BAYESIAN LEARNING AND EQUILIBRIUM IN GAMES

Drew Fudenberg

Summer School: Strategies and Dynamic in Games

Lima, January 2019

Outline

1. Overview: Learning and Equilibrium
2. Fictitous Play: A Special Kind of Bayesian Learning
3. Bayesian Learning in Extensive-Form Games
4. (Time permitting) Other learning models

Overview

- When and why will observed play approximate an equilibrium?
- What sort of equilibrium?
- Rationality, even common knowledge of rationality, is neither necessary nor sufficient for Nash equilibrium (NE).

Not sufficient:

- Only sufficient in games where the iterated deletion of dominated strategies yields a unique outcome.
- With multiple NE, no reason for play to look like any of the equilibria without some explanation for why players all expect the same equilibrium.

- Rationality is also not necessary either in theory (replicator dynamic can converge to NE) or data (convergence to approximation of NE seen in colonies of bacteria.)

- Yet equilibrium seems a decent approximation of the outcomes of some (not all!) experiments, and has been useful in empirical analyses of field data.

- To understand Nash equilibrium and related solution concepts, study the long-run behavior of non-equilibrium dynamic processes, and ask when they converge to equilibrium.

- Many sorts of adjustment processes, including biological evolution, have been said to involve "learning" in a broad sense.

- Today: narrower sense of "learning" in strategic-form games, where the models explicitly specify how individual agents use observations to change their behavior in a way I'll call "learning."

- I won't try to define "learning models": is a best response to yesterday's play "learning" or not?

Common themes in the learning-in-games literature:

Non-equilibrium adjustment.

- It is pointless to explain equilibrium in a game by assuming equilibrium in some larger adjustment game. And if "no one is making a mistake" –that is, if everyone's play in the learning process is optimal given the play of the others, learning process itself is in a Nash equilibrium.

- So to explain how equilibrium arises, we **must allow for** players who "are making a mistake" in the sense that their behavior isn't a best response to the behavior of the other.

- The issue is **not** whether a model can generate suboptimal play but rather whether players would notice that some other adjustment rule would be better.

Convergence not guaranteed in general games.

Learning can suggest equilibrium refinements.

Most papers: **Play repeatedly without playing a "repeated game."**

Explained by reference to a "**large population model**" with many "agents" in each "player role."

Large population: unlikely to play their current opponent again for a long time, even unlikely to play anyone who played anyone who played him. So not worth sacrificing current payoff to influence this opponent's future play.

Leading case **anonymous random matching**: Each period all agents are matched to play the game, and they only see play in their own matches.

This treatment is used in most game theory experiments.

Learning depends on what the players observe…

- Own payoffs, realizations of moves by Nature
- Actions and/or payoffs of agents they interact with
- Actions and payoffs of other agents (i.e. Market share, word-of-mouth, and other forms of "social learning"- won't have time to cover this here.)

And on whether their actions change what is observed

- If not, "passive learning": this is the case I'll start with,
- If actions change what is observed about others, players may have an incentive to "experiment" as in bandit problems.

Specifying Learning Rules

- Worst case/minmax considerations such as no-regret learnng
- Maximization of expected discounted utility (standard in economics)
- Exogenously specified, "boundedly rational" or "behavioral" (this may or may not differ from Bayesian… w.o. constraints on the prior and likelihood functions, Bayesianism is extremely general…)

Today: focus on Bayesian models. Start with strategic form games and fictitious play. (The "omitted material" at the end of the slides briefly discusess alternatives such as reinforcement learning and imitation.) Then discuss learning in extensive-form games.

Fictitious Play

Introduced by Brown [1951] as a way to compute equilibrium in two-player games. (hence "fictitious" play.)

Can also be used as a simple stylized model of learning. Easy to motivate and analyze, too simple to match experimental data; foundation for more complex models.

Motivation: Suppose that an agent is going to repeatedly play a fixed two-player strategic-form game. The agent knows the structure of the game (the strategy spaces and payoff functions) but not how the other side is going to play. (*can be adapted to less prior knowledge.)*

All that the agent observes is the outcome of play in their own matches, which in a strategic form game is the realized pure strategy profile s.

Doesn't observe what happens in other matches, or opponents' past play.

- Agent believes she is playing against a randomly drawn opponent from a large population, does not try to influence opponent's play.

- Because what the agent observes is independent of own action, no incentive for "experimentation."

- Agents act as if they are Bayesian expected utility maximizers, facing a stationary, but unknown, distribution of opponents strategies.

- Stationarity a reasonable first hypothesis in many situations, might expect players to stick with it when it is approximately right but to reject stationarity given sufficient evidence to the contrary- as when there is a strong time trend.

- Stationarity impies all observations equally informative. People often seem to give less weight to older data (displaying "recency bias") as if facing a hidden Markov process.

Aside on recency bias:

- See Erev and Haruvy *Handbook oF Experimental Economics vol. 2* for a survey of some of the psychology literature.

- Game dynamics with recency: Young [1993], Benaim Hofbauer Hopkins JET [2009], Fudenberg-Levine PNAS [2014], Block, Fudenberg and Levine *TE* [2018].

- Fudenberg-Peysakhovich EC [2014] document (large) recency bias in a "lemons" problem: even after 20 observations people reacted "a lot" to the next one. And show recency bias can be diminished by providing summary statistics of past play. Need to better understand what sort of feedback encourages and discourages recency.

- Hard to develop formal results with non-negligble recency bias, as the system doesn't settle down; may need to use simulations.
  --- *end aside*---

(*review from Monday's lecture)*
In FP, the belief updating has a special simple form:

- Player *i* has an initial weight function that gives a weight to each opponent's strategy.
- Add 1 to the weight of each opponent strategy each time it is played.
- Probability that player *i* assigns to -*i* playing s^{-i} is the weight on s^{-i} divided by the sum of the weights.
- Call this mixed strategy for –i player *i's assessment* (and use "beliefs" for distributions over opponent strategies.)

Fictitious play is any behavior rule that at each history specifies a static best response to these beliefs.

Bayesian interepretation:

- Player believes opponents' play is sequence of i.i.d. multinomial random variables with a fixed but unknown distribution.

- Prior beliefs: Dirichlet distribution (parameters are the initial weights)

- Key conceptual point: assessments are the expected value of beliefs over mixed strategies:

$$\gamma_t^i(s^{-i}) = \int_{\Sigma^{-i}} \sigma^{-i}(s^{-i})\mu_t^i[\sigma^{-i}]d\sigma^{-i}.$$

Digression/Reference : ***Dirichlet Priors and Multinomial Sampling***

Taken from DeGroot [1970] *Optimal Statistical Decisions*.

1) *The Multinomial Distribution:* Consider a sequence of *n* i.i.d. trials, where each period one of *k* outcomes occurs, with p_z denoting the probability of outcome *z*. Denote the outcome of the *n* trials by the vector κ, where κ_z is the number of the outcomes of type *z*. Then the distribution of the outcome is

$$f(\kappa) = \frac{n!}{\kappa_1 ! \cdot \kappa_k !} p_1^{\kappa_1} \cdots p_k^{\kappa_k} \quad \text{for} \quad \kappa \quad \text{such that} \sum_{z=1}^{k} \kappa_z = n.$$

2) *The Dirichlet Distribution:* Let Γ denote the gamma function. A random vector *p* has the Dirichlet distribution with parameter vector α if its density is given by

$$f(p) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \cdots p_k^{a_k - 1}$$

for all $p>0$ such that $\sum_{z=1}^{k} p_z = 1$.

(The gamma function here is just as the "integrating constant": for f to be a density it has to integrate to 1.)

Fact: If *p* has the Dirichlet distribution, then the expected probability of outcome z is $\int p_z f(p)dp = \alpha_z \,/\, \sum_{w=1}^{k} \alpha_w$.

So weights α correspond to relative probability of each outcome.

Two densities with the same expected values correspond to different ways the agent will update beliefs.

Fact: The Dirichlet distributions are a conjugate family for multinomial Sampling: if data is κ and prior is prior is Dirichlet with parameter α then posterior is Dirichlet with parameter $\alpha + \kappa$.

If player *i*'s date-t beliefs about player -*i's* mixed strategy have a Dirichlet distribution, player *i's* assessment of the probability that -*i* plays s^{-i} in period *t* is

$$\gamma_t^i(s^{-i}) = \int_{\Sigma^{-i}} \sigma^{-i}(s^{-i})\mu_t^i[\sigma^{-i}]d\sigma^{-i} = \alpha_z \,/\, \sum_{w=1}^{k} \alpha_w ,$$

This is the expected value of the component of σ^{-i} corresponding to s^{-i}.

So after observing sample κ, player *i*'s assessment of probability that the next observation is strategy *z* is

$$\frac{\alpha_z'}{\sum_{w=1}^{k} \alpha_w'} = \frac{\alpha_z + \kappa_z}{\sum_{w=1}^{k} (\alpha_w + \kappa_w)},$$ which is the formula for fictitious play.

The Interpretation of Cycles in Belief-Based Learning

Consider the following example (FK JET 1993)

	A	B
A	0,0	1,1
B	1,1	0,0

Consider FP, with 1 agent per side, initial weights (1, sqrt (2)) for each player.

- First period: both players think other will play B, so both play A.
- Next period weights are (2,sqrt (2)) and both play B; the outcome is the alternating sequence ((A,A),(B,B), (A,A), etc.)
- Empirical frequencies of each player's choices converge to (½, ½), which is the Nash equilibrium. So FP "works" for the purpose of computing equilibrium.

- This isn't a good model of learning: Realized play is always on the diagonal, and both players receive payoff 0 in every period even though each can guarantee a payoff of ½.

- Reason: the empirical joint distribution on pairs of actions does not equal the product of the two marginal distributions, so that the empirical joint distribution corresponds to correlated as opposed to independent play.

- We could? Should? expect players to notice the cycle and form more sophisticated beliefs.

- So in general we won't want to identify a cycle with its average.

- And we may want to worry about how sensitive the players are to correlations in the data.

- Very long slowly moving cycles seem harder to detect than the 2-cycles in the example.

Multi-Player Fictitious Play

A related modeling issue with three of more players:

In a 3-player game, must player 1's assessment about the play of his opponents correspond to a mixed strategy profile, or should the range of the assessment include correlated distributions?

Note that player 1's current assessment corresponsd to a correlated distribution even if player 1 is certain that the true distribution of his opponents' play in fact corresponds to independent randomizations.

