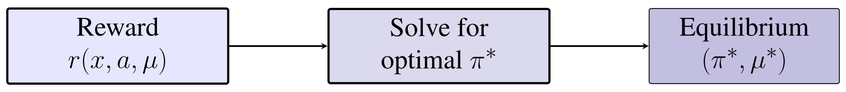
Average-Cost Mean-Field Games: Forward and Inverse Perspectives

Güneş Akbaş [1] and Şevket Kaan Alkır [2]
gunes.akbas@bilkent.edu.tr [1], kaan.alkir@bilkent.edu.tr[2]
Bilkent University - Department of Mathematics

1 Forward Reinforcement Learning



Mean-field Term

Let $(x_{1,t}, \ldots, x_{N,t})$ be exchangeable. Then

$$\mu_t^N := \frac{1}{N}\sum_{j=1}^{N} \delta_{x_{j,t}} \;\to\; \mu_t \quad \text{(a.s.) as } N \to \infty,$$

under Assumption (1). And conditioned on μ_t, agents are independent.

Representative-Agent Problem (Average Cost)

Fix a stationary population law μ (or a stationary flow $\{\mu_t\}$ that has converged). The representative agent chooses a stationary policy π to minimise the long-run average cost

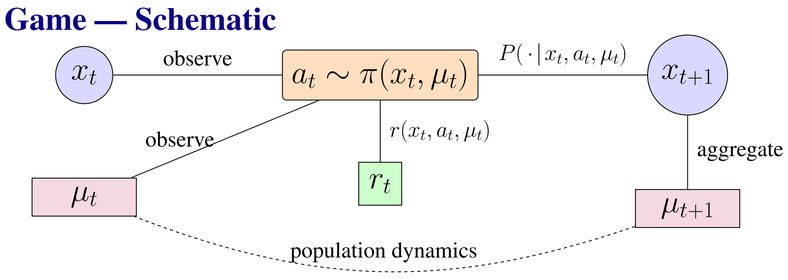
$$V_\mu(\pi, \mu_0) = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}^\pi\Big[r(x_t,\ a_t,\ \mu)\Big] \quad \textit{where} \quad x_0 \sim \mu_0.$$

Here, dynamics follows $x_{t+1} \sim P\big(\cdot \mid x_t, a_t, \mu\big)$, $a_t \sim \pi(\cdot \mid x_t)$. Now define

$$\rho(\mu) := \max_\pi V_\mu(\pi, \mu_0).$$

This criterion will serve as our reference for optimality.

Mean-Field Game — Schematic



State	$X_t \in S$ (discrete)
Population	$\mu_t \in \mathcal{P}(S)$ empirical distribution of states
Action	$A_t \in \mathcal{A}(X_t)$ chosen by policy $\pi(\cdot \mid X_t, \mu_t)$
Reward	$R_t = r(X_t, A_t, \mu_t)$
Dynamics	$X_{t+1} \sim P(\cdot \mid X_t, A_t, \mu_t)$

Coupling: each agent uses (X_t, μ_t), while μ_t aggregates everyone's states.

2 Mean-Field Average-Cost Soft Bellman Equation

Mean-Field Soft Average-Cost Bellman

For a fixed population law μ and parameter $\delta > 0$:

$$\rho(\mu) + h(x;\mu) \;=\; \max_{\pi(\cdot|x)} \Big\{ \sum_a \pi(a \mid x)\Big[r(x,a;\mu) + \sum_y P(y \mid x,a;\mu)\, h(y;\mu)\Big] + \delta\, H\big(\pi(\cdot \mid x)\big)\Big\},$$

where $H(\pi) := -\sum_a \pi(a)\log\pi(a)$.

