end{aligned}$$

by Lemma B.8.5 and monotonicity of log and exp. Analogously,

$$\begin{aligned}
\tilde{Q}_\eta(\mu, t, s, a) &\geq r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \eta \log \sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{Q^*(\mu, t+1, s', a') - \frac{\varepsilon}{2}}{\eta}\right) \\
&\to r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \max_{a' \in \mathcal{A}} Q^*(\mu, t+1, s', a') - \frac{\varepsilon}{2}\,.
\end{aligned}$$

Therefore, we can choose $\eta_t < \eta_{t+1}$ such that for all $\eta < \eta_t$ we have

$$\left|\tilde{Q}_\eta(\mu, t, s, a) - Q^*(\mu, t, s, a)\right| = \left|\tilde{Q}_\eta(\mu, t, s, a) - \left(r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \max_{a' \in \mathcal{A}} Q^*(\mu, t+1, s', a')\right)\right| < \varepsilon$$

which is the desired result. ■

We can now show that the soft action-value function converges uniformly to the action-value function as $\eta \to 0^+$.

Lemma B.8.7. *Any sequence of functions $(\mu \mapsto \tilde{Q}_{\eta_n}(\mu, t, s, a))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ converges uniformly to $\mu \mapsto Q^*(\mu, t, s, a)$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.*

Proof. First, we show that $\tilde{Q}_\eta(\mu, t, s, a)$ is monotonically decreasing in η for $\eta > 0$, i.e. $\frac{\partial}{\partial \eta} \tilde{Q}_\eta(\mu, t, s, a) \leq 0$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$. This is the case for $t = T - 1$ and arbitrary $s \in \mathcal{S}, a \in \mathcal{A}$, since $\tilde{Q}_\eta(\mu, T-1, s, a)$ is constant. Assume this holds for $t+1$, then for t and arbitrary $s \in \mathcal{S}, a \in \mathcal{A}$ we have

$$\begin{aligned}
\frac{\partial}{\partial \eta} \tilde{Q}_\eta(\mu, t, s, a) &= \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \log \sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, s', a')}{\eta}\right) \\
&+ \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \eta \frac{\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, s', a')}{\eta}\right) \left(-\frac{\tilde{Q}_\eta(\mu, t+1, s', a')}{\eta^2} + \frac{1}{\eta} \frac{\partial}{\partial \eta} \tilde{Q}_\eta(\mu, t+1, s', a')\right)}{\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, s', a')}{\eta}\right)}
\end{aligned}$$

$$\leq \max_{s' \in \mathcal{S}} \left(\log \sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, s', a')}{\eta} \right) - \frac{\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu,t+1,s',a')}{\eta} \right) \frac{\tilde{Q}_\eta(\mu,t+1,s',a')}{\eta}}{\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu,t+1,s',a')}{\eta} \right)} \right)$$

by induction hypothesis. Let $\xi_{a'} \equiv \frac{\tilde{Q}_\eta(\mu,t+1,s',a')}{\eta} \in \mathbb{R}$ and $s' \in \mathcal{S}$ arbitrary, then by Jensen's inequality applied to the convex function $\phi(x) = x \log x$ we have

$$\begin{aligned}
&\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \phi(\exp \xi_{a'}) \geq \phi\left(\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp \xi_{a'} \right) \\
\iff &\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \xi_{a'} \exp \xi_{a'} \geq \left(\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp \xi_{a'} \right) \log \left(\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp \xi_{a'} \right) \\
\iff &\log \left(\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp \xi_{a'} \right) - \frac{\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \xi_{a'} \exp \xi_{a'}}{\left(\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp \xi_{a'} \right)} \leq 0 \,,
\end{aligned}$$

such that $\tilde{Q}_\eta(\mu, t, s, a)$ is monotonically decreasing for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ by induction.