For example there is a difference between

(a) believing that your opponents are playing the correlated strategy
(1/2 (A,A), 1/2 (B,B))
and

(b) thinking they are either playing (A,A) or (B,B).

These two beliefs have the same (correlated), marginal over first period observations, get updated differently because they imply different correlations between period 1 and period 2. With beliefs (b), after one observation you know what they will be playing at all future dates.

No good reason to suppose initial assessments correspond to uncorrelated randomizations.

Deeper question: should support of the players' *prior beliefs* be the set of opponents' mixed strategy profiles, or allow for the possibility that opponents can consistently correlate their play?

If the support is mixed strategy profiles then persistent correlation will be ignored. Can dodge this by looking at two-player games, or looking at steady states, or both.

Important facts about FP:

- If actions converge they converge to a pure strategy NE.

- If time averages of empirical marginals converge the joint distribution is a NE. (*proof sketch: if player 2's marginal converges to* σ_2, *1's beliefs converge to* σ_2. *If the mixed action* σ_1 *corresponding to the limit of 1's empirical marginal is not a best response to* σ_2, *some* s_1 *is strictly better than any other strategy in the support of* σ_1 *for all large enough times. At such times 1 must play only* s_1)

- The above holds in any belief-based learning model that is "asymptotically empirical" (beliefs converge to empirical frequencies) and "asymptotically myopic" (eventually players choose actions that are myopic best responses to their beliefs.)

- Monderer, Samet, Sela [1995]: In FP asymptotically players expect to get at least their realized payoff to date- denote this by U_t^i .

- *Proof sketch*: let $k = \sum_{s^{-i}} \kappa_0^i(s^{-i})$: the length of the "fictitious history" implicit in the initial beliefs.

- Let $\hat{\sigma}_t^i$ be a best response to time-t beliefs γ_t^i.

- Then playing the best response to time-t belief is at least as good as playing the action played last period, so at time t player expects to get

$$u^i(\hat{\sigma}_t^i, \gamma_t^i) \geq u^i(\hat{\sigma}_{t-1}^i, \gamma_t^i) = \frac{\left(u^i(\hat{\sigma}_{t-1}^i, s_{t-1}^{-i}) + (t+k-1)u^i(\hat{\sigma}_{t-1}^i, \gamma_{t-1}^i)\right)}{(t+k)}$$

Expanding this backwards shows the player expects to get at least

$$\frac{\sum_{\tau=1}^{t-1} u^i(\hat{\sigma}_\tau^i, s_\tau^{-i}) + k u^i(\hat{\sigma}_1^i, \gamma_1^i)}{t+k} = \frac{tU_t^i + k u^i(\hat{\sigma}_1^i, \gamma_1^i)}{t+k} \to U_t^i.$$

- If the time averages converge, so do the beliefs, and so the payoff the players expect to get converge to the NE payoffs. (*note that this was true even in the example where their actual payoffs were below the NE!*)

- Early work showed empirical averages converge under FP s in 2x2 games and zero-sum games. (*though this needn't mean play converges.*)

Shapley [1964]: empirical averages under FP do not converge in this game:

$$\begin{matrix} 0,0 & 1,0 & 0,1 \\ 0,1 & 0,0 & 1,0 \\ 1,0 & 0,1 & 0,0 \end{matrix}$$

Under FP play follows the best response cycle, is never on the diagonal. Shapley gave direct proof that exact FP slows down fast enough that time averages don't converge.

This ended early literature on FP.

MSS proof of non-convergence: along the best response cycle the realized payoffs are 1, so the sum of what the two players expect to get converges to 1. But the sum of the NE payoffs is 2/3, QED.

Stochastic (or "Smooth") Fictitious Play

(Fudenberg-Kreps [1993])

Like FP but with a smooth (continuous) "stochastic best response function" that assigns a mixed strategy response to each belief.

Advantages:

- If beliefs converge behavior does too; not the case with standard fictitious play as shown by the 2-cycle example,
- Allows convergence to mixed-strategy equilibria in fictitious play-like models: Actual play in FP can't converge to a mixed equilibrium.
- Avoids the discontinuity in standard fictitious play, where a small change in the data can lead to an abrupt change in behavior.

- Discontinuous responses may not be descriptively realistic, can lead to "frequent switches" and so poor worst-case performance.

- Stochastic ficitious play is ε-consistent, where ε can be made arbitrarily small.

- Stochastic responses can be given a "Harsanyi-purification" foundation based on private payoff shocks.

Aside on payoff shocks

Suppose the *i's* payoff to profile *s* at time *t* is $u_t^i(s) = v^i(s) + e_t^i(s^i)$ where the $\{e_t^i\}$ are i.i.d. over time and players, and have a strictly positive density.

For each distribution σ^{-i} over the actions of *i*'s opponents, player *i*'s *best-response distribution* $\overline{BR}^i(\sigma^{-i})$ is given by

$$\overline{BR}^i(\sigma^{-i})(s^i) = \text{Prob}[\eta^i \text{ s.t. } s^i \text{ is a best response to } \sigma^{-i}]$$

Because there e is a unique best response distribution for almost every type, the smoothed best-response distribution is a <u>function.</u>

And these functions are continuous, and have a fixed point- a "*Nash distribution"* where each player's best response distribution is a best response to the others.

Harsanyi's purification theorem (*paraphrase*): For generic payoffs, any NE of the underlying game is the limit of pure strategy equilibria of these perturbed games as the distribution of payoff shocks becomes concentrated on 0.

And conversely the fixed points of the perturbed games converge to the fixed points of the unperturbed game (Hofbauer and Sandholm [2002]).

/*end aside*

Now back to learning: Wth payoff shocks if a player's assessment converges his behavior will too.

This is also true of smooth BR correspondences generated by non-linear payoff perturbations.

- Observation: If v^i is a smooth, strictly differentiable, concave function on the interior of Σ^i whose gradient becomes infinite at the boundary, then argmax $u^i(\sigma) + v^i(\sigma^i)$ is a smooth best response function that assigns positive probability to each of i's pure strategies.

- Canonical example of a perturbation function v is entropy: $v(p) = \sum_n p_n \ln(p_n)$; this generates logit best responses.

- Fudenberg, Iijima and Strzalecki [2015] characterizes the revealed-preference implications of a subclass of these perturbations, and show they correspond to ambiguity-aversion by an agent who is afraid of making the wrong choice.

Long run Dynamics of Smooth FP

For now stick with one agent per player role, even though this undercuts the idea that agents don't treat this as a repeated game- will then extend to large populations.

Ideas:

- Limit system deterministic because weight on shocks is 1/t, so use variant of LOLN.

- Limit is a continuous time system after time rescaling.

Consider a discrete-time stochastic process on a compact set in $\Re^n$:

$\theta_{t+1} - \theta_t = (F(\theta_t) + \eta_{t+1} + b_{t+1}) / (t+1)$, where

- F is C^2,
- the η_t are noise terms with bounded variance (and sometimes bounds on additional moments) satisfying $E[\eta_{t+1} \mid \theta_t, \ldots \theta_1] = 0$, and
- the b_t converge to 0 a.s.

Important: the η_t don't need to be independent or even exchangeable, just martingale differences.

Stochastic Approximation relates the behavior of this discrete-time stochastic system to that of the deterministic continuous-time system $\dot{\theta} = F(\theta)$.

Extended Intuition in an example:

state space [-1,1], with $F(0) = 0$, $\theta F(\theta) < 0$ for all $\theta \neq 0$.

0 is globally stable in the continuous-time dynamics, so the general theorem reduces to the conclusion that the discrete-time system converges to 0 with probability 1.

Can prove with Lyapunov function $V(\theta) = \theta^2$.

Note: $V(\theta) = \theta^2$ is <u>not</u> a supermartingale:
$E[V(\theta_{t+1}) \mid \theta_t = 0] > V(0) = 0.$

But outside of any fixed neighborhood of 0 we eventually have
$E[V(\theta_{t+1}) \mid \theta_t] < V(\theta_t)$ for *t* sufficiently large.

Intuition: the drift reduces V. Noise increases it, since V is convex, but the impact of each noisy observation diminishes at rate 1/t, so that the drift term eventually dominates in any region where V bounded away from zero.

Now back to more general stochastic approximation.

Consider

$$\theta_{t+1} - \theta_t = (F(\theta_t) + \eta_{t+1} + b_{t+1}) / (t+1).$$

First step: show that almost surely the sample path lies in some invariant set of the continuous-time process.

Definitions

The ω-limt set of a sample path $\{\theta_t\}$ is the set of long-run outcomes: y is in the ω-limit set if there is an increasing sequence of times $\{t_k\}$ such that $\theta_{t_k} \to y$ as $k \to \infty$.

A *flow* on X is a continuous function $\Phi : X \times R \to X$ such that $\Phi_0(x) = x$ and $\Phi_s(\Phi_t(x)) = \Phi_{t+s}(x)$.
(Take X to be a subset of finite dimensional Euclidean space.)

A semi-flow: same as a flow, but time is non-negative.

The solution of a differential equation is a semi-flow.

Extend flow to image of sets: $\Phi_t(A) = \{\Phi_t(x) : x \in A\}$

An *invariant set* of a continuous-time flow Φ: $\Phi_t(A) \subseteq A$ for all t .

Assume F is C^2, $E[\eta_{t+1} \mid h_t] = 0$ a.s., where h_t is history at end of t, noise term has uniformly bounded variance (a.s.), and b_t converges to 0 a.s.

Proposition: (Benaim and Hirsch [1999]) With probability one, the ω-limit set of any realization of the discrete- time process is an invariant set of the continuous-time process; this set is compact, connected, and contains no proper subsets that are attractors for the continuous- time process. (So it is connected and "internally chain recurrent.")

In application to smooth fictitious play, the state space will be the player's beliefs, and the map F is the e *smooth best response dynamic* $\overline{BR}(\theta) - \theta$.

