Learning: Reinforcement, Fictitious Play and EWA
學習理論：制約、計牌與EWA

Joseph Tao-yi Wang (王道一)

Lecture 9, EE-BGT

Outline: Estimating Learning (Econometrics, Ch. 18)

1. Directional Learning (DL): Selten and Stoecker (1986)
2. Reinforcement Learning (RL)
3. Belief Learning (BL)
4. EWA Learning: Camerer and Ho (ECMA 1999)
   - Experience-Weighted Attraction – a Hybrid of RL and BL

Directional Learning Theory: Selten and Stoecker (1986)

- Subjects adjust their behavior in response to previous outcome
  - Finitely Repeated Prisoner's Dilemma (PD)
  - SPE: Always Defect
- Stylized Facts
  - Tacit Cooperation Until Close to End
  - Want to Defect 1st (then Keep Defect)
- Decision: Which Round to Defect

	C	D
C	2, 2	0, 3
D	3, 0	1, 1

Directional Learning Theory: Selten and Stoecker (1986)

- Play N Supergames with a different opponent each time
  - Adjust next intended deviation period:
- If Deviated First:
  - May gain if deviated later
- If Deviated Later:
  - May gain if deviate early
- If Deviate in the Same Round:
  - May gain if deviate 1 period earlier

	C	D
C	2, 2	0, 3
D	3, 0	1, 1

The Data: Table B1 of Selten and Stoecker (1986)

- n=35 subjects play 25 supergames (of 10-round PD)
  - Play the same opponent within 10 rounds of PD, but
  - Randomly rematch in between: `selten-stoecker.dta`
- Intended Deviation Period of each supergame: `self`
  - `self/other` = 1-10 (period)
  - `self/other` = 11 (later than opponent, but unobserved)
  - `self/other` = 12 (never deviate)
- Deviate `before/same/after` their opponent

Simple Linear Regression

- Predict difference in `self` with `before/same/after`
  - `d.self` - Difference in `self`
  - `l.before` - Lagged `before`
  - `l.same` - Lagged `same`
  - `l.after` - Lagged `after`
- STATA Command:

```
xtset i t
regress d.self l.before l.same l.after, nocon
```

No constant term

Results

```
. xtset i t
       panel variable: i (strongly balanced)
        time variable: t, 1 to 25
                delta: 1 unit

. regress d.self l.before l.same l.after, nocon
      Source |       SS       df       MS              Number of obs = 528
-------------+------------------------------           F( 3, 525) = 29.36
       Model |  93.535929     3   31.178643           Prob > F = 0.0000
    Residual |  557.464071   525  1.06183633           R-squared = 0.1437
-------------+------------------------------           Adj R-squared = 0.1388
       Total |        651   528  1.23295455           Root MSE = 1.0305
                                                       Number of obs = 1000

------------------------------------------------------------------------------
      D.self |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      before |
         L1. |   .3645833   .0743666     4.90   0.000     .2184906    .5106761
             |
        same |
         L1. |  -.1626506   .0799788    -2.03   0.042    -.3197683   -.0055329
             |
       after |
         L1. |  -.6117647   .0790322    -7.74   0.000    -.767023    -.4565064
------------------------------------------------------------------------------
```

Players deviate 0.36 periods later if before

Players deviate 0.16 periods earlier if same

Players deviate 0.61 periods earlier if after

Pursue-Evade Game (Rosenthal et al. 2003)

- Data: 100 pairs of 50 rounds
  `pursue_evade_sim.dta`
- Payoff Table

	L	**R**
L	1, -1	0, 0
R	0, 0	2, -2

- Player 1 - Pursuer: **L** (left) or **R** (right)
  - $\mathrm{y1} = 0$ if Pursuer choose **L**; $\mathrm{y1} = 1$ if Pursuer choose **R**
- Player 2 - Evader: **L** (left) or **R** (right)
  - $\mathrm{y2} = 0$ if Evader choose **L**; $\mathrm{y2} = 1$ if Evader choose **R**

Pursue-Evade Game (Rosenthal et al. 2003)