Softmax Closed Form

$$\rho(\mu) + h(x;\mu) \;=\; \delta\, \log \sum_a \exp\Big(\tfrac{1}{\delta}\big[r(x,a;\mu) + \sum_y P(y \mid x,a;\mu)\, h(y;\mu)\big]\Big).$$

Optimal Stochastic Policy

$$\pi^*(a \mid x;\mu) \;=\; \frac{\exp\big(\frac{1}{\delta}\big[r(x,a;\mu) + \sum_y P(y \mid x,a;\mu)\, h(y;\mu)\big]\big)}{\sum_{a'} \exp\big(\frac{1}{\delta}\big[r(x,a';\mu) + \sum_y P(y \mid x,a';\mu)\, h(y;\mu)\big]\big)}.$$

3 Fixed-Point Iteration for Mean-Field Equilibrium

(1) Best-Response / Softmax Policy For a population measure μ and entropy weight $\delta > 0$, define

$$\begin{aligned}\pi_\mu(\cdot \mid x) &= \operatorname{softmax}_{\frac{1}{\delta}}\big(Q_\mu(x,\cdot)\big) \\ &= \arg\max_{\pi(\cdot|x)\in\Delta(\mathcal{A})} \Big[\sum_{a\in\mathcal{A}} \pi(a \mid x)\big(r(x,a;\mu) + \sum_{y\in\mathcal{X}} P(y \mid x,a;\mu)\, h_\mu(y)\big) + \delta\, H\big(\pi(\cdot \mid x)\big)\Big]\end{aligned}$$

with h_μ solving the average-cost Bellman equation under μ.

(2) Population-Update Map Using π_μ, propagate the population forward:

$$\mu'(y) = \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} P\big(y \mid x,a,\mu\big)\, \pi_\mu(a \mid x)\, \mu(x), \qquad y \in \mathcal{X}.$$

A *mean-field equilibrium* is a fixed point $\mu^* = \mu'$ of this two-step operator.

Corollary (Policy Lipschitz)

For the soft-max policy

$$\pi_\mu(a \mid x) \;\propto\; \exp\big(Q^*_\mu(x,a)/\tau\big),$$

there exists a constant $C_\pi > 0$ such that

$$\|\pi_\mu - \pi_{\mu'}\|_{\ell^1} \;\le\; C_\pi\, \|\mu - \mu'\|_{\ell^1}.$$

Algorithm 1 Contraction Fixed-Point Iteration for Mean-Field Equilibrium

Require: Initial measure $\mu_0 \in \mathcal{P}(\mathcal{X})$, temperature $\delta > 0$, tolerance $\varepsilon > 0$

$\mu \leftarrow \mu_0$
repeat
  $h \leftarrow \textsc{BellmanSolve}(\mu)$ // span-contraction step
  $\pi \leftarrow \textsc{Softmax}_{1/\delta}(h)$
  $\mu' \leftarrow \textsc{PopUpdate}(\mu, \pi)$
  $\Delta \leftarrow \|\mu' - \mu\|_1$; $\mu \leftarrow \mu'$
until $\Delta \le \varepsilon$
return Equilibrium measure μ and policy π

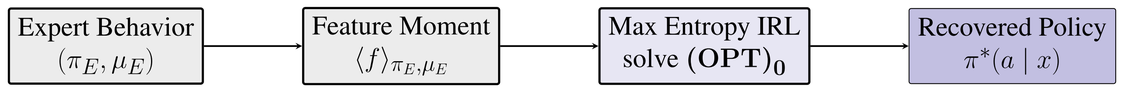
References

[1] Óscar Hernández-Lerma and Jean B. Lasserre (1996), *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer, New York.

[2] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi, *Q-Learning in Regularized Mean-field Games*, *Dynamic Games and Applications*, vol. 13, no. 1, pp. 89–117, Mar. 2023.

[3] Bolin Gao and Lacra Pavel, A game-theoretic approach to apprenticeship learning, in *On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning*, arXiv preprint arXiv:1704.00805, 2017.

1 Inverse Reinforcement Learning



A widely used assumption in IRL for MFGs is that the reward function can be expressed as a linear combination of feature vectors, which depend on the state, action, and mean-field term:

$$\mathcal{R} := \Big\{ r(x,a,\mu) = \langle \theta, f(x,a,\mu)\rangle \;\mid\; \theta \in \mathbb{R}^k,\ f : \mathcal{X}\times\mathcal{A}\times\mathcal{P}(\mathcal{X}) \to \mathbb{R}^k \Big\}$$

where $f(x,a,\mu) \in \mathbb{R}^k$ is the feature vector that includes information about the state-action pair (x,a) and the mean-field term μ.