Illustration: 2 players, 2 actions each.

$$\theta_{t+1} - \theta_t = \begin{bmatrix} (a_{1,t+1} - \theta_{1,t}) / (t+1) \\ (a_{2,t+1} - \theta_{2,t}) / (t+1) \end{bmatrix}$$

$$= \begin{bmatrix} (\overline{BR}_1(\theta_{2,t}) - \theta_{1,t}) / (t+1) \\ (\overline{BR}_2(\theta_{1,t}) - \theta_{2,t}) / (t+1) \end{bmatrix} + \begin{bmatrix} (a_{1,t+1} - \overline{BR}_1(\theta_{2,t})) / (t+1) \\ (a_{2,t+1} - \overline{BR}_2(\theta_{1,t})) / (t+1) \end{bmatrix}.$$

$$= \begin{bmatrix} (\overline{BR}_1(\theta_{2,t}) - \theta_{1,t}) / (t+1) \\ (\overline{BR}_2(\theta_{1,t}) - \theta_{2,t}) / (t+1) \end{bmatrix} + \eta_t$$

(Following BH this equates beliefs with the empirical distribution; turns out to be ok as the priors don't matter asymptotically.) The noise terms η_t have conditional expectation of zero, but are not in general i.i.d. or even exchangeable.

The Benaim-Hirsch proposition implies that if SFP eventually converges to a point or a cycle, the point or cycle should be a closed orbit of the continuous-time best reply dynamics.

Moreover, the noise will eventually "kick" the system away from any unstable states, at least if there is "enough noise."

Say that a steady state is *linearly unstable* if at least one of its associated eigenvalues has positive real part.

Proposition: (Pemantle[1990]): Suppose that $b_t \equiv 0$ and that the distribution of the noise term η_t is such that for every unit vector e_i, $E(\max(0, e_i \circ \eta_t) > c > 0$. (*full-enough-support assumption)*

Then if θ^* is linearly unstable for the continuous time process, $P\{\lim_{t\to\infty} \theta_t = \theta^*\} = 0$. (Brandiere and Duflo [1996] extend to $b_t \to 0$.)

Theorem: Smooth fictitious play converges to the Nash distribution in any game where the (unique) Nash distribution is a global attractor for the continuous-time dynamics.

One way to show the Nash distribution is a global attractor is to construct a strict Lyapunov function as FK do for 2x2 games with a unique mixed equilibrium.

What about 2x2 games with 2 strict equilibria and one mixed?

Benaim Hirsch show that SFP can't cycle in 2x2 games as it is "volume contracting." So it must converge to a steady state. Which?

Proposition (Benaim and Hirsch [1999]) If every strategy profile has positive probability at every state, and θ^* is an asymptotically stable equilibrium of the continuous time process, then $P[\theta_t \rightarrow \theta_*] > 0$.

The mixed equilibrium is unstable, and when payoff perturbations are small the only stable equilibria are near the NE.

Conclusion: In battle of the sexes, the two pure equilibria have positive probability and the mixed equilibrium has probability 0.

Hofbauer-Sandholm [2002] provide related results for "potential games:" games where we can transform payoffs w/o changing best responses so that the game is a team problem: $u_i(s) = u_j(s)$ for all *i*,*j*,*s*. Prof. Hofbauer will discuss this tomorrow.

Note that non-strategic behavior, as assumed in FP and SFP, doesn't make sense with one agent per role. Fudenberg-Takahashi [2011] consider a large populations of agents each of whom only sees outcomes of own matches.

Paper considers 3 matching strucures: 1) Two populations (one per player role), all agents matched to play each period; 2) one population, all agents matched to play each period; 3) One population, "asynchronous clocks": a pair of agents is selected at random each period, so ex-post some agents may play more often than others.

The structure of the argument similar in all 3 cases:

a) Derive a limit continuous time deterministic system with heterogeneous beliefs using stochastic approximation.

b) Argue that the system converges to homogeneous (identical) beliefs.

c) So convergence, local stability etc. on the smaller space implies same on the larger one.

Aynchronous clocks

To start assume each pair has the same probability of playing.

- When an agent is chosen, he uses a smooth best response given his beliefs. After the play, he updates his assessment based on the realized strategy of his partner. If a player is not chosen, he keeps the same assessment as before.

- The difference between this system and the usual SFP is that each agent only updates his assessment when he is drawn to play- and different agents may have played a different number of times.

- We want a 1/t step size to appeal to stochastic approximation theorems.

- Track the fraction of time each agent has played with an auxiliary state variable; this controls the difference in step sizes.

- That is set $\bar{\kappa}_{i,n} = \sum_s \kappa_{i,n}(s)$, and set $y_{i,n} = \frac{\bar{\kappa}_{i,n} + 1}{n+1}$.

- When all matches are equally likely then in the long run each agent will play 2/M of the time, so the semiflow has auxiliary equation $\dot{y}_i = y_i - 2/M$

- And the main state updates according to

$$\theta_i(\mathrm{t}) = (2/\mathrm{M})(1/(\mathrm{y}_i(\mathrm{t})))\left(1/(\mathrm{M}\text{-}1)\sum_{j\neq i} \overline{BR}_i(\theta_j(t)) - \theta_i(t)\right) .$$

Show that the conditions for stochastic approximation to be valid are satisfied.

Then analyzing the limit continuous time system show that if the number of agents $M > K+1$, where K is the Lipschitz constant of the smooth best response function, then all agents end up with the same beliefs.

- *Intuition:* agents i,j both are matched with the other M-2 agents. The difference in beliefs can be magnified by at most the slope K of BR, so the difference is bounded by decreasing exponential if M large compared to K.

- Counterexamples show that some bound on M is needed- lemma and theorems are false for small M.

Then show that when M>K+1 the system a.s. converges to a collection of states that is internally chain recurrent for the lower-dimensional system where all agents have the same beliefs. This means that a cycle that is stable in ordinary SFP stable here too.

Now suppose different pairs interact with different probabilities q_{ij} .

We show that beliefs and play converge when the population is large compate to K and the matching probabilities "not too different."

Specifically if $K\Delta < 1$ where $\Delta = \max_{1 \le i < j \le M} \sum_{k=1}^{M} | q_{ij} - q_{ik} | /2$.

This rules out most interesting network structures, e.g. 0 interaction with others who are "far away." We don't know what happens on more general networks..

Learning in Extensive form games

Detour: Experimentation and Learning in Bandit Problems

1-armed bandit: Choose "Out" (Safe arm): get *M* with certainty.
"In" (Risky arm) has fixed but unknown probability distribution.

Example:

- Risky arm pays 2 with probability *p* , 0 with probability (1-*p*).
- Prior $\mu(p=1)=.4$, $\mu(p=0)=.6$, so the expected value of risky arm is $2E[p]=.8$.
- Suppose $M=.9$ and discount factor $\delta=0$, then risky arm is never tried.
- Now suppose discount factor $\delta > 2/9$, and consider the policy

" Try In once, then switch to Out forever if get 0."
This policy yields

$$.4(2) + .6((1-\delta) * 0 + .9\delta) = .8 + .54\delta > .9$$

So do better to try the risky arm at least once as an "experiment."

Now suppose *p* is uniform on [0,1] (so E[p]=.5) and $M = 4/3$ (so myopic player uses the safe arm.)

- Here one observation isn't fully revealing.
- Posterior after *a* successes and *b* failures (0's): Beta (*a+1,b+1*)
- Posterior expected value of *p* is $(a+1)/(a+b+2)$. (*special case of the updating in FP)* .
- So after 1 success, E[*p*]= 2/3.

- Consider the (not optimal) policy "Try once, if get a success stay in forever else play the safe arm forever."

- This yields

$$.5\ \ 2(1-\delta)+\delta[2\ /\ 3\ *\ 2+1\ /\ 3\ *\ 0]\ \ +$$
$$.5\ *\ [(1-\delta)0+4\delta\ /\ 3]=1+\delta$$

- So worth using the risky arm at least once if $1+\delta>4\ /\ 3\leftrightarrow\delta>1\ /\ 3.$

- Reason: "information" or "option" value- the myopically optimal action isn't played to "experiment."

- To consider the optimal policy use dynamic programming: $V(\mu_t,M)=\max\{M,E_{\mu_t}[(1-\delta)r_t+\delta V(\mu_{t+1},M)]\}$.

- *Gittins index:* smallest M s.t. $V(\mu_t, M) = M$.

  Also equals best average discounted payoff for any stopping time τ , that is

$$\sup_{\tau>0}\left(\mathbf{E}\left\{\sum_{t=0}^{\tau-1}\delta^t u(t)\right\} / \mathbf{E}\left\{\sum_{t=0}^{\tau-1}\delta^t\right\}\right)$$

- k-armed bandit: finite number k of <u>independent</u> arms, each period the agent must pick one of them.

- Most tractable case of bandit. Independence assumption is restrictive but sometime defensible, and the solution provides more general intuitions.

- **Theorem (**Gittins) The solution is to use in each period the arm with the highest index. Reduces the k-dimensional optimization problem to k one-dimensional problems.

- For any fixed δ , experimentation ends in finite time with probability one: there is a T such that at all t>T the arm chosen is myopically optimal. This is because the expected amount learned- e.g. the variance of next period's expected beliefs- goes to 0.

- There can be positive probability that play "locks on" to the safe arm, even though it is suboptimal.

- But the probability of this goes to 0 as $\delta \to 1$ (see e.g. Rothschild [1974].)

- And with time averaging can get first best payoff by experimenting infinitely often but a vanishing fraction of the time.

- Big picture: we expect learning to be more complete when agents are more patient.

Learning in extensive form games:

- Agents observe (at most) the terminal nodes that are reached in their own plays of the game.

- Don't observe how the opponents would have played at information sets that were not reached in that play of the game, may not observe outcomes in other matches.

- So incorrect beliefs about off-path play could persist, and play might converge to a non-Nash outcome that is a *self-confirming equilibrium* (SCE) unless agents "experiment enough" with off path actions.

- If agents do "experiment enough", can rule out non-Nash equilibria and some Nash equilibria as well: get refinements of NE from results about where in the tree the agents experiment and with which actions.

- Many concepts in the literature related to incomplete learning and SCE, going back at least to Hahn [1978].

- Some have explicit learning foundations, others don't.

- Thinking about the learning foundations can help guide the specification of the equilibrium features, and explicit foundations are needed to study the impact of experimentation.

- Learning models and equilibrium concepts vary in many ways.

- First study equilibrium concepts that correspond to little or no experimentation. Then look at at explicit learning models with either myopic or patient players.

Notation

- $I+1$ players in the *game*, $i = I + 1$ is nature.
- Finite game tree with nodes $x \in X$, information sets *h*.
- Terminal nodes $z \in Z$; player $i's$ payoff function u_i is a function of *z*.
- S_i is the set of pure strategies for player *i*, $s \in S$ denotes a strategy profile for all players including nature.
- Each *s* determines a probability distribution $p(\bullet | s)$ over terminal nodes.
- Players know the extensive form of the game, except that they may not know the distribution of Nature's move. If Nature's move is unknown, players' beliefs about it are treated in the same way as their beliefs about the strategies of other players.