Furthermore, $\mathcal{M}$ is compact and both $\tilde{Q}_\eta$ and Q are compositions, sums, products and finite maxima of continuous functions in μ and therefore continuous in μ by the standing assumptions. Since $(\mu \mapsto \tilde{Q}_{\eta_n}(\mu, t, s, a))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ converges pointwise to $\mu \mapsto Q^*(\mu, t, s, a)$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ by Lemma B.8.6, by Dini's theorem the convergence is uniform. ■

Now that $\tilde{Q}_\eta$ converges uniformly against Q, we can show that RelEnt MFE have vanishing exploitability by replicating the proof for Boltzmann MFE.

Lemma B.8.8. *Any sequence of functions $(\mu \mapsto Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, s, a))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ converges pointwise to $\mu \mapsto Q^*(\mu, t, s, a)$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.*

Proof. The proof is the same as in Lemma B.8.1. The only difference is that we additionally choose $n_2 \in \mathbb{N}$ in each induction step such that for all $n > n_2$ we have

$$\left| \tilde{Q}_\eta(\mu, t, s, a) - Q^*(\mu, t, s, a) \right| \leq \frac{\Delta Q_{\min}^{s', \mu}}{4}$$

for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$, which is possible, since by Lemma B.8.7, $\tilde{Q}_\eta$ converges uniformly against Q. As long as we choose $n' \equiv \max(n_1, n_2, \max_{s' \in \mathcal{S}, a' \in \mathcal{A}} n_{s',a'})$, the rest of the proof will apply. ■

Lemma B.8.9. *Any sequence of functions $(\mu \mapsto Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, s, a))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ fulfills equicontinuity for large enough n: For any $\varepsilon > 0$ and any $\mu \in \mathcal{M}$, we can choose a $\delta > 0$ and an integer $n' \in \mathbb{N}$ such that for all $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu, \mu') < \delta$ and for all $n > n'$ we have*

$$\left| Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, s, a) - Q^{\tilde{\Phi}_{\eta_n}(\mu')}(\mu', t, s, a) \right| < \varepsilon \,.$$

Proof. To obtain the desired property, we replicate the proof of Lemma B.8.2 by setting $\mathcal{F} = (\mu \mapsto Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, s, a))_{n \in \mathbb{N}}$. Any bounds for $\tilde{Q}_\eta$ can be instantiated by the corresponding bound for Q and then bounding the distance between both by uniform convergence. The only differences lie in bounding the terms

$$\left| (\tilde{\Phi}_{\eta_n}(\mu)(a_{\text{sub}} \mid s') - (\tilde{\Phi}_{\eta_n}(\mu')(a_{\text{sub}} \mid s') \right|$$

where the action-value function has been replaced with the soft action-value function. Since $\tilde{Q}_{\eta_n}$ uniformly converges to Q, we instantiate additional requirements $N_{t,s,a}^{s'}, \tilde{N}_{t,s,a}^{s'}$ to let $n > N_{t,s,a}^{s'}, n > \tilde{N}_{t,s,a}^{s'}$ large enough such that η is sufficiently small enough.

The first difference is to obtain

$$\left| \tilde{Q}_{\eta_n}(\mu', t, s, a) - \tilde{Q}_{\eta_n}(\mu, t, s, a) \right| < \frac{\Delta Q_{\min}^{s', \mu}}{4}$$

for all $\mu' \in \mathcal{M}, t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ with $d_{\mathcal{M}}(\mu, \mu')$ sufficiently small. We choose $\hat{\delta}^3_{t,s,a}$ slightly stronger than in the original proof, such that if $d_{\mathcal{M}}(\mu, \mu') < \hat{\delta}^3_{t,s,a}$, we have

$$|Q^*(\mu', t, s, a) - Q^*(\mu, t, s, a)| < \frac{\Delta Q^{s',\mu}_{\min}}{12}.$$

We must then additionally choose $N^{s'}_{t,s,a} \in \mathbb{N}$ for each induction step via uniform convergence from Lemma B.8.7 such that as long as $n > N^{s'}_{t,s,a}$, we have

$$\left|\tilde{Q}_{\eta_n}(\mu, t, s, a) - Q^*(\mu, t, s, a)\right| < \frac{\Delta Q^{s',\mu}_{\min}}{12}.$$

This implies the required inequality

$$\begin{aligned}\left|\tilde{Q}_{\eta_n}(\mu', t, s, a) - \tilde{Q}_{\eta_n}(\mu, t, s, a)\right| &\leq \left|\tilde{Q}_{\eta_n}(\mu', t, s, a) - Q^*(\mu', t, s, a)\right| + |Q^*(\mu', t, s, a) - Q^*(\mu, t, s, a)| \\ &\quad + \left|Q^*(\mu, t, s, a) - \tilde{Q}_{\eta_n}(\mu, t, s, a)\right| < \frac{\Delta Q^{s',\mu}_{\min}}{4}\end{aligned}$$

and we can proceed as in the original proof.