Assumption 1 *We assume that for any policy π and mean-field term μ, the state process is positive Harris recurrent and aperiodic, guaranteeing the convergence of empirical averages [1, Thm. 13.3.3]. In other words, the state process exhibits ergodic behavior under any policy and mean-field term, ensuring long-run statistical stability. Assuming the expert population distribution μ_E is known is standard, as it can be consistently estimated from long-run trajectories. [1, Thm. 13.3.3]*

Definition 1 *average-reward expected feature vector under the mean-field equilibrium (π_E, μ_E) for $x_0 \sim \mu_E$ is defined as,*

$$\langle f \rangle_{\pi_E,\mu_E} := \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}^{\pi_E,\mu_E}\left[f(x_t, a_t, \mu_E)\right]. \tag{1}$$

2 Entropy-Regularized Inverse Learning

Definition 2 *Average-reward causal entropy $H(\pi)$ of the policy $\pi \in \Pi$ is defined as follows*

$$H(\pi) := \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}^{\pi,\mu_E}\left[-\log \pi(a_t \mid x_t)\right].$$

This principle selects the least biased policy among those consistent with expert demonstrations [2] . We define the average-cost maximum causal entropy IRL problem:

$$\begin{aligned}(\mathbf{OPT})_0 \ \text{maximize}_\pi \quad & H(\pi) \\ \text{subject to} \quad & \pi(a \mid x) \ge 0 \ \forall (x,a) \in \mathcal{X}\times\mathcal{A} \\ & \textstyle\sum_{a\in\mathcal{A}} \pi(a \mid x) = 1 \ \forall x \in \mathcal{X} \\ & \mu_E(x) = \textstyle\sum_{(a,y)\in\mathcal{A}\times\mathcal{X}} p(x \mid y,a,\mu_E)\, \pi(a \mid y)\, \mu_E(y) \ \forall x \in \mathcal{X} \\ & \lim_{T\to\infty} \tfrac{1}{T}\textstyle\sum_{t=0}^{T-1} \mathbb{E}^{\pi,\mu_E}[f(x_t,a_t,\mu_E)] = \langle f\rangle_{\pi_E,\mu_E}\end{aligned}$$

3 Convex Reformulation via Occupation Measures

Definition 3 *We define the state-action occupation measure ν_π for any policy $\pi \in \Pi$ and corresponding state occupation measure as*

$$\nu_\pi(x,a) := \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}^{\pi,\mu_E}\left[\mathbf{1}_{\{(x_t,a_t)=(x,a)\}}\right] \qquad \nu_\pi^{\mathcal{X}}(x) := \sum_{a\in\mathcal{A}} \nu_\pi(x,a).$$

Under ergodicity, ν_π defines a valid probability measure over $\mathcal{X}\times\mathcal{A}$.

Lemma 1 *Suppose $\pi \in \Pi$ is a feasible point for the optimization problem* (OPT)*. Then:*

1. ***State Marginals:*** $\nu_\pi^{\mathcal{X}}(x) = \mu_E(x) = \sum_{(y,a)} p(x \mid y,a,\mu_E)\, \nu_\pi(y,a)$ *for all $x \in \mathcal{X}$.*
2. ***Entropy:*** $H(\pi) = \sum_{(x,a)} -\log\left(\frac{\nu_\pi(x,a)}{\mu_E(x)}\right) \nu_\pi(x,a)$.
3. ***Feature Expectation:*** $\langle f\rangle_{\pi,\mu_E} = \sum_{(x,a)} f(x,a,\mu_E)\, \nu_\pi(x,a)$.