- A probability measure μ_i over Π_{-i}, the set of other players' behavior strategies, describes player i's beliefs about his opponents' play.
- Start with a simple case: myopic agents who don't experiment.
- Single player per role, or if many all have the same beliefs.
- Conjecture about opponents' play correspond to a mixed strategy profile: that is for each player *I*, μ_i is a point mass on some σ_{-i} . This follows from Kuhn's theorem in 2-player games but otherwise requires independence.

- Players see the realized terminal node
- This corresponds to "unitary independent self-confirming equilibrium:"

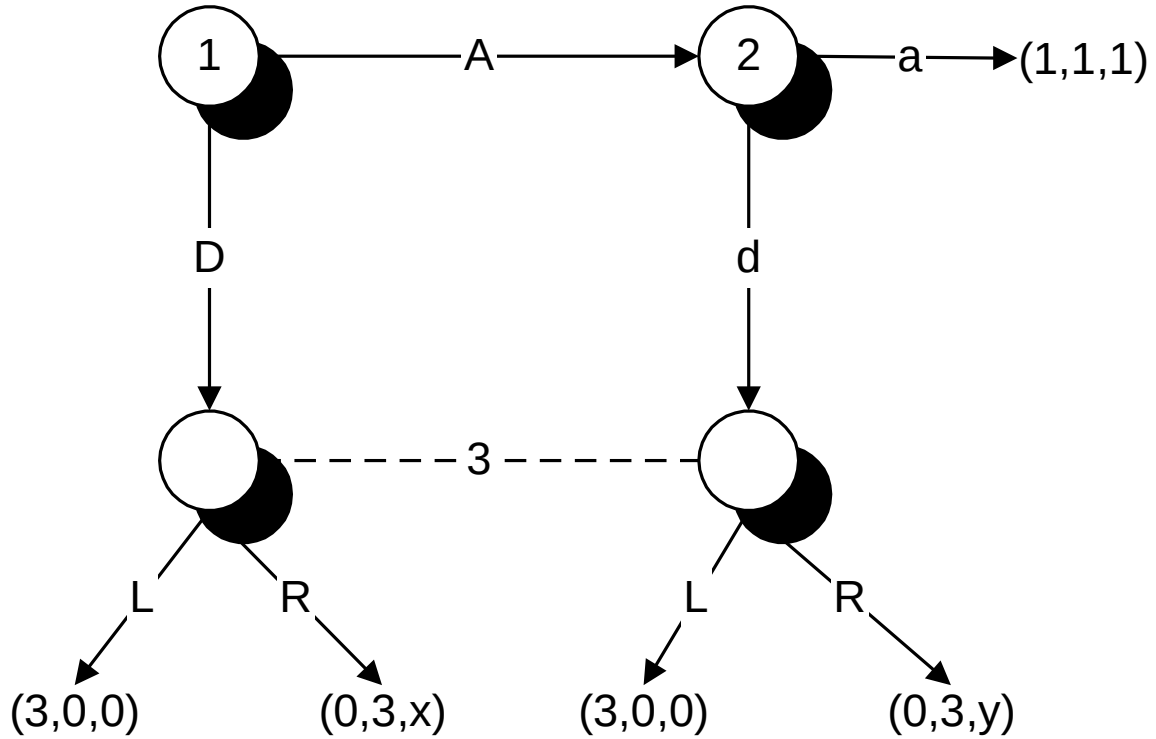
Definition: strategy profile σ is a *unitary independent self-confirming equilibrium* (SCE) if for each player i there is a conjecture $\bar{\sigma}_{-i}$ such that for each s_i with $\sigma_i(s_i) > 0$

(a) s_i is a best response to $\bar{\sigma}_{-i}$, and

(b) $p(\bullet|(\sigma_i, \bar{\sigma}_{-i})) = p(\bullet|\sigma)$.

- In two-player games, independent unitary SCE is outcome-equivalent to Nash equilibrium- change off-path play of each player 2 to match others' beliefs.

- But with more players there can be independent unitary SCE that differ from Nash because 2 players disagree about the play of a 3rd:



(A,a,L) is a SCE. It can also arise from a learning process where players 1 and 2 update beliefs about opponents' play as in fictitious play: As long as the prior makes them both play A, they get no data on 3's play, and hence don't update.

- Bayesian interpretation: each player's beliefs about the opponents are a product measure, so seeing player 2's action doesn't change 1's beliefs about 3's play. This is why 1 and 2 can "agree to disagree."

- But (A,a) is not a Nash equilibrium outcome: Nash equilibrium requires players 1 and 2 to make the same (correct) forecast of player 3's play, and if both make the same forecast, at least one of the players must choose *D*.

- It wouldn't matter that 1 and 2 have different beliefs about 3's play at information set *h* if only player 1 can cause that information set to be reached.

  Let $\bar{H}(s)$ be the path of s.

  Defn: A game has *observed deviators* if for all players *i*, all strategy profiles *s*, and all $\hat{s}^i \neq s^i$, $h \in \bar{H}(\hat{s}^i, s^{-i}) \setminus \bar{H}(s)$ implies that there is no $\hat{s}^{-i}$ with $h \in \bar{H}(s^i, \hat{s}^{-i})$.

- Implies that if a deviation by player *i* leads to an information set off the equilibrium path, there is no deviation by *i*'s opponents that leads to the same information set.

- Games of perfect information satisfy this condition, as do all multistage games with observed actions.

- Always satisfied in two-player games of perfect recall: With two players, both players must know who deviated. Satisfied in most economic examples but not in the "horse" game.

Theorem (Fudenberg Levine 93a**)** In games with observed deviators, the outcome of any independent unitary self-confirming equilibria is the outcome of a Nash equilibrium. (*this allows the strategy profiles to differ off-path*)

Proof Sketch:

- Unitary beliefs: a single set of beliefs for each player *i.*
- Independent beliefs: these beliefs correspond to a mixed strategy profile. (*uses Kuhn's theorem on equivalence of mixed and behavior strategies.*)
- For Nash equilibrium, beliefs about play 2 or more steps off the equilibrium path don't matter- all that matters are beliefs about play at information sets that some player could cause to be reached with a unilateral deviation.

- Observed deviators: at most one player's beliefs about play at an off path information set are relevant- namely the beliefs of the player who could cause that information set to be reached.

- So construct a new strategy from the original one by changing play at each information set *h* one step off the equilibrium path to match the beliefs of the relevant player $i_r(h)$.

- Because the information set is off-path, the player who moves at $i_r(h)$ is indifferent about the change (as far as Nash equilibrium is concerned.)

- Player $i_r(h)$ *'s* stragey is now a best response to actual play at $i_r(h)$.

- And the incentives of the other players aren't altered by this change.

- So new profile is a NE with same outcome as the SCE we started with.

Now a more general definition inspired by large-population learning models:

Definition: σ is an *independent heterogeneous self-confirming equilibrium* (SCE) if for each player i and each s_i with $\sigma_i(s_i) > 0$ there is a conjecture $\bar{\sigma}_{-i}$ such that

(a) s_i is a best response to $\bar{\sigma}_{-i}$, and

(b) $p(\bullet | (\sigma_i, \bar{\sigma}_{-i})) = p(\bullet | \sigma)$

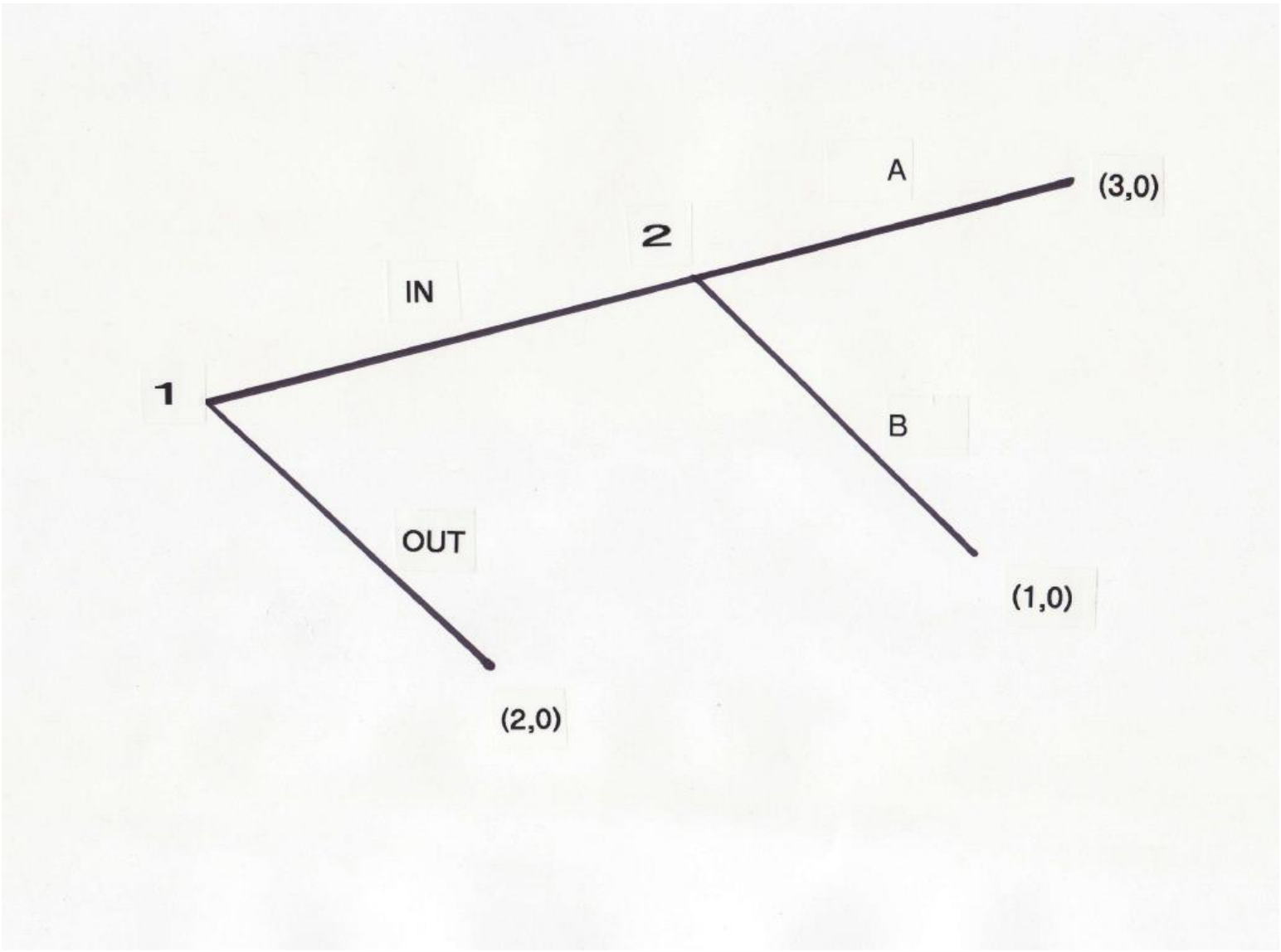
- Like previous version, this reduces to Nash in one-shot simultaneous-move games.
- But allows player *i* to rationalize each strategy in the support of σ_i with a different $\bar{\sigma}_{-i}$.