The second difference lies in choosing $\delta^{4,s'}_{t,s,a}$. Note that $\tilde{Q}_{\eta_n}$ is still bounded by M_Q, see Lemma B.7.1. However, since $\tilde{Q}_{\eta_n}$ might no longer be Lipschitz with the same constant as Q^*, we choose an additional integer $\tilde{N}^{s'}_{t,s,a} \in \mathbb{N}$ for each induction step by Lemma B.8.7, such that as long as $n > \tilde{N}^{s'}_{t,s,a}$, we have

$$\left|\tilde{Q}_{\eta_n}(\mu, t, s, a) - Q^*(\mu, t, s, a)\right| \leq \Delta^{s'}_Q \equiv \frac{\frac{\varepsilon_{t,s,a}}{16M_Q|\mathcal{A}|}}{4R^{\max}_q|\mathcal{A}| \cdot \frac{1}{\eta^{s'}_{\min}} \exp\left(\frac{2M_Q}{\eta^{s'}_{\min}}\right)}$$

for any $\mu' \in \mathcal{M}, t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$. The required bound then follows immediately from

$$\begin{aligned}&|(\Phi_{\eta_n}(\mu)(a_{\text{sub}} \mid s') - (\Phi_{\eta_n}(\mu')(a_{\text{sub}} \mid s')| \\ &\leq R^{\max}_q \sum_{a' \neq a_{\text{sub}}} \left|\exp\left(\frac{\tilde{Q}_{\eta_n}(\mu', t, s', a') - \tilde{Q}_{\eta_n}(\mu', t, s', a_{\text{sub}})}{\eta}\right) - \exp\left(\frac{\tilde{Q}_{\eta_n}(\mu, t, s', a') - \tilde{Q}_{\eta_n}(\mu, t, s', a_{\text{sub}})}{\eta}\right)\right| \\ &\leq R^{\max}_q \sum_{a' \neq a_{\text{sub}}} \left|\frac{1}{\eta}\exp\left(\frac{\xi_{a'}}{\eta}\right)\right| \left|(\tilde{Q}_{\eta_n}(\mu', t, s', a') - \tilde{Q}_{\eta_n}(\mu', t, s', a_{\text{sub}})) - (\tilde{Q}_{\eta_n}(\mu, t, s', a') - \tilde{Q}_{\eta_n}(\mu, t, s', a_{\text{sub}}))\right| \\ &\leq R^{\max}_q|\mathcal{A}| \cdot \frac{1}{\eta^{s'}_{\min}} \exp\left(\frac{2M_Q}{\eta^{s'}_{\min}}\right) \left(\left|\tilde{Q}_{\eta_n}(\mu', t, s', a') - \tilde{Q}_{\eta_n}(\mu, t, s', a')\right| + \left|\tilde{Q}_{\eta_n}(\mu, t, s', a_{\text{sub}}) - \tilde{Q}_{\eta_n}(\mu', t, s', a_{\text{sub}})\right|\right) \\ &\leq R^{\max}_q|\mathcal{A}| \cdot \frac{1}{\eta^{s'}_{\min}} \exp\left(\frac{2M_Q}{\eta^{s'}_{\min}}\right) \cdot \left(2K_Q d_{\mathcal{M}}(\mu, \mu') + 4\Delta^{s'}_Q\right) \\ &\leq R^{\max}_q|\mathcal{A}| \cdot \frac{1}{\eta^{s'}_{\min}} \exp\left(\frac{2M_Q}{\eta^{s'}_{\min}}\right) \cdot (2K_Q d_{\mathcal{M}}(\mu, \mu')) + \frac{\varepsilon_{t,s,a}}{16M_Q|\mathcal{A}|} < \frac{\varepsilon_{t,s,a}}{8M_Q|\mathcal{A}|}\end{aligned}$$