- Two Players: $i = 1, 2$
- Rounds: $t = 1, 2, \ldots, \mathrm{T} = 50$
- Two Actions: $s_i^0 = \mathbf{L}$, $s_i^1 = \mathbf{R}$
- Relabel as Actions $j = 0$ (**L**) and $j = 1$ (**R**)
- Strategy of Players i in round t is $s_i(t)$
  - Strategy of Players $-i$ in round t is $s_{-i}(t)$
- Players i's Payoff in round t is $\pi_i(s_i(t),\ s_{-i}(t))$

	L	R
L	1, -1	0, 0
R	0, 0	2, -2

Learning

- Attraction to action $j = 0, 1$ after round t is $A_i^j(t)$
- Initial Attractions to action $j = 0, 1$ is $A_i^j(0)$
  - Normalize one of initial attractions to 0 for each player
- Choice Probability obtained by logistic transformation

$$P_i^j(t) = \frac{\exp\left[\lambda A_i^j(t-1)\right]}{\exp\left[\lambda A_1^j(t-1)\right] + \exp\left[\lambda A_0^j(t-1)\right]}$$

□Irrelevant ($\lambda = 0$)
□Important (λ large)

  - $i = 1, 2$; $j = 0, 1$; $t = 1, 2, \ldots, \mathrm{T}$; $\lambda =$ Sensitivity to attractions

Reinforcement Learning (RL): Erev and Roth (1998)

- Update attractions in response to previous payoffs
  - Choices "reinforced" only by previous payoffs

$$\underline{\underline{A_i^j(t)}} = \phi \underline{\underline{A_i^j(t-1)}} + I(s_i(t) = s_i^j)\boxed{\pi_i(s_i^j, s_{-i}(t))}$$

  - $i = 1,\ 2;\ j = 0,\ 1;\ t = 1,\ 2, \ldots, \mathrm{T}$
- Recency parameter:
  - $\phi = 0$: Only most recent payoff is remembered
  - $\phi = 1$: All past payoffs have equal weight

Reinforcement Learning (RL): Erev and Roth (1998)

- Normalize Initial Attractions $\mathrm{A}_1{}^1(0) = 0$, $\mathrm{A}_2{}^1(0) = 0$
- Estimate Initial Attractions $A_1{}^0(0)$, $A_2{}^0(0)$, as well as
- Recency parameter ϕ and Sensitivity parameter λ
  - In STATA using Maximum Likelihood
  - (See code in package)

Results of Reinforcement Learning (RL)

Number of obs = 5000
Wald chi2(0) = .
Log likelihood = -6863.0929
Prob > chi2 = .

	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
phi _cons	$\boxed{.7676348}$	.0845556	9.08	0.000	.601909	.9333607
lam _cons	$\boxed{.1095563}$	.0156976	6.98	0.000	.0787895	.1403231
A10_start _cons	$\boxed{1.389702}$	1.457448	0.95	0.340	-1.466843	4.246247
A20_start _cons	$\boxed{-8.310659}$	2.286308	-3.63	0.000	-12.79174	-3.829578

Recency ϕ far from 1: Forget past quickly, significant

Sensitivity λ > 0: Attractions matter, significant

$A_1^0(0) > 0$: Pursuers start at **L**, not significant

$A_2^0(0) < 0$: Evaders start at **R**, significant

Simple Belief Learning (RL): Cournot Learning

- Attractions increase by action corresponding payoffs given opponent actions
  - BR to opponent action in previous round

$$\underline{\underline{A_i^j(t)}} = \underline{\underline{A_i^j(t-1)}} + \boxed{\pi_i(s_i^j, s_{-i}(t))}$$

  - $i = 1, 2; j = 0, 1; t = 1, 2, \ldots, \mathrm{T}$
  - Normalize Initial Attractions $A_1{}^1(0) = 0$, $A_2{}^1(0) = 0$
  - Only need to estimate Initial Attractions $A_1{}^0(0)$, $A_2{}^0(0)$ and λ using Maximum Likelihood (Too simple?!)