Intepretation: Many agents in the role of each player, and different agents in the role of player *i* may have observed play at different nodes. (*heterogeneity is a non-issue for Nash equilibrium.)*

Heterogeneous beliefs are important in many game theory experiments (see Fudenberg Levine [1997]), possibly also important in the field.

- Following game allows a simple example of the impact of heterogeneous beliefs:

  No Nash equilibrium with outcome distribution
  (½ *Out*, ½ (*In*, *A*))


A
(3,0)
2
IN
1
B
OUT
(1,0)
(2,0)

What factors leads SCE to differ from Nash equilibrium?

- *Heterogeneous Beliefs*

- *Correlated Beliefs*: beliefs about play at unreached info sets isn't a product measure. Can generate additional SCE outcomes: a player can be deterred from playing an out-of-equilibrium action by correlated subjective uncertainty about the response. (even if the player is certain that the actual play of opponents is uncorrelated.)

- "*Inconsistent*" beliefs: 1 and 2 disagree about the play of 3 at some info set *h* that both 1 and 2 can unilaterally cause to be reached. (*irrelevant in games with observed deviators.*)

- Theorem: that's it- shut down these 3 channels and SCE is outcome-equivalent to NE.

Adding Prior Information about Payoffs+ Rationality to SCE

- In SCE the only constraint on beliefs is what players observe about others' play- players aren't required to use information about opponents' payoff functions.

- May be a good approximation of some field situations and for experiments in which subjects are given no information about opponents' payoffs.

- In other cases, players do have some prior information about their opponents' payoffs.

- In experiments, giving subjects information about other players' payoff functions can make a difference. (see e.g. Prasnikar-Roth [1992]).

- This difference corresponds to the distinction between SCE and Rationalizable Self Confirming Equilibrium or RSCE. (Dekel, Fudenberg, Levine 1999).

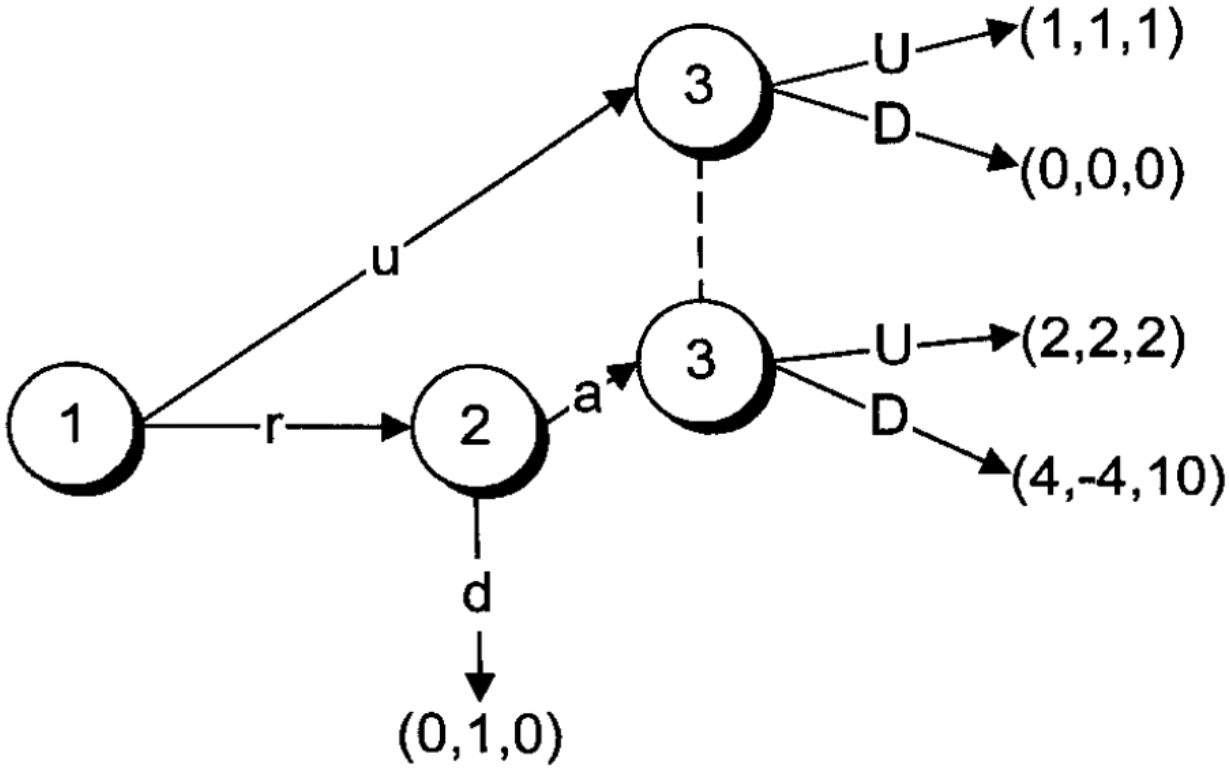
- RSCE is “unitary”- a single belief for all players, and all players see the same distribution on terminal nodes.

- RSCE imposes some off-path optimality restrictions. It coincides with backwards induction in two-stage games of perfect information, but in longer games it is much weaker and more like SCE.

- RSCE has implications beyond the intersection of SCE and rationalizability

- These come from the assumption that the outcome path is public information.

Try to explain this w/o the formal definition



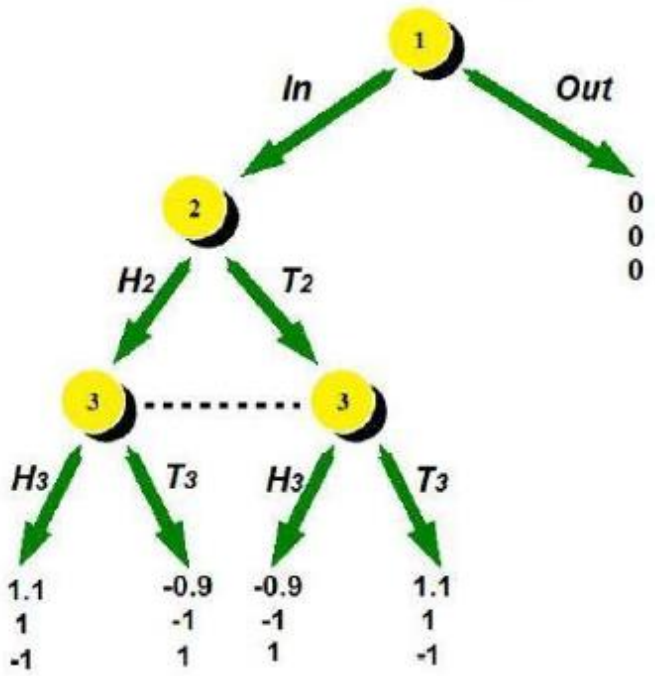
(u,U) is a Nash outcome (so self-confirming) not RSCE: If player 1 knows 2 knows 3 is playing up, can use this knowledge and his knowledge of player 2's payoffs to deduce that 2 will play a.

RPCE: extends RSCE to the case where players only see a partition of the terminal nodes, as in sealed-bid 1st price auctions where players see winning bid but not losing ones.

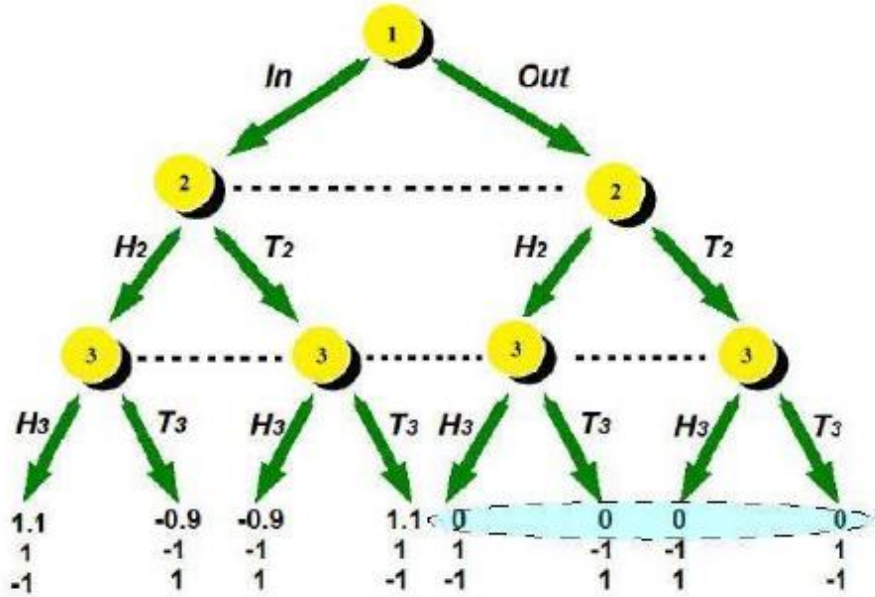
Here even with unitary beliefs there isn't a commonly observed distribution on terminal nodes.

In game A on the next page, players 2 and 3 do not observe each other's play when 1 plays *Out*, so there is no reason for player 1 to expect their play to resemble a Nash equilibrium, so an impatient player 1 might choose to play *Out*, fearing that player 2 would lose to player 3.

In game B players 2 and 3 observe each other's play, whatever player 1's action is. Thus they should be playing as in the Nash equilibrium of the matching pennies game, and 1 knows this, so she should play *In*.