as in the original proof by letting $d_{\mathcal{M}}(\mu, \mu') < \delta^{4,s'}_{t,s,a}$ and choosing

$$\delta^{4,s'}_{t,s,a} = \frac{\varepsilon_{t,s,a}\eta^{s'}_{\min}}{16M_Q|\mathcal{A}|^2 R^{\max}_q \cdot \exp\left(\frac{2M_Q}{\eta^{s'}_{\min}}\right) \cdot 2K_Q}.$$

The rest of the proof is analogous. We obtain the additional requirement $n > N^{s'}_{t,s,a}$, $n > \tilde{N}^{s'}_{t,s,a}$ for some integers $N^{s'}_{t,s,a}, \tilde{N}^{s'}_{t,s,a}$ and each $t \in \mathcal{T}, s \in \mathcal{S}, s' \in \mathcal{S}, a \in \mathcal{A}$. By choosing $n' \equiv \max_{t \in \mathcal{T}, s \in \mathcal{S}, s' \in \mathcal{S}, a \in \mathcal{A}} \max(N^{s'}_{t,s,a}, \tilde{N}^{s'}_{t,s,a})$, the desired result holds as long as $n > n'$. ■

From this property, we again obtain the desired uniform convergence via compactness of $\mathcal{M}$.

Lemma B.8.10. *Any sequence of functions* $(\mu \mapsto Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, s, a))_{n\in\mathbb{N}}$ *with* $\eta_n \to 0^+$ *converges uniformly to* $\mu \mapsto Q^*(\mu, t, s, a)$ *for all* $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$.

Proof. Fix $\varepsilon > 0, t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$. Then, there exists by Lemma B.8.9 for any point $\mu \in \mathcal{M}$ both $\delta(\mu)$ and n' such that for all $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu, \mu') < \delta(\mu)$ for all $n > n'$ we have

$$\left|Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, s, a) - Q^{\tilde{\Phi}_{\eta_n}(\mu')}(\mu', t, s, a)\right| < \frac{\varepsilon}{3}$$

which via pointwise convergence from Lemma B.8.8 implies

$$|Q^*(\mu, t, s, a) - Q^*(\mu', t, s, a)| \leq \frac{\varepsilon}{3}\,.$$

Since $\mathcal{M}$ is compact, it is separable, i.e. there exists a countable dense subset $(\mu_j)_{j\in\mathbb{N}}$ of $\mathcal{M}$. Let $\delta(\mu)$ be as defined above and cover $\mathcal{M}$ by the open balls $(B_{\delta(\mu_j)}(\mu_j))_{j\in\mathbb{N}}$. By the compactness of $\mathcal{M}$, finitely many of these balls $B_{\delta(\mu_{n_1})}(\mu_{n_1}), \ldots, B_{\delta(\mu_{n_k})}(\mu_{n_k})$ cover $\mathcal{M}$. By pointwise convergence from Lemma B.8.8, for any $i = 1, \ldots, k$ we can find integers m_i such that for all $n > m_i$ we have

$$\left|Q^{\tilde{\Phi}_{\eta_n}(\mu_{n_i})}(\mu_{n_i}, t, s, a) - Q^*(\mu_{n_i}, t, s, a)\right| < \frac{\varepsilon}{3}\,.$$

Taken together, we find that for $n > \max(n', \max_{i=1,\ldots,k} m_i)$ and arbitrary $\mu \in \mathcal{M}$, we have

$$\begin{aligned}\left|Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, s, a) - Q^*(\mu, t, s, a)\right| &< \left|Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, s, a) - Q^{\tilde{\Phi}_{\eta_n}(\mu_{n_i})}(\mu_{n_i}, t, s, a)\right| \\ &\quad + \left|Q^{\tilde{\Phi}_{\eta_n}(\mu_{n_i})}(\mu_{n_i}, t, s, a) - Q^*(\mu_{n_i}, t, s, a)\right| \\ &\quad + |Q^*(\mu_{n_i}, t, s, a) - Q^*(\mu, t, s, a)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} < \varepsilon\end{aligned}$$

for some center point μ_{n_i} of a ball containing μ from the finite cover. ■

As a result, a sequence of RelEnt MFE with $\eta \to 0^+$ is approximately optimal in the MFG.

