Feedback Control for Learning in Games

Gurdal ARSLAN & Jeff SHAMMA

Mechanical and Aerospace Engineering

UCLA
Setup: Repeated Games

- Time $k = 1, 2, 3, \ldots$
- Player $i$:
  - Strategy: $p_i(k) \in \Delta$
  - Action: $a_i(k) = \text{rand}[p_i(k)]$
  - Payoff: $U_i(a_i, a_{-i}) \sim a_i^\top M_i a_{-i}$
  - Play: $p_i(k) = F(\text{information up to time } k)$
- Assume players do not share utilities!

How can simple rules lead players to mixed strategy Nash equilibrium?

- Separate issues: Will they? should they? compute NE?
Prior Work & Convergence

- (Stochastic) Fictitious Play
- No Regret
- New approaches: Multirate, Joint weak calibration, Regret testing, ...
- Convergence results:
  - Special cases: NE
  - Correlated equilibria
  - Convex hull of NE
  - “Dwell” near NE
Non-convergence Results

• Shapley game vs Fictitious Play

• Crawford (1985): *wide class of learning mechanisms must fail to converge mixed strategies*

• Jordan anticoordination game: 3 players, each with 2 moves.

• Hart & Mas-Colell (2003): Consider larger class & show

  Uncoupled + Jordan anticoordination = non-convergence
Preview

• Introduce new uncoupled dynamics based on “feedback control”.

• Demonstrate how convergence to mixed strategy NE can be enabled (including Shapley & Jordan games).

• Best/Better response variants.

• Action/Payoff based versions.

• Two/Multi-player cases.
Feedback Control

- \( K = \) controller = sequential decision maker
- \( P = \) process with approximate model \( P_{\text{model}} \)
- Think of “standing upright”
What’s the Connection?

• FB $\rightarrow$ GT: \[
\max_{K \in \mathcal{K}} \min_{d \in \mathcal{D}} \ \text{Performance}(P, K)
\]

- New initiatives in “cooperative control” (combat systems, networks, self-assembly, automata teams...) require general sum formulation.

• GT $\rightarrow$ FB:

$DM_i$ is in feedback with $DM_j$
Typical Controller: PID

• Proportional + Integral + Derivative

\[ u(t) = K_p e(t) + K_i \int_0^t e(s) \, ds + K_d \frac{de(t)}{dt} \]

- \( K_p \Rightarrow \) current error
- \( K_i \Rightarrow \) error history
- \( K_d \Rightarrow \) error change

• “Workhorse” of traditional control design.
• Model of human motion control, homeostasis, …
Derivative Action

\[ u(t) = -K_p(e(t + \tau)) \approx -K_p(e(t) + \tau \frac{de(t)}{dt}) \]

• React to predicted error

• Example: “Balancing”:
  \[ \ddot{y} = u \]

  \[ u = -K_P y \Rightarrow \text{oscillations} \]

  \[ u = -(K_P y + K_D \dot{y}) \Rightarrow \text{convergence} \]
Repeated Games in Continuous Time

- Empirical frequencies:

\[ q_i(k + 1) = q_i(k) + \frac{1}{k + 1} (\text{rand}(p_i(k)) - q_i(k)) \]

\[ \downarrow \]

\[ \frac{d}{dt} q_i(t) = -q_i(t) + p_i(t) \]

- ODE method of stochastic approximation:

   *Deterministic continuous time analysis*

   \[ \downarrow \]

   *Probabilistic discrete time conclusions*
Derivative Action FP (DAFP)

• Define smoothed best response:

\[ \beta(v) = \arg \min_{s \in \Delta} s^T M v + \tau \mathcal{H}(s) \]

• FP:

\[ p_i(t) = \beta_i(q_{-i}(t)) \]

• Derivative action FP:

\[ p_i(t) = \beta_i(q_{-i}(t) + \gamma \frac{dq_{-i}}{dt}(t)) \]

• “First order” model of adversary: Moving target.
Ideal vs Approximate

- Ideal $\Rightarrow$ Implicit Equations
  \[ \dot{q}_1 = -q_1 + \beta_1 (q_2 + \gamma \dot{q}_2) \]
  \[ \dot{q}_2 = -q_2 + \beta_2 (q_1 + \gamma \dot{q}_1) \]

- Approximate:
  \[ \dot{q}_1 = -q_1 + \beta_1 (q_2 + \gamma \dot{q}_2^{\text{est}}) \]
  \[ \dot{q}_2 = -q_2 + \beta_2 (q_1 + \gamma \dot{q}_1^{\text{est}}) \]

- Use of ideal differentiators can always lead to NE (a misleading conclusion).
Approximate Differentiator

- Define:
  \[ \frac{dr}{dt} = \lambda(q - r), \quad \lambda \gg 1 \]

- Asymptotically
  \[ \left| \frac{dq}{dt} - \frac{dr}{dt} \right| \leq \frac{1}{\lambda} \left| \frac{d^2q}{dt^2} \right|_{\text{max}} \]

- Two-player implementation:

  \[
  \begin{align*}
  \dot{q}_1 &= -q_1 + \beta_1(q_2 + \gamma \dot{r}_2) \\
  \dot{q}_2 &= -q_2 + \beta_2(q_1 + \gamma \dot{r}_1) \\
  \dot{r}_1 &= \lambda(q_1 - r_1) \\
  \dot{r}_2 &= \lambda(q_2 - r_1)
  \end{align*}
  \]
Local Convergence of DAFP

• Theorem: Consider a two-player game with a NE $q^*$.

