Lecture 6: Stochastic Approximation & Learning Algorithms

Last Lecture Recap

Last time: Repeated N-player finite games

- P_i uses mixed strategy $\mathbf{x}_i^k$ at iteration k, depending on
  - Observation w_i^k
  - Internal variable z_i^k
  - Update rule f_i
- **Goal:** Learning process that iteratively updates z_i^k and $\mathbf{x}_i^k$

Game Elements in This Lecture

Players N players playing repeatedly at $k = 0, 1, 2, \ldots$

Strategy Use action j-th or e_{ij} with $\mathbb{P}(e_i^k = e_{ij}) = x_{ij}^k$ where $\mathbf{x}_i^k \in \Delta_i$ at iteration k

Cost $\bar{J}_i(\mathbf{x}^k)$ evaluated at each iteration

Focus on:

- Intuition (origin of alg.)
- Informational reqs
- Steady-state is at NE (or related to NE)
- Idea of convergence (CT-ODE)

6.1 Stochastic approximation - ODE method (Appx A, Ch 9)

DT - Discrete-Time Stochastic Process. A generic learning process has the form:

$$\begin{cases} \mathbf{z}_i^{k+1} = \mathbf{z}_i^k + \gamma^k f(\mathbf{z}_i^k, w_i^k) \\ \mathbf{x}_i^k = \sigma(\mathbf{z}_i^k) \end{cases}$$

where:

- $\mathbf{z}_i^k$: internal state/variable at iteration k
- $\gamma^k > 0$: step size (learning rate) at iteration k
- $f_i(\cdot, \cdot)$: update function (depends on state and observation)
- w_i^k: information/observation at iteration k
- $\sigma(\cdot)$: strategy mapping from internal state to mixed strategy

How to study the long-run behavior ($k \to \infty$)?.

Problem: Stochastic processes are complex to analyze directly!

Solution: Use **Stochastic Approximation**

- Connect DT stochastic process to CT-ODE (continuous-time ordinary differential equation)
- Analyze the simpler ODE instead

DT Stochastic Process. **General form:**

$$\mathbf{z}^{k+1} = \mathbf{z}^k + \gamma^k \left[f(\mathbf{z}^k) + \xi^k \right] \quad \longrightarrow \text{deterministic + noise}$$

where:.

1. $\{\mathbf{z}^k\}$: stochastic process
2. $\{\gamma^k\}$: diminishing step size

- $\gamma^k \geq 0$ for all k
- $\sum_{k=0}^{\infty} \gamma^k = \infty$ (infinite travel)
- $\lim_{k\to\infty} \gamma^k = 0$ (vanishing step size)

3. $\{\xi^k\}$: perturbations with **martingale difference property**

$$\mathbb{E}[\xi^k \mid \mathcal{F}_k] = 0 \quad \text{(zero-mean, conditioned on all past info)}$$

4. $f(\mathbf{z}^k)$: mean update direction

$$f(\mathbf{z}^k) = \mathbb{E}\left[\frac{1}{\gamma^k}(\mathbf{z}^{k+1} - \mathbf{z}^k) \mid \mathcal{F}_k\right]$$

Under some further assumptions, long run behaviour of the DT stochastic process can be described by the long run behaviour of the CT–ODE

$$\frac{\mathbf{z}^{k+1} - \mathbf{z}^k}{\gamma_k} \approx \dot{\mathbf{z}}(t_k) = f(\mathbf{z}(t_k)), \qquad t_k := \sum_{s=0}^{k-1} \gamma^s$$

as if DT is a perturbation of Euler discretization with variable step size.

Theorem.

Let $\bar{\mathbf{z}}$ be an **(asymptotically) stable equilibrium** for ODE.

If $\{\gamma^k\}$ goes to 0 at a **suitable rate**,

then $\{\mathbf{z}^k\}$ sequence **converges almost surely (a.s.)** to $\bar{\mathbf{z}}$.