Lemma B.8.11. *For any sequence* $(\pi_n^*, \mu_n^*)_{n\in\mathbb{N}}$ *of* η_n*-RelEnt MFE with* $\eta_n \to 0^+$ *and for any* $\varepsilon > 0$ *there exists integer* $n' \in \mathbb{N}$ *such that for all integers* $n > n'$ *we have*

$$J^{\mu_n^*}(\pi_n^*) \geq \max_{\pi} J^{\mu_n^*}(\pi) - \varepsilon\ .$$

Proof. By Lemma B.8.10, we have $\left|Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, s, a) - Q^*(\mu, t, s, a)\right| \to 0$ uniformly. Therefore, for any $\varepsilon > 0$, there exists by uniform convergence an integer n' such that for all integers $n > n'$ we have

$$Q^{\pi_n^*}(\mu_n^*, t, s, a) \geq Q^*(\mu_n^*, t, s, a) - \varepsilon = \max_{\pi\in\Pi} Q^{\pi}(\mu_n^*, t, s, a) - \varepsilon\,,$$

and since by Lemma B.3.1, we have

$$J^{\mu_n^*}(\pi_n^*) = \sum_{s\in\mathcal{S}} \mu_0(s) \cdot \sum_{a\in\mathcal{A}} Q^{\pi_n^*}(\mu_n^*, t, s, a) \geq \sum_{s\in\mathcal{S}} \mu_0(s) \cdot \max_{\pi\in\Pi} \sum_{a\in\mathcal{A}} Q^{\pi}(\mu_n^*, t, s, a) - \varepsilon = \max_{\pi\in\Pi} J^{\mu_n^*}(\pi) - \varepsilon\,,$$

the desired result follows immediately. ■

By repeating the previous argumentation for Boltzmann MFE with Lemma B.5.6 and replacing Lemma B.8.4 with Lemma B.8.11, we obtain the desired result for RelEnt MFE. □

C Relative entropy mean field games

We show that the necessary conditions for optimality hold for the candidate solution. (For further insight, see also Neu et al. (2017), Haarnoja et al. (2017) and references therein.) Fix a mean field $\mu \in \mathcal{M}$ and formulate the induced problem as an optimization problem, with $\rho_t(s)$ as the probability of our representative agent visiting state $s \in \mathcal{S}$ at time $t \in \mathcal{T}$, to obtain

$$\begin{aligned}
\max_{\rho,\pi} \quad & \sum_{t=0}^{T-1} \sum_{s\in\mathcal{S}} \rho_t(s) \sum_{a\in\mathcal{A}} \pi_t(a \mid s) r(s,a,\mu_t) \\
\text{subject to} \quad & \rho_{t+1}(s') = \sum_{s\in\mathcal{S}} \rho_t(s) \sum_{a\in\mathcal{A}} \pi_t(a \mid s) p(s' \mid s,a,\mu_t) \qquad \forall s' \in \mathcal{S}, t \in \{0,\dots,T-2\}, \\
& 1 = \sum_{s\in\mathcal{S}} \rho_t(s) \qquad \forall t \in \{0,\dots,T-1\}, \\
& 1 = \sum_{a\in\mathcal{A}} \pi_t(a \mid s) \qquad \forall s \in \mathcal{S}, t \in \{0,\dots,T-1\}, \\
& 0 \le \rho_t(s), 0 \le \pi_t(a \mid s) \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in \{0,\dots,T-1\}, \\
& \mu_0(s) = \rho_0(s) \qquad \forall s \in \mathcal{S}.
\end{aligned}$$