1
In
Out
0
0
0
2
H2
T2
3
3
H3
T3
H3
T3
1.1
1
-1
-0.9
-1
1
-0.9
-1
1
1.1
1
-1

Game A


1
In
Out
2
2
H2
T2
H2
T2
3
3
3
3
H3
T3
H3
T3
H3
T3
H3
T3
1.1
1
-1
-0.9
-1
1
-0.9
-1
1
1.1
1
-1
0
1
-1
0
-1
1
0
-1
1
0
1
-1

Game B

Modelling experimentation

SCE models the long-run outcomes when players do little or no experimentation with off path actions.

Fudenberg and Kreps [1988, *JET* 1995], Jehiel and Samet *JET* [2004], Laslier and Walliser *JET* [2004] look at "boundedly rational" experimentation.

FK 88: "belief-based learning" in the spirit of fictitious play, with the extra assumption that agents experiment at rate $1/t$; i.e. there is a lower bound on the probability of each action, and the bound goes to 0 over time at rate $1/t$.

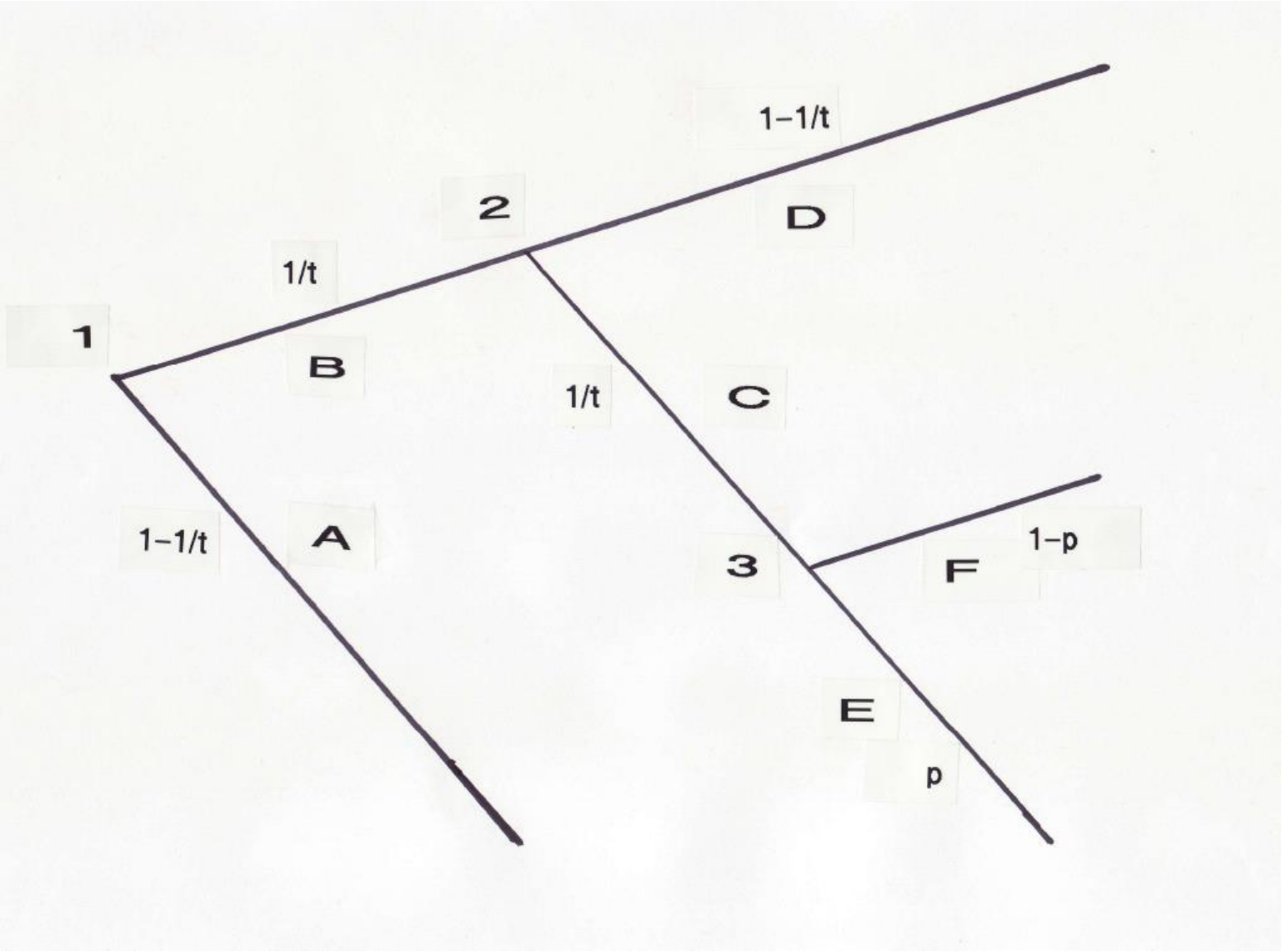
To rule out convergence to non-Nash outcomes, it is enough that players have correct beliefs about play at any "relevant" information set-information sets that can be reached if any one player deviates from the equilibrium path.

With the "$1/t$ experimentation rule," these information sets are reached infinitely often, because $\sum_{t=1}^{\infty} 1/t = \infty$. (0-1 *laws*)

- So from the law of large numbers and asymptotic empiricism, beliefs at relevant information sets become correct.

- Rational Bayesians don't randomize, and even when patient may never play some actions.

- And the $1/t$ rule needn't lead to correct beliefs at nodes that take 2 or more deviations to reach, because $\sum_{t=1}^{\infty} 1/t^2 \neq \infty$.


1–1/t
2
D
1/t
1
B
1/t
C
1–1/t
A
1–p
3
F
E
p

How much experimentation will players will actually do?

One way to pin this down is to derive experimentation from dynamic programming.

Fudenberg and Levine [1993, 2006], Fudenberg and He [2018, 2019]:

- Continuum of agents, with steady inflow of new agents who do not know the prevailing distribution of strategies, accompanied by an outflow of equal size. (*FL assume agents have finite lifetimes T, FH assume lifetimes have a geometric distribution with parameter* γ.)

- Doubly infinite time periods- no initial time.

- Agents have no prior knowledge of other agents' payoff functions.

- Each time the game is played, the agent observes only the terminal node.

- Agents believe the system is in a steady state.
- ***Assume prior beliefs over strategies are*** *non-doctrinaire*: represented by a continuous density function that is strictly positive at interior points. (Thus no action has probability 0, and given enough observations of a steady state beliefs would converge to the truth. This allows priors to go to zero on the boundary, as is the case for many Dirichlet priors.)
- Updates beliefs about strategies using Bayes rule.
- Non-doctrinaire prior implies non-doctrinaire posterior.
- The agents face dynamic programming problems.
- These problems have deterministic optimal policies that map histories to strategies. (*Note: agents with the same policy may meet different opponents, and so have different histories and play different strategies.)*

- A state of the system is the fraction of agents with each possible history.
- When agents have fixed lifetimes the state vector is finite-dimensional.
- With geometric lifetimes the state vector has countably infinite dimension but has "thin tails."
- With a continuum of agents, we can define the random-matching process to be deterministic.
- Then the deterministic policies of the agents generate a deterministic map from states today to states tomorrow: the current state determines the current distribution of strategies used, agents are randomly matched and observe their outcomes, and some agents leave and are replaced.
- Steady states are fixed points of this map.

- Steady states exist even though individual agents learn, because departing agents take their information with them. (*proof via fixed point theorem*: *because there are few agents with long histories, the set of states is a locally convex Hausdorf space in the sup-norm topology, and the "update" map is continuous.*)
- Steady state for short lifetimes determined by the priors.

- Interesting open problem: characterize steady states for intermediate lifetimes in some interesting examples. Here long-lived, better informed players might be able to take advantage of the new entrants.

- But results focus on limits as lifetimes long, so most players have lots of observations of play.

- If agents are myopic and don't experiment very much, they can maintain mistaken beliefs about how opponents would respond to deviations.

- So even when agents play many times, there can be non-Nash steady states that are self-confirming equilibria. But because agents learn the path of play and eventually stop experimenting, steady states must be SCE, possibly with heterogeneous beliefs.

  Theorem: A limit of steady states for lifetimes $T \to \infty$ (FL 93b) or continuation probabilities $\gamma \to 1$ (FH 2018, 19) must be a self-confirming equilibrium.

- I*ntuition/proof sketch*

a) If s^i played a positive fraction of the time when agents live a long time (e.g. as $T \to \infty$ or $\gamma \to 1$), then it must be played by a positive fraction of the population a positive fraction of their life.

b) Most agents who have played s^i many times have approximately correct beliefs about the path what happens when they do:

- LOLN: most players have accurate samples.
- Diaconis and Freedman *Annals Statistics* [1990]: posteriors converge to empirical distribution at a rate that depends only on sample size. (This uses non-doctrinaire priors.)

c) Agents eventually stop experimenting and play myopic BR to beliefs. This is related to what happens in bandit problems.

In extensive-form games there is an additional complication that comes from the assumption that players know the extensive form:

Example: 1 chooses L or R, 2 simultaneously chooses A or B. 3 only gets to play if following (R,B).

Suppose player 1 has played both L and R many times, and equally frequently, and that empirical cdf is (¼ (L, A), ¼ (L,B), ½ (R,A)).

1 has learned that 2 mixes (¾,A, ¼ B) so 1 assigns probability ¼ to 3 being reached if 1 plays R, but 1 has no observations on 3's play. So 1 has an "information value" for R, even though R has been played many times.

Here 1 knows some samples are "unrepresentative," so there can be nontrivial expected information value from an action that has been played many times unlike in a bandit problem. But LOLN says such samples are rare (if 2's play really is independent of 1's) .

- Experimentation and Learning by Patient Bayesians

  - Study the *patiently stable steady states* where the agents are both long-lived (so have lots of data) and patient (so have a reason to experiment). (*formally these are limits as first lifetimes grow long and then discount factor tends to 1; needed in proofs we still don't know if it's needed for the results.*)

  - Fudenberg and Levine *Ema* [1993] showed that these patiently stable steady states must correspond to Nash equilibria.

  - Note that even with patient agents the steady states depend on the prior: If everyone enters the system assigning high probability to a particular strict equilibrium, that's what will occur.

Why are patiently stable steady states Nash?

- Easy to prove under "1/t experimentation."

- But for generic beliefs rational players don't randomize in decision problems. And it's not obvious whether players off the equilibrium path want to experiment at all. (In fact we'll see that sometimes they won't.)