Note: For CT-ODE analysis $\longrightarrow$ use **Linearization** and **Lyapunov theory**

6.2 Best-response and Perturbed (smooth) best-response

Problem: NE is a Fixed Point of BR, which is a set-valued map.

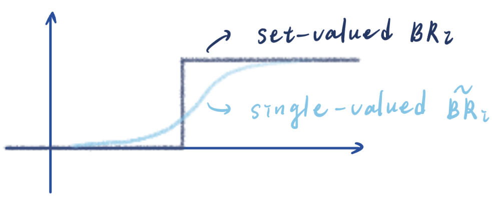
$$\boxed{1} \quad \mathbf{x}_i^* \in \mathrm{BR}_i(\mathbf{x}_{-i}^*) \quad \forall\, i \in \mathcal{I}, \quad \Leftrightarrow \quad \mathbf{x}^* \in \mathrm{BR}(\mathbf{x}^*)$$

where

$$\underbrace{\mathrm{BR}_i(\mathbf{x}_{-i})}_{\text{set-valued map}} = \arg\min_{\mathbf{x}_i \in \Delta_i} \bar{J}_i(\mathbf{x}_i, \mathbf{x}_{-i}) = \arg\max_{\mathbf{x}_i \in \Delta_i} U_i(\mathbf{x}_i, \mathbf{x}_{-i})$$

To avoid working with set-valued maps, introduce perturbed best-response:

$$\boxed{2} \quad \widetilde{\mathrm{BR}}_i(\mathbf{x}_{-i}) = \arg\max_{\mathbf{x}_i \in \Delta_i} \underbrace{\left[\bar{U}_i(\mathbf{x}_i, \mathbf{x}_{-i}) - \varepsilon v_i(\mathbf{x}_i)\right]}_{\widetilde{U}_i(\mathbf{x}_i, \mathbf{x}_{-i})}$$



- $\varepsilon > 0$ (small): perturbation
- $v_i(\mathbf{x}_i)$: **strictly convex** in $\mathbf{x}_i$
- $\widetilde{U}_i(\mathbf{x}_i, \mathbf{x}_{-i}) = \bar{U}_i(\mathbf{x}_i, \mathbf{x}_{-i}) - \varepsilon V_i(\mathbf{x}_i)$

Smooth / perturbed best response.

$\widetilde{\mathrm{BR}}_i$ works like a softmax function:

it transforms the set-valued fixed-point condition $\boxed{1}$ into a single-valued fixed point $\boxed{2}$

$$\boxed{1} \quad \mathbf{x}^* \in \mathrm{BR}(\mathbf{x}^*) \qquad \Leftrightarrow \qquad \boxed{2} \quad \mathbf{x}^* = \widetilde{\mathrm{BR}}(\mathbf{x}^*)$$

NE distribution.

Let $\mathbf{x}^*(\varepsilon)$ be a Nash equilibrium of the perturbed game, i.e.

$$\mathbf{x}_i^*(\varepsilon) = \widetilde{\mathrm{BR}}_i\big(\mathbf{x}_{-i}^*(\varepsilon)\big).$$

As $\varepsilon \to 0$, the NE distribution $\mathbf{x}^*(\varepsilon)$ converges to a Nash equilibrium $\mathbf{x}^{NE}$ of the original (unperturbed) game:

$$\mathbf{x}^*(\varepsilon) \longrightarrow \mathbf{x}^{NE}.$$

6.3 Two iterative algorithms for Repeated Finite Games:

Iterative Algorithm Design for repeated games.

- Intuition / origin
- Information: w_i^k
- Goal I: steady state of the algorithm is an NE (or closely related to an NE)
- Goal II: convergence to an NE

Question: Given that NE are fixed points of perturbed $\widetilde{\mathrm{BR}}$, how to design a learning/update rule from $\mathbf{x}^k$ to $\mathbf{x}^{k+1}$ that will converge to such a fixed point?

