Lecture 7: Stochastic Approximation & Learning Algorithms

Last Lecture Recap

Last time: Repeated N-player finite games

- P_i at iteration k,

$$\begin{cases} w_i^k & \text{observation/ information} \\ z_i^k & \text{internal variable} \\ x_i^k & \text{mixed strategy} \end{cases} \qquad \forall\, i \in \mathcal{I}$$

- Discrete-time (DT) stochastic process:

$$\begin{cases} z_i^{k+1} = z_i^k + \gamma f_i(z_i^k, w_i^k) \\ x_i^k = \sigma(z_i^k) \end{cases}$$

- **Goal:** DT process converges to a stationary (steady) state that is an NE (or related).

Game Elements in This Lecture

Players N players, $\mathcal{I} = \{1, \dots, N\}$, playing repeatedly $k = 0, 1, 2, \dots$
Strategy Player i uses mixed strategy $\mathbf{x}_i^k \in \Delta_i$ over finite actions $\{e_{i1}, \dots, e_{im_i}\}$
Cost Stage payoff $\bar{U}_i(\mathbf{x}^k)$ or cost $\bar{J}_i(\mathbf{x}^k)$ at each iteration k

7.1 Repeated N-player finite games

Convergence analyzed by stochastic approximation via CT-ODE (mean ODE)

BR play.

$$\begin{cases} w_i^k & \text{own } \bar{U}_i(\text{or } \bar{J}_i) \text{ and } x_{-i}^k \text{ others' mixed strategy at iteration } k \\ z_i^k = x_i^k & \text{no separate internal state (compared to fictitious play)} \\ x_i^{k+1} = x_i^k + \gamma^k \left(\widetilde{\mathrm{BR}}_i(x_{-i}^k) - x_i^k \right) \end{cases}$$

where $\widetilde{\mathrm{BR}}_i(x_{-i}^k)$ is the perturbed (smoothed) BR map ($\epsilon \to 0$)

$\Rightarrow$ fixed point is NE distribution (a perturbed NE)

SA $\longrightarrow$ CT-ODE:

$$\dot{x}_i = \underbrace{\widetilde{\mathrm{BR}}_i(x_{-i}) - x_i}_{N\text{-coupled ODEs}} \quad \forall\, i$$

Fictitious Play.

$$\begin{cases} w_i^k & \text{own } \bar{U}_i(\text{or } \bar{J}_i) \text{ and } \{e_{-i}^k\}_{k \geq 0} \quad \text{set of all others' actions used before (history)} \\ z_i^k = \{\hat{x}_{-i}^k\} & \text{empirical frequencies of play for player } i' \neq i \\ \hat{x}_{i'}^{k+1} = \dfrac{1}{k+1} \displaystyle\sum_{k'=0}^{k} e_{i'}^{k'} = \hat{x}_{i'}^k + \dfrac{1}{k+1}(e_{i'}^k - \hat{x}_{i'}^k) \end{cases}$$

SA $\longrightarrow$ CT-ODE:

$$\dot{\hat{x}}_i = \widetilde{\mathrm{BR}}_i(\hat{x}_{-i}) - \hat{x}_i$$

same as BR-Play in terms of approximation ("beliefs")

7.2 Reinforcement Learning algorithms (Ch.9.4)

Information Structure. At iteration k:

$$P_i \text{ observes payoff } \pi_i^k \text{ as information } w_i^k \xrightarrow{\text{update}} \text{internal state } z_i^k \xrightarrow{\text{generate}} \text{strategy } x_i^{k+1}$$

where

- **Information:** $\pi_i^k = U_i(e_i^k, e_{-i}^k)$ realized payoff (scalar)
- **Internal state:** $z_i^k = \begin{bmatrix} \vdots \\ z_{ij^{\text{th}}\text{action}}^k \\ \vdots \end{bmatrix}_{m \text{ components in total}}$

7.2.1 Payoff-RL algorithm (Erev-Roth)

Elementwise update z_{ij}^{k+1}.

$$\begin{cases} z_{ij}^{k+1} = z_{ij}^k + \pi_i^k & \text{if action } j \text{ selected} \\ z_{ij'}^{k+1} = z_{ij'}^k & \text{else, } \forall\, j' \neq j \end{cases}$$