Note that if the agent follows the mean field policy of the other agents, we have $\rho_t = \mu_t$. The optimized objective is just the expectation $\mathbb{E}\left[\sum_{t=0}^{T-1} r(S_t, A_t)\right]$. As in Belousov and Peters (2019), we change this objective to include a KL-divergence penalty weighted by the state-visitation distribution $\rho_t(\cdot)$ by introducing the temperature $\eta > 0$ and prior policy $q \in \Pi$ to obtain

$$\begin{aligned}
\max_{\rho_t,\pi_t} \quad & \sum_{t=0}^{T-1} \sum_{s\in\mathcal{S}} \rho_t(s) \sum_{a\in\mathcal{A}} \pi_t(a \mid s) r(s,a,\mu_t) - \eta \sum_{t=0}^{T-1} \sum_{s\in\mathcal{S}} \rho_t(s) D_{\mathrm{KL}}(\pi_t(\cdot \mid s) \parallel q_t(\cdot \mid s)) \\
\text{subject to} \quad & \rho_{t+1}(s') = \sum_{s\in\mathcal{S}} \rho_t(s) \sum_{a\in\mathcal{A}} \pi_t(a \mid s) p(s' \mid s,a,\mu_t) \qquad \forall s' \in \mathcal{S}, t \in \{0,\dots,T-2\}, \\
& 1 = \sum_{s\in\mathcal{S}} \rho_t(s) \qquad \forall t \in \{0,\dots,T-1\}, \\
& 1 = \sum_{a\in\mathcal{A}} \pi_t(a \mid s) \qquad \forall s \in \mathcal{S}, t \in \{0,\dots,T-1\}, \\
& 0 \le \rho_t(s), 0 \le \pi_t(a \mid s) \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in \{0,\dots,T-1\}, \\
& \mu_0(s) = \rho_0(s) \qquad \forall s \in \mathcal{S}.
\end{aligned}$$

We ignore the constraints $0 \le \pi_t(a \mid s)$ and $0 \le \rho_t(s)$ and see later that they will hold automatically. This results in the simplified optimization problem

$$\begin{aligned}
\max_{\rho_t,\pi_t} \quad & \sum_{t=0}^{T-1} \sum_{s\in\mathcal{S}} \rho_t(s) \sum_{a\in\mathcal{A}} \pi_t(a \mid s) r(s,a,\mu_t) - \eta \sum_{t=0}^{T-1} \sum_{s\in\mathcal{S}} \rho_t(s) D_{\mathrm{KL}}(\pi_t(\cdot \mid s) \parallel q_t(\cdot \mid s)) \\
\text{subject to} \quad & \rho_{t+1}(s') = \sum_{s\in\mathcal{S}} \rho_t(s) \sum_{a\in\mathcal{A}} \pi_t(a \mid s) p(s' \mid s,a,\mu_t) \qquad \forall s' \in \mathcal{S}, t \in \{0,\dots,T-2\}, \\
& 1 = \sum_{s\in\mathcal{S}} \rho_t(s) \qquad \forall t \in \{0,\dots,T-1\}, \\
& 1 = \sum_{a\in\mathcal{A}} \pi_t(a \mid s) \qquad \forall s \in \mathcal{S}, t \in \{0,\dots,T-1\}, \\
& \mu_0(s) = \rho_0(s) \qquad \forall s \in \mathcal{S},
\end{aligned}$$

for which we introduce Lagrange multipliers $\lambda_1(t,s)$, $\lambda_2(t)$, $\lambda_3(t,s)$, $\lambda_4(s)$ and the Lagrangian

$$L(\rho,\pi,\lambda_1,\lambda_2,\lambda_3,\lambda_4) = \sum_{t=0}^{T-1} \sum_{s\in\mathcal{S}} \rho_t(s) \sum_{a\in\mathcal{A}} \pi_t(a \mid s) \left(r(s,a,\mu_t) - \eta \log \frac{\pi_t(a \mid s)}{q_t(a \mid s)} \right)$$