Key idea: If $\mathbf{x}^*$ is a fixed point of a map F (e.g. BR, $\widetilde{\mathrm{BR}}$), then a relaxation step with step size α_k moves you along the segment from $\mathbf{x}^k$ to $\mathbf{x}^*$:

$$\mathbf{x}^{k+1} = (1-\alpha_k)\mathbf{x}^k + \alpha_k \mathbf{x}^*$$
$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k(\mathbf{x}^* - \mathbf{x}^k)$$

Use this intuition to get iterative algorithms for $\mathbf{x}^* = f(\mathbf{x}^*)$ to get iterative algorithms based on

- $\boxed{1}$ Best-Response Play
- $\boxed{2}$ Perturbed Best-Response Play

6.3.1 Best-Response Play and its variants (Ch 9.2)

BR Play Algorithm.

$$\begin{aligned}
&\boxed{1} && \mathbf{x}_i^* \in \mathrm{BR}_i(\mathbf{x}_{-i}^*) \\
&\Rightarrow && \mathbf{x}_i^{k+1} \in \mathbf{x}_i^k + \gamma^k\big(\mathrm{BR}_i(\mathbf{x}_{-i}^k) - \mathbf{x}_i^k\big) \\
\text{(ODE–SA)} \ &\Rightarrow && \dot{\mathbf{x}}_i \in \underbrace{\mathrm{BR}_i(\mathbf{x}_{-i}) - \mathbf{x}_i}_{\text{(set-valued CT dynamics)}} \quad \boxed{1^*}
\end{aligned}$$

Perturbed / Smooth BR Play Algorithm.

$$
\begin{aligned}
\boxed{2} \quad \mathbf{x}_i^* &= \widetilde{\mathrm{BR}}_i(\mathbf{x}_{-i}^*) = \arg\max_{\mathbf{x}_i \in \Delta_i} \left[\bar{U}_i(\mathbf{x}_i, \mathbf{x}_{-i}) - \varepsilon\, \nu_i(\mathbf{x}_i) \right] \\
\Rightarrow \quad \mathbf{x}_i^{k+1} &= \mathbf{x}_i^k + \gamma^k \left(\widetilde{\mathrm{BR}}_i(\mathbf{x}_{-i}^k) - \mathbf{x}_i^k \right) \\
(\text{ODE–SA}) \quad \Rightarrow \quad \dot{\mathbf{x}}_i &= \underbrace{\widetilde{\mathrm{BR}}_i(\mathbf{x}_{-i}) - \mathbf{x}_i}_{\text{single-valued CT dynamics}} \qquad \boxed{2^*}:
\end{aligned}
$$

Note: This ODE is **coupled** across players. e.g. when $N = 2$, $\mathcal{I} = \{1, 2\}$

$$
\begin{cases}
\dot{\mathbf{x}}_1 = \widetilde{\mathrm{BR}}_1(\mathbf{x}_2) - \mathbf{x}_1, \\
\dot{\mathbf{x}}_2 = \widetilde{\mathrm{BR}}_2(\mathbf{x}_1) - \mathbf{x}_2.
\end{cases}
$$

Steady-State Value of the Algorithm.

$$
\mathbf{x}_i^k = \bar{\mathbf{x}}_i = \widetilde{\mathrm{BR}}_i(\bar{\mathbf{x}}_{-i}), \qquad \forall\, i.
$$

DT view from $\boxed{1^*}$

$$
\mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \gamma_k \left(\widetilde{\mathrm{BR}}_i(\mathbf{x}_{-i}^k) - \mathbf{x}_i^k \right).
$$

At steady state, $\mathbf{x}_i^k = \bar{\mathbf{x}}_i$ and $\mathbf{x}_{-i}^k = \bar{\mathbf{x}}_{-i}$, so

$$
\bar{\mathbf{x}}_i = \bar{\mathbf{x}}_i + \gamma_k \left(\widetilde{\mathrm{BR}}_i(\bar{\mathbf{x}}_{-i}) - \bar{\mathbf{x}}_i \right).
$$

CT view from $\boxed{2^*}$:

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

At equilibrium $\dot{\mathbf{x}}_i = 0$:

$$
0 = \widetilde{\mathrm{BR}}_i(\bar{\mathbf{x}}_{-i}) - \bar{\mathbf{x}}_i \quad \Longleftrightarrow \quad \bar{\mathbf{x}} \text{ is an NE distribution.}
$$

6.3.2 Fictitious Play and its Variants (Ch 9.3)

Idea of Fictitious Play.