- Instead of direct bounds on experimentation, proof uses an indirect approach: In a steady state, most players who use a strategy s_i do so because it maximizes their current period's expected payoff: If they have played s_i many times, they do not expect to learn much about its consequences, so its "option value" is low. *(This step uses the order of limits…)* And then show this leads to contradiction at non-Nash states.

Sketch of proof:

1) If the steady state σ is not Nash-> some player *i* can gain at least k by deviating. Let *E* be all profiles where that player can gain at least this *k*.

2) Prior assigns positive probability to *E*. *(why?)* LOLN says agents are unlikely to have samples that both hit an information set many times and are very different from the steady state play there, so Diaconis-Freedman implies most agents' posterior probability of *E* is bounded away from 0.

3) Show this implies that a sufficiently patient player *i* has a positive option value for some $s_i' \notin \text{support}(\sigma)$.

Conclusion: patient players experiment enough to rule out non Nash states.

This does not say that all NE are limits of steady states with patient players.

FL *AER* [2006]

"Simple games": perfect information, and no player moves more than once on any path.

Simplifies inference and optimal experimentation: no reason to take action 1 to learn about the consequences of action 2.

Show that for some non-doctrinaire priors there is no off-path experimentation- hence needn't get backward induction.

But off-path play isn't completely arbitrary: nodes one step off the path are reached infinitely often, and so play there looks like a SCE.

Defn: Node x is *one step off the path of* π if x isn't reached under π, and is an immediate successor of a node that is reached under π.

Defn: Profile π is a *subgame-confirmed Nash equilibrium* if it is a Nash equilibrium and if, in each subgame beginning one step off the path, the restriction of π to the subgame is self-confirming in that subgame.

Theorem In simple games with no own ties, any subgame-confirmed Nash equilibrium that is nearly pure is path-equivalent to a patiently stable state. (*may need to choose the priors carefully..)*

- No own ties: no player has a pair of actions that lead to the same payoffs for him. Implies unique BI solution, also implies that players will not randomize on the equilibrium path.

- Nearly pure: no randomization on path, only Nature randomizes off the path. Not necessary in games of length 3 or less, don't know if needed in general. (General result would need bounds on experimentation at off-path nodes when there is mixing on the equilibrium path).

Proof strategy:

- Specify non-doctrinaire and independent beliefs that assign probability near 1 to a nghbd of subgame confirmed equilibrium $\hat{\pi}$.

- Fix a hypothetical steady state is close to the equilibrium, show that regardless of the discount factor a vanishingly small fraction of off-path agents experiment as lifetime $T \to \infty$, because they are reached with such low probability that experimentation doesn't pay.

- Use fixed point argument to show there is a steady state in the ngbd of $\hat{\pi}$, and that there is a sequence of fixed points that converge to a strategy that is path-equivalent to $\hat{\pi}$ as $T \to \infty$.

Lemmas:

a) Define 'option value" of reaching a node to be the expected future value of the associated information, ignoring the current period's payoff. Show that this option value is bounded by a term that is proportional to the subjective probability that the node is reached.

b) Almost all agents eventually stop experimenting, so along a sequence of steady states that converge to $\hat{\pi}$ as $T \to \infty$ or $\gamma \to 1$, the probability of nodes that are off-path under $\hat{\pi}$ goes to 0.

c) If steady state gives a node low probability, then few agents get samples of <u>any length</u> that make it look non-negligible.

■ Proof of c):

i) when priors are concentrated on strategies that give the node probability near 0, small samples can't make its posterior probability large, and

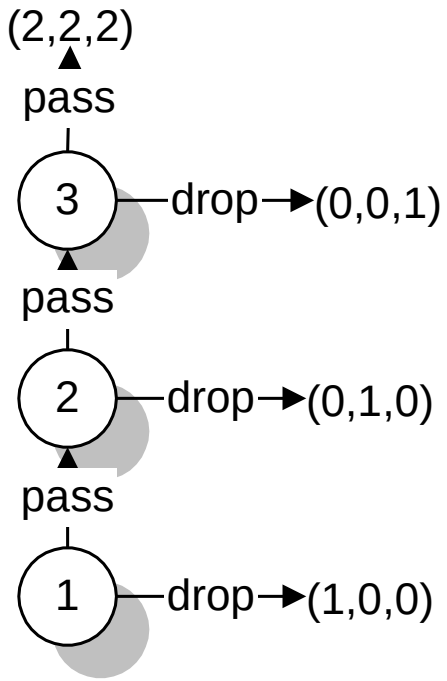
ii) "uniform LOLN": If reaching a node is unlikely then there is n* such that it is unlikely that any sample of length more than n* makes the posterior probability of the node large. (Like usual strong law, only uniform over a range of sample sizes.)

Implications of subgame- confirmed equilibria:

- In a simple game with no more than two consecutive moves, self-confirming equilibrium for any player moving second implies optimal play by that player, so subgame-confirmed Nash equilibrium implies subgame perfection.

- This can fail when there are paths of length three:

Example (Three Player Centipede Game)



- Unique subgame-perfect equilibrium: all players pass.
- (drop, drop, pass) is subgame-confirmed: since 2 drops, doesn't learn 3's play.

- A 4-node centipede game shows that subgame-confirmed is not equivalent to equivalent to Kalai-Neme *IJGT* [1992] 1-step perfection.

- Main takeaway: wrong beliefs off the path are more persistent than wrong beliefs on the path.

- Fanciful application: which superstitions are stable?

Illustration: The second of Hammurabi's laws is

"If any one bring an accusation against a man, and the accused go to the river and leap into the river, if he sink in the river his accuser shall take possession of his house. But if the river prove that the accused is not guilty, and he escape unhurt, then he who had brought the accusation shall be put to death, while he who leaped into the river shall take possession of the house that had belonged to his accuser."

- Looks like an attempt to provide incentives for truthful accusations.

- Seems based on superstition that the guilty are more likely to drown than the innocent. If people are this superstitious, why didn't Hammurabi simply assert that those who are guilty will be struck dead by lightning, while the innocent will not be?

- Our explanation: Suppose players are indoctrinated into the social norm as children and start playing the game with believing that it is likely that the social norm is true.

- Since the players are relatively patient, when young they optimally decide to commit a few crimes.

- "Lightning-strike norm" unravels.

The “Hammurabi social norm” is to not commit crimes and to only accuse the guilty; enforced by belief that anyone who makes false accusation drowns.

If older people adhere to this norm, then young player 1’s commit crimes, are accused of crimes, and are punished, so they correctly learn that (given the behavior of accusers) crime does not pay, and as they grow older stop committing crimes.

Accusers wrongly believe that the probability of the suspect drowning depends on who they accuse. Since there are few crimes, accusers only get to play infrequently, which reduces the value of experimenting with false accusations. In the U.S. today, the probability of being called as a witness at a trial is small… probably same was true then…

Fudenberg-He [2018, 2019]

Necessary conditions for patient stability. 2018 paper studies signaling games where each sender's type is fixed at birth; 2019 studies a more general class of games.

Derive an equilibrium refinement from conditions on the relative frequency of experiments.

Key: Assume independent priors. This lets us apply the Gittins index.

Signalling Game Notation

- Finite set of *types* for the sender, prior λ, λ(θ)>0 for all $\theta \in \Theta$.
- Finite set M of *messages* for the sender.
- Finite set A of *actions* for the receiver.
- *utility functions* u_S, u_R (or sometimes u_1, u_2) .
- Agents know λ (*our results also hold if they don't*).
- Behavior strategies $\pi_S(\cdot \mid \theta)_{\theta \in \Theta}$, $\pi_R(\cdot \mid m)_{m \in \mathcal{M}}$.
- $\Delta(X) :=$ all probability distributions over a finite set X.
- $BR(P, m) := \bigcup_{p \in P} \{ a \in \arg\max u_R(a, m, p) \}$: union of best responses to $p \in P \subseteq \Delta(\Theta)$

Definition: $\theta' \succsim_{m'} \theta''$ (θ' *is more compatible with* m' *than* θ'') if whenever is a weak best response for θ'' against some rational receiver strategy, it is a strict best response for θ' against that strategy (*e.g. strictly better than choosing any* $m'' \neq m'$ *.)*

Proposition: compatability is transitive, and it is irreflexive "on the relevant messages." (I.e. except for messages that are strictly dominant for both types, or strictly dominated for both types.)

Proposition If $\theta' \succsim_{m'} \theta''$, then in every perfect Bayesian equilibrium π^*, $\pi_S^*(m' \mid \theta') \geq \pi_S^*(m' \mid \theta'')$.

Proof: If $\pi_S^*(m' \mid \theta'') > 0$ then m' is at least weakly optimal for θ'', so if $\theta' \succsim_{m'} \theta''$ then m' is the (unique) strict best response for θ' .

So in equilibrium if $\pi_S^*(m' \mid \theta') > 0$ and $\theta' \succ_{m'} \theta''$, then $p(\theta'' \mid m') / p(\theta' \mid m') \leq \lambda(\theta'') / \lambda(\theta)$.

Type-compatible equilibrium imposes this monotonicity at off-path messages.

For given strategy profile π^*, let $u_1^*(\theta)$ denote the payoff to type θ.

Set $J(m,\pi^*) := \left\{\theta \in \Theta : \max_{a\in A} u_1(\theta,m,a) > u_1(\theta;\pi^*)\right\}$: types who could possibly do better than their equilibrium payoff by sending *m*.
And define

$$P(m,\pi^*) := \cap\left\{P_{\theta' \triangleright \theta''} : \theta' \succsim_m \theta'' \text{ and } \theta' \in J(m,\pi^*)\right\}$$

Definition: π^* is a *type-compatible equilibrium* (TCE) if

1. It is a Nash equilibrium and

2. It satisfies the *compatibility criterion* that

$\pi_2(\cdot \mid s) \in \Delta(\mathrm{BR}(P(s, \pi^*), s))$ for all s.

Theorem Every patiently stable profile in a signalling game is a TCE.

Intuition for the role of type compatibility:

Claim : If $\theta' \succ_{m'} \theta''$, the types have the same beliefs, and m' has the highest Gittins index for θ'', then it also has the highest index for θ'.