Belief Learning (RL): Standard Fictitious Play

- Attractions is action-corresponding average payoffs
  - Counting cards and BR to opponent actions from all rounds
- All Initial Attractions are zero: $A_i^j(0) = 0$, $j = 0, 1$

$$A_i^j(1) = \pi_i(s_i^j, s_{-i}(1)),\ A_i^j(2) = \frac{1}{2}\left[\pi_i(s_i^j, s_{-i}(1)) + \pi_i(s_i^j, s_{-i}(2))\right]$$

$$A_i^j(3) = \frac{1}{3}\left[\pi_i(s_i^j, s_{-i}(1)) + \pi_i(s_i^j, s_{-i}(2)) + \pi_i(s_i^j, s_{-i}(3))\right]$$

- …, $A_i^j(t) = \frac{1}{t}\sum_{\tau=1}^{t} \pi_i(s_i^j, s_{-i}(\tau))$

Belief Learning (RL): Experience Weight

- Express attractions based on Experience $N(t)$
  - Observation Equivalents: Experience accumulated up to t
- Initial Experience is zero: $N(0) = 0$
- Iteratively define $N(t) = N(t-1) + 1,\ t = 1,\ldots,T$
- All Initial Attractions are zero: $A_i^j(0) = 0\ ,\ j = 0,\ 1$
- Iteratively define (for $j = 0,\ 1;\ t = 1,\ldots,T$)

$$A_i^j(t) = \frac{1}{N(t)}\left[N(t-1)A_i^j(t-1) + \pi_i(s_i^j, s_{-i}(t))\right]$$

  - Special Case of $N(t) = t$ is Standard Fictitious Play!

Belief Learning (RL): Weighted Fictitious Play

- Another Special Case is Weighted Fictitious Play
  - With Recency parameter ϕ
- Initial Experience is zero: $N(0) = 0$
- Iteratively define $N(t) = \phi N(t-1) + 1,\ t = 1,\ldots,T$
- All Initial Attractions are zero: $A_i^j(0) = 0\ ,\ j = 0,\ 1$
- Iteratively define (for $j = 0,\ 1;\ t = 1,\ldots,T$)

$$A_i^j(t) = \frac{1}{N(t)}\left[\phi N(t-1) A_i^j(t-1) + \pi_i(s_i^j, s_{-i}(t))\right]$$

- Weights are 1, ϕ, ϕ^2, ϕ^3, etc.

Belief Learning (RL): Weighted Fictitious Play

- Attractions is action-corresponding average payoffs weighted by recency (exponentially discounted)
- All Initial Attractions are zero: $A_i^j(0) = 0$, $j = 0,\ 1$

$$A_i^j(1) = \pi_i(s_i^j, s_{-i}(1)),$$

$$A_i^j(2) = \frac{1}{\phi + 1}\left[\phi\pi_i(s_i^j, s_{-i}(1)) + \pi_i(s_i^j, s_{-i}(2))\right]$$

$$A_i^j(3) = \frac{\phi^2\pi_i(s_i^j, s_{-i}(1)) + \phi\pi_i(s_i^j, s_{-i}(2)) + \pi_i(s_i^j, s_{-i}(3))}{\phi^2 + \phi + 1}, \text{ etc.}$$

Belief Learning (RL): Weighted Fictitious Play

- In general, initial attractions and $N(0)$ need not be zero
  - Normalize Initial Attractions $A_1^1(0) = 0$, $A_2^1(0) = 0$
  - And:
- Estimate Initial Attractions $A_1^0(0)$, $A_2^0(0)$, $N(0)$ as well as Recency parameter ϕ and Sensitivity parameter λ
  - In STATA using Maximum Likelihood (See code in package)
- Standard Fictitious Play if $\phi = 1$
- Cournot Learning if $\phi = 0$

Results of Belief Learning (Weighted Fictitious Play)

Log likelihood = -6808.3011 Prob > chi2 = .

Recency ϕ away from 0 and 1: Neither Cournot nor standard fictitious play

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
phi						
_cons	.9531451	.0188062	50.68	0.000	.9162856	.9900046
lambda						
_cons	.424634	.0355367	11.95	0.000	.3549833	.4942846
A10_start						
_cons	1.684257	.4686631	3.59	0.000	.7656939	2.602819
A20_start						
_cons	-4.255275	.7702032	-5.52	0.000	-5.764845	-2.745704
N_start						
_cons	.4723948	.1498293	3.15	0.002	.1787348	.7660549

Sensitivity $\lambda > 0$: Attractions matter, significant

${A^0}_1(0) > 0$: Pursuers start at **L**, significant

${A^0}_2(0) < 0$: Evaders start at **R**, significant

Start with 0.5 observation equivalents of experience