Instead of knowing others' mixed strategies (as in BR play):

$$
\mathbf{x}_{-i}^k = \{\mathbf{x}_{i'}^k,\ i' \neq i\}
$$

player P_i will approximate them using empirical averages (frequencies) of play.

Strategy level. Instead of knowing the whole vector

$$
\mathbf{x}_{-i}^k = \{\mathbf{x}_{i'}^k,\ i' \neq i\},
$$

player P_i approximates the mixed strategy of every $i' \neq i$ from the empirical average (frequency) of past actions.

Action level. Instead of the probabilities $x_{i'j}^k = \Pr(P_{i'}$ selects action j at iteration $k)$, player P_i uses the empirical average (frequency) of how many times player $P_{i'}$ has used action j up to iteration k.

$$
\Rightarrow \quad P_i \text{ denotes this approximation by } \hat{\mathbf{x}}_{i'}^k \approx \mathbf{x}_{i'}^k, \qquad \forall\, i' \neq i.
$$

Empirical Mixed Strategy (Information w_i^k).

Information available to P_i at iteration k:

$$w_i^k = \{e_{i'}^t : i' \neq i,\ 0 \leq t \leq k\} \quad \text{(history of all other players' actions)}.$$

For each $i' \neq i$, define the empirical mixed strategy at iteration $k+1$:

$$\boxed{3} \qquad \hat{\mathbf{x}}_{i'}^{k+1} = \frac{1}{k+1} \sum_{t=0}^{k} e_{i'}^t,$$

where $e_{i'}^t$ is the unit-vector of action j of player i' at iteration t.

write it recursively so we don't need to track the whole history:

$$\boxed{4} \qquad \hat{\mathbf{x}}_{i'}^{k+1} = \hat{\mathbf{x}}_{i'}^k + \frac{1}{k+1}\left(e_{i'}^k - \hat{\mathbf{x}}_{i'}^k\right), \qquad \forall\, i' \neq i.$$

Now P_i use $\{\hat{\mathbf{x}}_{i'}^k\}$ as its internal variable $\mathbf{z}_i^k$ (beliefs about others):

$$\{\hat{\mathbf{x}}_{i'}^k\}_{i' \neq i} =: \hat{\mathbf{x}}_{-i}^k =: \mathbf{z}_i^k$$

Fictitious Play Algorithm for Player i.

P_i plays a (perturbed) best response to the fictitious/approximated mixed strategy profile $\hat{\mathbf{x}}_{-i}^k$.

Internal processing (belief update from $\boxed{4}$):

$$\hat{\mathbf{x}}_{i'}^{k+1} = \hat{\mathbf{x}}_{i'}^k + \frac{1}{k+1}\left(e_{i'}^k - \hat{\mathbf{x}}_{i'}^k\right), \qquad \forall\, i' \neq i.$$

Strategy update (best response to beliefs):

$$\mathbf{x}_i^k = \widetilde{\mathrm{BR}}_i(\hat{\mathbf{x}}_{-i}^k), \qquad k = 0, 1, 2, \ldots$$

All Other Players Also Use Fictitious Play.