$$
\begin{aligned}
&- \sum_{t=0}^{T-1} \sum_{s' \in \mathcal{S}} \lambda_1(t, s') \left(\rho_{t+1}(s') - \sum_{s \in \mathcal{S}} \rho_t(s) \sum_{a \in \mathcal{A}} \pi_t(a \mid s) p(s' \mid s, a, \mu_t) \right) \\
&- \sum_{t=0}^{T-1} \lambda_2(t) \left(1 - \sum_{s \in \mathcal{S}} \rho_t(s) \right) \\
&- \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}} \lambda_3(t, s) \left(\sum_{a \in \mathcal{A}} \pi_t(a \mid s) - 1 \right) \\
&- \sum_{s \in \mathcal{S}} \lambda_4(s) \left(\mu_0(s) - \rho_0(s) \right)
\end{aligned}
$$

with the artificial constraint $\lambda_1(T-1, s) \equiv 0$, which allows us to formulate the following necessary conditions for optimality. For $\nabla_{\pi_t(a|s)} L$ and all $s \in \mathcal{S}, a \in \mathcal{A}, t \in \{0, \dots, T-1\}$, we obtain

$$
\begin{aligned}
\nabla_{\pi_t(a|s)} L &= \rho_t(s) \left(r(s, a, \mu_t) - \eta \log \frac{\pi_t(a \mid s)}{q_t(a \mid s)} - \eta + \sum_{s' \in \mathcal{S}} \lambda_1(t, s') p(s' \mid s, a, \mu_t) \right) - \lambda_3(t, s) \stackrel{!}{=} 0 \\
\implies \pi_t^*(a \mid s) &= q_t(a \mid s) \exp \left(\frac{r(s, a, \mu_t) - \eta + \sum_{s' \in \mathcal{S}} \lambda_1(t, s') p(s' \mid s, a, \mu_t) - \frac{\lambda_3(t,s)}{\rho_t(s)}}{\eta} \right) .
\end{aligned}
$$

For $\nabla_{\lambda_3} L$ and all $s \in \mathcal{S}, t \in \{0, \dots, T-1\}$, by inserting π_t^* we obtain

$$
\begin{aligned}
\nabla_{\lambda_3(t,s)} L &= 1 - \sum_{a \in \mathcal{A}} \pi_t(a \mid s) \stackrel{!}{=} 0 \\
\iff 1 &= \sum_{a \in \mathcal{A}} q_t(a \mid s) \exp \left(\frac{r(s, a, \mu_t) - \eta + \sum_{s' \in \mathcal{S}} \lambda_1(t, s') p(s' \mid s, a, \mu_t) - \frac{\lambda_3(t,s)}{\rho_t(s)}}{\eta} \right)
\end{aligned}
$$

which is fulfilled by choosing

$$
\lambda_3^*(t, s) = \eta \rho_t(s) \log \sum_{a \in \mathcal{A}} q_t(a \mid s) \exp \left(\frac{r(s, a, \mu_t) - \eta + \sum_{s' \in \mathcal{S}} \lambda_1(t, s') p(s' \mid s, a, \mu_t)}{\eta} \right)
$$

since it fulfills the required equation

$$
\begin{aligned}
&\sum_{a \in \mathcal{A}} q_t(a \mid s) \exp \left(\frac{r(s, a, \mu_t) - \eta + \sum_{s' \in \mathcal{S}} \lambda_1(t, s') p(s' \mid s, a, \mu_t) - \frac{\lambda_3^*(t,s)}{\rho_t(s)}}{\eta} \right) \\
&= \sum_{a \in \mathcal{A}} q_t(a \mid s) \exp \left(\frac{r(s, a, \mu_t) - \eta + \sum_{s' \in \mathcal{S}} \lambda_1(t, s') p(s' \mid s, a, \mu_t)}{\eta} \right) \\
&\quad \cdot \left(\sum_{a \in \mathcal{A}} q_t(a \mid s) \exp \left(\frac{r(s, a, \mu_t) - \eta + \sum_{s' \in \mathcal{S}} \lambda_1(t, s') p(s' \mid s, a, \mu_t)}{\eta} \right) \right)^{-1} = 1 .
\end{aligned}
$$