1) \[(FP) \text{ stable at } q^* \Rightarrow (DAFP, \text{ with any } \gamma < 1) \text{ stable at } q^* \]

2) \[(FP) \text{ unstable at } q^*, \text{ with } \max_i \frac{a_i}{a_i^2 + b_i^2} < \frac{1}{\max_i a_i} \Rightarrow (DAFP, \text{ with some } \gamma < 1) \text{ stable at } q^* \]

where $a_i + jb_i$ are the eigenvalues of linearized (FP)
Jordan Anticoordination Revisited

• Unique mixed NE is unstable under \( (FP) \)
• \( \max_i \frac{a_i}{a_i^2 + b_i^2} < \frac{1}{\max_i a_i} \), hence stabilizable by \( (DAFP) \)
Extensions to “Gradient Play”

• “Better Response” = GP
  \[ \dot{q}_1(t) = -q_1(t) + \nabla \Delta [q_1(t) + M_1 q_2(t)] \]
  \[ \dot{q}_2(t) = -q_2(t) + \nabla \Delta [q_2(t) + M_2 q_1(t)] \]

• DAGP :
  \[ \dot{q}_1(t) = -q_1(t) + \nabla \Delta [q_1(t) + M_1 (q_2(t) + \gamma \dot{r}_2(t))] \]
  \[ \dot{q}_2(t) = -q_2(t) + \nabla \Delta [q_2(t) + M_2 (q_1(t) + \gamma \dot{r}_1(t))] \]

• Theorem: Similar … using eigenvalues of
  \[ \mathcal{N}^T \begin{pmatrix} 0 & M_1 \\ M_2 & 0 \end{pmatrix} \mathcal{N} \]

• Shapley & Jordan games convergent.
Crawford & Conlisk


  - Two-player zero-sum games
  - Play in “rounds” (..., R-1, R, R+1, ...)
  - On R+1 use adjust mixed strategy with “forecast” payoff based on intervals R & R-1
$$q_i(k + 1) = q_i(k) + \frac{1}{k + 1} \left( \text{rand}(p_i(k)) - q_i(k) \right)$$

$$r_i(k + 1) = r_i(k) + \frac{1}{k + 1} \lambda (q_i(k) - r_i(k))$$

with...

$$p_i(k) = \beta_i \left( q_{-i}(k) + \gamma \lambda (q_{-i}(k) - r_{-i}(k)) \right)$$

—or—

$$p_i(k) = \prod_{\Delta \varepsilon} \left( q_i(k) + M_i \left( q_{-i}(k) + \gamma \lambda (q_{-i}(k) - r_{-i}(k)) \right) \right)$$

• **Theorem**: Local attractor in continuous time $\Rightarrow$ Positive probability of convergence to NE in discrete-time.

• ...as opposed to **Zero** probability.
Payoff Based Rules

• Use “stimulus response”

\[
\begin{align*}
\hat{U}_i^\ell(k+1) &= \begin{cases} 
\hat{U}_i^\ell(k) + \frac{1}{k+1} \frac{1}{p_i^\ell(k)} (U_i(k) - \hat{U}_i^\ell(k)), & \text{if } a_i(k) = \ell; \\
\hat{U}_i^\ell(k), & \text{otherwise.}
\end{cases}
\end{align*}
\]

\[
W_i(k + 1) = W_i(k) + \frac{\lambda}{k + 1} (\hat{U}_i(k) - W_i(k))
\]

\[
p_i(k) = \text{softmax} \left( \frac{1}{\gamma}(\hat{U}_i(k) + \gamma \lambda (\hat{U}_i(k) - W_i(k))) \right)
\]

—or—

\[
p_i(k) = \prod_{\Delta \epsilon} (q_i(k) + \hat{U}_i(k) + \gamma \lambda (\hat{U}_i(k) - W_i(k)))
\]

• **Theorem**: Positive probability of convergence to NE.
Jordan Anticoordination: Payoff Based DAGP

\[ \gamma = 1, \lambda = 50, \varepsilon = 0.1 \]
Multiplayer Games

• Immediate extensions in case of “pair-wise utility” structure:

\[ U_i(p_i, p_{-i}) = \sum_{j \neq i} p_i^T M_{ij} p_j \]

• Otherwise, must inspect “joint-action” version of FP.
Concluding Remarks

• Feedback control motivates the use of auxiliary dynamics to enable NE convergence.

• Other “controller” structures possible (all mixed strategy equilibria “stabilizable”)

• DAFP & DAGP respect “graph” structures.

• Key concerns:
  – Natural?
  – Strategic?