Assume every player i' uses the Fictitious Play strategy update above. Then, at iteration k, the probability of selecting action j

$$\boxed{5} \quad \mathbb{P}\left(e_{i'}^k = e_{i'j} \mid \hat{\mathbf{x}}^k\right) = x_{i'j}^k$$

The conditional expectation of the action vector is

$$\begin{aligned} \boxed{6} \quad \mathbb{E}\left[e_{i'}^k \mid \hat{\mathbf{x}}^k\right] &= \sum_j e_{i'j}\, \mathbb{P}\left(e_{i'}^k = e_{i'j} \mid \hat{\mathbf{x}}^k\right) \\ &= \sum_j x_{i'j}^k e_{i'j} = \mathbf{x}_{i'}^k = \widetilde{\mathrm{BR}}_{i'}(\hat{\mathbf{x}}_{-i'}^k). \end{aligned}$$

Recall: stochastic approximation.

$$\mathbf{z}^{k+1} = \mathbf{z}^k + \gamma^k\big(f(\mathbf{z}^k) + \xi^k\big) \quad \text{where} \quad \begin{cases} \gamma^k > 0,\ \sum_k \gamma^k = \infty,\ \gamma^k \to 0, \\ \mathbb{E}[\xi^k \mid \mathcal{F}_k] = 0. \end{cases}$$

$$\underbrace{\mathbb{E}\left[\frac{\mathbf{z}^{k+1} - \mathbf{z}^k}{\gamma^k} \,\middle|\, \mathcal{F}_k\right]}_{\text{compute it for } \boxed{4}} = f(\mathbf{z}^k) \quad \rightsquigarrow \quad \dot{\mathbf{z}} = f(\mathbf{z}).$$

Let

$$\gamma^k = \frac{1}{k+1}, \qquad \mathbf{z}^k = \hat{\mathbf{x}}^k,$$

and recall the belief update $\boxed{4}$ (for each player i):

$$\hat{\mathbf{x}}_i^{k+1} = \hat{\mathbf{x}}_i^k + \frac{1}{k+1}\big(e_i^k - \hat{\mathbf{x}}_i^k\big).$$

Then

$$\frac{\hat{\mathbf{x}}_i^{k+1} - \hat{\mathbf{x}}_i^k}{\gamma^k} = e_i^k - \hat{\mathbf{x}}_i^k.$$

Taking conditional expectation (given all past info, summarized by $\hat{\mathbf{x}}^k$):

$$\begin{aligned}
\mathbb{E}\left[\frac{\hat{\mathbf{x}}_i^{k+1} - \hat{\mathbf{x}}_i^k}{\gamma^k} \,\middle|\, \hat{\mathbf{x}}^k\right] &= \mathbb{E}[e_i^k \mid \hat{\mathbf{x}}^k] - \hat{\mathbf{x}}_i^k \\
&= \underbrace{\mathbb{E}[e_i^k \mid \hat{\mathbf{x}}^k]}_{\boxed{6}\ = \mathbf{x}_i^k = \widetilde{\mathrm{BR}}_i(\hat{\mathbf{x}}_{-i}^k)} - \hat{\mathbf{x}}_i^k \\
&= \widetilde{\mathrm{BR}}_i(\hat{\mathbf{x}}_{-i}^k) - \hat{\mathbf{x}}_i^k \;=:\; f_i(\hat{\mathbf{x}}^k).
\end{aligned}$$

By the SA–ODE correspondence,

$$\dot{\hat{\mathbf{x}}}_i = \widetilde{\mathrm{BR}}_i(\hat{\mathbf{x}}_{-i}) - \hat{\mathbf{x}}_i, \qquad \forall\, i.$$

Comparison with BR Play.

Perturbed BR Play:

$$\dot{\mathbf{x}}_i = \widetilde{\mathrm{BR}}_i(\mathbf{x}_{-i}) - \mathbf{x}_i, \qquad \forall\, i.$$

Fictitious Play:

$$\dot{\hat{\mathbf{x}}}_i = \widetilde{\mathrm{BR}}_i(\hat{\mathbf{x}}_{-i}) - \hat{\mathbf{x}}_i, \qquad \forall\, i.$$

- BR Play: state $\mathbf{x}_i$ = actual mixed strategy.
- FP: state $\hat{\mathbf{x}}_i$ = belief / empirical frequency.

In both cases the CT–ODE is a relaxation toward a fixed point $\mathbf{x}^* = \widetilde{\mathrm{BR}}(\mathbf{x}^*)$, i.e. a (perturbed) NE, but BR play moves the strategies, while FP moves the beliefs.