Vector form:

$$\boxed{1} \quad \mathbf{z}_i^{k+1} = \mathbf{z}_i^k + \pi_i^k \mathbf{e}_i^k$$

where $e_i^k = e_{ij}^k$ indicating $\begin{bmatrix} \vdots \\ 1 \\ \vdots \end{bmatrix}$ j-th action selected (realized payoff as reinforcement signal)

Elementwise update x_{ij}^k.

$$\text{Prob}(\text{action } j) = x_{ij}^k = \frac{z_{ij}^k}{\sum_{j'=1}^{m_i} z_{ij'}^k}$$

Vector form:

$$\boxed{2} \quad \mathbf{x}_i^k = \sigma(\mathbf{z}_i^k) = \frac{\mathbf{z}_i^k}{\sum_{j'=1}^{m_i} z_{ij'}^k}$$

where σ is the probability map $\sigma : \mathbb{R}^{m_i} \to \Delta_i$, $0 \leq x_{ij} \leq 1$, $\sum_j x_{ij} = 1$

From DT process to CT mean ODE.

$$\text{P-RL DT process} \begin{cases} z_i^{k+1} = z_i^k + \pi_i^k e_i^k \\ x_i^k = \sigma(z_i^k) = \dfrac{z_i^k}{\sum_{j'=1}^{m_i} z_{ij'}^k} \end{cases}$$

$$\xRightarrow{\text{SA}} \boxed{3} \quad \text{P-RL ODE} \begin{cases} \dot{z}_i = U_i(x_{-i}) \\ x_i = \sigma(z_i) \end{cases}$$

Variants.

- **Arthur P-RL:** uses step-sizes (special)
- **Coucheney:** σ is $\widetilde{\text{BR}}$ map (special)

7.2.2 Q-Learning (Leslie and Collins)

Elementwise update z_{ij}^{k+1}.

$$\begin{cases} z_{ij}^{k+1} = z_{ij}^k + \dfrac{\gamma_i^k}{x_{ij}^k}(\pi_i^k - z_{ij}^k) & \text{if action } j \text{ selected} \\ z_{ij'}^{k+1} = z_{ij'}^k & \text{else, } \forall\ j' \neq j \end{cases}$$

where $\dfrac{\gamma_i^k}{x_{ij}^k}$ is step-size with normalization, $(\pi_i^k - z_{ij}^k)$ is Temporal Difference (TD) error (realized - expected)

Vector form:

$$\boxed{4} \quad \mathbf{z}_i^{k+1} = \mathbf{z}_i^k + \gamma_i^k[\pi_i^k I - \text{diag}(\mathbf{z}_i^k)]\text{diag}(1/\mathbf{x}_i^k)\mathbf{e}_i^k$$

Elementwise update x_{ij}^k.

$$\text{Prob(action } j) = x_{ij}^k = \widetilde{\text{BR}}_i(z_{ij}^k)$$

Vector form:

$$\boxed{5} \quad \mathbf{x}_i^k = \sigma(\mathbf{z}_i^k) = \widetilde{\text{BR}}_i(\mathbf{z}_i^k)$$

where σ is softmax ("logit map"), a special perturbed BR by avg entropy-like perturbation

From DT process to CT mean ODE.

$$\text{Q-RL DT process} \begin{cases} z_i^{k+1} = z_i^k + \gamma_i^k[\pi_i^k I - \text{diag}(z_i^k)]\text{diag}(1/x_i^k)e_i^k \\ x_i^k = \widetilde{\text{BR}}_i(z_i^k) \end{cases}$$

$$\overset{\text{SA}}{\Longrightarrow} \boxed{6} \quad \text{Q-RL ODE} \begin{cases} \dot{z}_i = U_i(x_{-i}) - z_i \\ x_i = \sigma(z_i) = \widetilde{\text{BR}}_i(z_i) \end{cases}$$

Convergence based on $\boxed{5} \hookrightarrow$ zero-sum 2P / potential games

P-RL vs Q-RL.

- **P-RL:** Scores grow unbounded ($\dot{z}_i = U_i$), simple normalization σ
- **Q-RL:** Forgetting term ($-z_i$) ensures convergence to Q-values, perturbed BR σ (softmax)