Finally, inserting λ_3^* and π^*, for $\nabla_{\rho_t(s)} L$ we obtain

$$
\begin{aligned}
\nabla_{\rho_t(s)} L &= \sum_{a \in \mathcal{A}} \pi_t(a \mid s) \left(r(s, a, \mu_t) - \eta \log \frac{\pi_t(a \mid s)}{q_t(a \mid s)} + \sum_{s' \in \mathcal{S}} \lambda_1(t, s') p(s' \mid s, a, \mu_t) + \lambda_2(t) \right) - \lambda_1(t-1, s) \\
&= \sum_{a \in \mathcal{A}} \pi_t(a \mid s) \left(\eta + \lambda_2(t) + \frac{\lambda_3(t, s)}{\rho_t(s)} \right) - \lambda_1(t-1, s) \stackrel{!}{=} 0
\end{aligned}
$$

which implies

$$
\lambda_1^*(t-1, s) = \eta + \lambda_2(t) + \eta \log \sum_{a \in \mathcal{A}} q_t(a \mid s) \exp \left(\frac{r(s, a, \mu_t) - \eta + \sum_{s' \in \mathcal{S}} \lambda_1(t, s') p(s' \mid s, a, \mu_t)}{\eta} \right) .
$$

We can subtract $\lambda_2(t)$ and shift the time index to obtain the soft value function $\tilde{V}_\eta(\mu, t, s)$ defined via terminal condition $\tilde{V}_\eta(\mu, T, s) \equiv 0$ and the recursion

$$\tilde{V}_\eta(\mu, t, s) = \eta \log \sum_{a \in \mathcal{A}} q_t(a \mid s) \exp\left(\frac{r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} \tilde{V}_\eta(\mu, t+1, s') p(s' \mid s, a, \mu_t)}{\eta}\right)$$

since then, by normalization the optimal policy for all $s \in \mathcal{S}, a \in \mathcal{A}, t \in \{0, \dots, T-1\}$ is equivalent to

$$\begin{aligned}\pi_t^*(a \mid s) &= \frac{q_t(a \mid s) \exp\left(\frac{r(s,a,\mu_t) + \sum_{s' \in \mathcal{S}} \lambda_1(t,s') p(s' \mid s, a, \mu_t)}{\eta}\right)}{\sum_{a' \in \mathcal{A}} q_t(a' \mid s) \exp\left(\frac{r(s,a',\mu_t) + \sum_{s' \in \mathcal{S}} \lambda_1(t,s') p(s' \mid s, a', \mu_t)}{\eta}\right)} \\ &= \frac{q_t(a \mid s) \exp\left(\frac{r(s,a,\mu_t) + \sum_{s' \in \mathcal{S}} \tilde{V}_\eta(\mu,t+1,s') p(s' \mid s, a, \mu_t)}{\eta}\right)}{\sum_{a' \in \mathcal{A}} q_t(a' \mid s) \exp\left(\frac{r(s,a',\mu_t) + \sum_{s' \in \mathcal{S}} \tilde{V}_\eta(\mu,t+1,s') p(s' \mid s, a', \mu_t)}{\eta}\right)}.\end{aligned}$$

To obtain a recursion in $\tilde{Q}_\eta$, define

$$\tilde{Q}_\eta(\mu, t, s, a) \equiv r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \eta \log \sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, s', a')}{\eta}\right)$$

with terminal condition $\tilde{Q}_\eta(\mu, T, s, a) \equiv 0$ to obtain

$$\pi_t^*(a \mid s) = \frac{q_t(a \mid s) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,s,a)}{\eta}\right)}{\sum_{a' \in \mathcal{A}} q_t(a' \mid s) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,s,a')}{\eta}\right)}$$

which is the desired result as π^* fulfills all constraints and determines ρ uniquely. For the uniform prior $q_t(a \mid s) = 1/|\mathcal{A}|$, we obtain the maximum entropy solution.

References

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003, 2016.

Lloyd Shapley. Some topics in two-person games. *Advances in game theory*, 52:1–29, 1964.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Naci Saldi, Tamer Basar, and Maxim Raginsky. Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1352–1361, 2017.

Boris Belousov and Jan Peters. Entropic regularization of markov decision processes. *Entropy*, 21(7):674, 2019.