In view of above results, let us now reformulate $(\mathbf{OPT})_0$ in terms of state-action occupation measure [3] . Although $(\mathbf{OPT})_0$ is nonconvex, it can be reformulated as an equivalent convex program over occupation measures. This new formulation is denoted by $(\widehat{\mathbf{OPT}})$, which is convex and yield the same solution as $(\mathbf{OPT})_0$:

$$\begin{aligned}(\widehat{\mathbf{OPT}}) \ \text{maximize}_\nu \quad & \textstyle\sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} -\log\left(\frac{\nu(x,a)}{\mu_E(x)}\right) \nu(x,a) \\ \text{subject to} \quad & \textstyle\sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} f(x,a,\mu_E)\, \nu(x,a) = \langle f\rangle_{\pi_E,\mu_E} \\ & \mu_E(x) = \textstyle\sum_{(y,a)\in\mathcal{X}\times\mathcal{A}} p(x \mid y,a,\mu_E)\, \nu(y,a) \ \forall x \in \mathcal{X} \\ & \nu^{\mathcal{X}}(x) = \mu_E(x) \ \forall x \in \mathcal{X} \\ & \nu(x,a) \ge 0 \ \forall (x,a) \in \mathcal{X}\times\mathcal{A}.\end{aligned}$$

The optimization problem $(\widehat{\mathbf{OPT}})$ is convex: its objective combines a strongly concave entropy term with a linear component, and all constraints are linear in the occupation measure ν. Using convex duality [4] and Sion's minimax theorem [5], the problem $(\widehat{\mathbf{OPT}})$ admits the dual formulation:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^k \boldsymbol{\beta},\boldsymbol{\theta}\in\mathbb{R}^{\mathcal{X}}} \left\{ \log \sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} e^{k_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}}(x,a)} - \langle \boldsymbol{\alpha},\ \langle f\rangle_{\pi_E,\mu_E}\rangle - \sum_{x\in\mathcal{X}} \boldsymbol{\theta}_x\, \mu_E(x) \right\}$$

Here, $k_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}}(x,a) := \langle \boldsymbol{\alpha}, f(x,a,\mu_E)\rangle + \boldsymbol{\theta}_x + \sum_{z\in\mathcal{X}} \boldsymbol{\beta}_z\, (p(z \mid x,a,\mu_E) - \mu_E(z))$. For fixed dual variables, the inner maximization over ν is solved by the Boltzmann distribution, and the corresponding recovered policy is given by:

$$\nu^*(x,a) := \frac{e^{k_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}}(x,a)}}{Z_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}}}, \qquad \pi^*(a \mid x) := \frac{\nu^*(x,a)}{\mu_E(x)}$$

which uniquely maximizes the entropy-regularized objective. This yields a fully tractable dual with no duality gap.

Theorem 1 *The dual objective h is L-smooth and $\rho(D)$-strongly convex over any compact subset $D \subset \mathbb{R}^m$, where $m = k + 2|\mathcal{X}|$. Specifically,*

$$L := 2M^2\sqrt{|\mathcal{X}|\cdot|\mathcal{A}|},$$

with M a uniform bound on the gradients of the function $k_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}}(x,a)$. Moreover, if the set

$$\{(f(x,a,\mu_E),\ p(\cdot \mid x,a,\mu_E),\ e(\cdot \mid x,a)) : (x,a) \in \mathcal{X}\times\mathcal{A}\}$$

spans $\mathbb{R}^k \times \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}}$, then h is uniformly strongly convex on D.

Now we can introduce the gradient descent algorithm for finding the minimizer of h as follows.

Algorithm 1 Gradient Descent for Parameter Optimization

Require: $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \boldsymbol{\theta}_0) \in \mathbb{R}^m$ and $\delta \in \left(0, \frac{1}{L}\right]$ and $K \in \mathbb{N}$

Set $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}) \leftarrow (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$
for $k = 0, 1, \ldots, K-1$ **do**
  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}) \leftarrow (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}) - \delta\, \nabla h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})$
end for
Compute $\nu^*_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}}$ from $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})$
return $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})$ and $\nu^*_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}}$

References

[1] Meyn, S. P. & Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer-Verlag London, Stochastic Modelling and Applied Probability.

[2] Ziebart, B. D. (2008), *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*, Ph.D. thesis, Carnegie Mellon University.

[3] Syed, U. & Schapire, R. E. (2007), A game-theoretic approach to apprenticeship learning, in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1449–1456.

[4] Dupuis, P. & Ellis, R. S. (1997), Formulation of Large Deviation Theory in Terms of the Laplace Principle. In: *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley-Interscience, ch. 1, pp. 1–47.

[5] Sion, M. (1958), On general minimax theorems, *Pacific Journal of Mathematics*, 8, pp. 171–176.