Given probability distribution ν over responses to m, every stopping time τ for message m induces a distribution over discounted receiver actions observed before stopping:
For each receiver action a, set

$$\sigma_m(\tau,\beta)(a) = \frac{E_\nu \sum_{t=0}^{\infty} \beta^t P_\nu[\tau \geq t \text{ and see } a \text{ in period } t]}{\sum_{t=0}^{\infty} \beta^t P_\nu[\tau \geq t]} \quad .$$

- Example: Suppose ν is (½ $\pi_R(a_1 \mid m) = 1$, ½ $\pi_R(a_2 \mid m) = 1$): either the receiver always plays a_1 or they always play a_2.

- And suppose τ says to stop at the first observation of a_2 .

  Then only two positive-probability histories:

  Pr{single a_2 then stop} =Pr{ a_1 every period, never stop}=.5

  So $\sigma_m(\tau, \beta)(a_2) = \dfrac{.5}{1 + .5\sum_{t=1}^{\infty} \beta^t} = \dfrac{1-\beta}{2-\beta}$.

- We show that show that for each sender type θ, the expression that defines the Gittins index for m is the same as the payoff against the receiver strategy induced by τ.

- So if θ' and θ" have same beliefs, $\theta' \succ_{m'} \theta''$, and m' has the highest Gittins index for type θ", it also has the highest index for θ'.

In the learning model the two types may use different policies, get different information, and so not have same posterior beliefs about receiver play.

- We extend the observation above to aggregate sender play using the idea that all types face the same population of receivers using a "coupling" argument based on "pre-progammed responses."

A pre-programmed response sequence for a given message m is an abritrary list of receiver responses to it.

Fix a profile of response sequences, one for each *m*, and imagine that the kth time i plays *m*, the receiver plays the kth element of the corresponding sequence.

Given a sender's policy (their map from past histories to current signals) each pre-programmed response sequence determines the history they will observe and her play in every period.

Show by induction on n that the period where s is played for the nth time by the more compatible type is no later than the period it is played by the less compatible type.

Because both types face the same distribution of response sequences, this shows that the more compatible type in aggregate plays the signal more.

Last steps:

- Use assumptions of nondoctrinaire priors and $\max_{a \in BR(\Delta(\Theta),m)} u_S(\theta', m, a) > u_S^*(\theta')$ to show that a patient θ' plays m "many times," because eventually it learns the payoffs to its equilibrium messages, so it will experiment many times with potentially better messages.

- Show that when m is played many times, most receivers correctly believe that more compatible types plays m more than less compatible types do, so the posterior odds ratios for a more vs. less compatible type exceeds the prior. (Fudenberg, Imhof and He [2017]). (*this uses an extra assumption on the priors- they can't go to 0 at the boundary more than polynomially quickly. This nests the Dirichlet priors used that correspond to fictitious play)*

Comparison with Simple Games

- Here sequential equilibrium makes strong predictions: generically unique.
- These restrictions aren't implied by rational learning: every *subgame confirmed* equilibrium is patiently stable in simple games. So patient learning allows more outcomes than does sequential equilibrium.
- In signalling games: patient learning makes *stronger* predictions than sequential equilibrium.
- Reasons:

  Here the relative probabilities of experiments matters as there are non-singleton information sets. And different types of senders have different incentives to experiment.

  Note also that receivers don't need to experiment; they see sender's type at the end of each round.

Ongoing work FH (2019): Extend to "Player-Compatible Equilibrium"

Step 1: Define a "compatibility order" for "parallel players" in general games.

These are players with the same action sets whose payoffs depend only weakly each other's actions.

$\left(a_i^* \mid i\right) \succsim \left(a_j^* \mid j\right)$ if for every strictly mixed profile α_{-ij} s.t. a_j^* is weakly optimal for j vs. $(\hat{\alpha}_i, \alpha_{-ij})$ for some $\hat{\alpha}_i \in \Delta^\circ(\mathbb{A}_i)$, a_i^* is strictly optimal for i *vs.* (α_j, α_{-ij}) for all strictly positive α_j .

This reduces to type compatibility in signaling games when we view each type as a player.

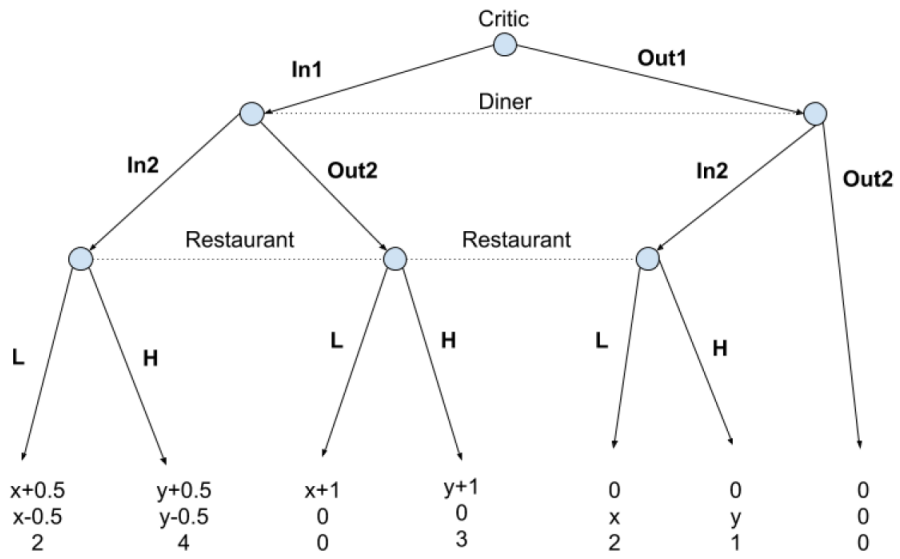
Step 2: A *player-compatible equilibrium* (PCE) is, roughly, a limit of trembling-hand perfect equilibria with the extra requirement that along the limit types that are more compatible with an action play it more.

Idea: trembles correspond to experiments and more compatible experiments are played more often.

Player compatible equilibria can be justified with the limits of Bayesian learning models we are also exploring cases where it can be justified using related quasi-Bayesian methods like *Thompson sampling* and *upper confidence intervals.*

PCE makes comparative statics predictions across game parameters in games were other equilibrium concepts don't.

Necessary condition, not sufficient. The exact implications of patient learning for equilibrium refinements are still open..



- Critic's utility minimized when Diner chooses In2 with prob 1.
- Critic gets higher utility from going to a crowded restaurant than the diner from going to an empty restaurant, given same food quality.
- So (**In1**|Critic) $\succ$ (**In2**|Diner).

Omitted Material 1: Reinforcement Learning.

A branch of the learning literature studies non-Bayesian models inspired by or taken from psychology.

A leading example is reinforcement learning (REL), where agents update based only on received payoffs.

Consider a 2 player 2x2 game where player 1 chooses U or D and 2 plays L or R. In FP, when 2 plays L, 1 updates beliefs in the same way whether he himself played U or D. So if he played U and 2 played R and this gave a good payoff but (D,R) would have been better, it tends to make 1 want to switch.

Two steps to defing a REL process:

a) what is reinforced? Actions, strategies, rules?

b) how?

Simple(st?) version is *Cumulative Proportional Reinforcement* (CPR)

- Normalize so that all utilities are positive, and give initial weights $CU_k(1)$ to each action *k*.
- Then update the score of the action that was played by its realized payoff, do not update other scores.
- Probability of each action is proportional to its score: $\Pr_k(t) = CU_k(t) / \sum CU_{k'}(t)$

- Since score is a function of the realized payoffs, 1's response to "2 played ***R***" can depend on 1's own action.

- If players told the structure of the game, and they are rational, they shouldn't condition on own action; in lab some subjects seem to do so.

- On the other hand, REL agents don't respond at all to what they could have gotten by playing something else- that is they ignore their regret- and it seems that most people don't do that.

- Also, subjects play differently if told the play of others and not just own payoffs-which goes against standard REL models.

Roth-Erev [1996] and Cheung-Friedman [1997] look at reinforcement learning and fictitious play respectively. (Roth Erev model is an extension of CPR with more parameters) Each find that their ex-ante preferred model fits better

Camerer-Ho [1999] "EWA learning" nests both reinforcement learning and fictitious play as special cases by adding a parameter that allows for different weight on actual payoff and "hypothetical reinforcement" or regret. They find that the best fit is "in the middle."

My (not their) main take away: these 3 models all fit the data reasonably well in cases where an agent using it would do a reasonably good job of optimizing, and these are cases where play is not changing very quickly over the course of the experiment.

In games where play has a strong trend (like “beauty contest game” : guess 2/3 the average) none of the models CH consider do well.

Salmon *Econometrica* [2001] shows little power in the tests used to distinguish the learning models. Following Camerer Ho, he uses assumes all agents use the same rule.

Wilcox *Econometrica* [2006] shows that fitting a representative-agent model to learning data tends to bias estimates in favor of REL.

Intuition: if all players are belief learners, but they are heterogeneous, pooled estimation of the model gives prediction errors that are correlated with past strategy choices of players, because those past choices carry idiosyncratic parameters. So players’ own past choices have relevant into, and the past payoffs used in REL depend on past choices…

Omitted Material 2: Learning by Mis-specified Bayesians

A Bayesian with a full support prior over multinomial probabilities is asymptotically empirical" and will eventually learn the true distribution.

Not true if the the prior is mis-specified and assigns probability 0 to a neighborhood of the truth.

Berk [1966]: when the agent is trying to learn a parameter from a series of exogenous and exchangeable signals but none of the parameters the agent considers possible corresponds to the true distribution, the posterior concentrates a.s. (with respect to the true distribution) on the subset of the parameter set that minimizes Kullback-Leibler divergence of beliefs to the true distribution.

Esponda-Pouzo [2016] Berk-Nash equilibrium

Each period a state is drawn according to a fixed distribution. Then each player privately observes her own signal (i.e. "type.") Then players simultaneously choose actions and each player observes her consequence (which includes at least her payoff.)

Agents needn't know either opponents' strategies or the distribution of Nature's moves or even the signal generating function. Instead, they have subjective beliefs about the consequences they will see as a function of their actions.

Berk-Nash equilibrium: each agent's strategy (map from type to action) is optimal given beliefs, and beliefs minimize the KL divergence from what the agent sees.

EP provide a learning-theoretic foundation for this equilibrium concept.

- Single agent in each player role. (appendix relaxes this).

- Agents believe they are facing a fixed steady state distribution.
- Each agent *i* has full support prior on finite-dimensional parameter space Θ^i .

- Each possible observation has positive probability under at least one parameter value.

- "Richness" condition on the subjective model: If a feasible event is deemed impossible by some parameter value, then that parameter value is not isolated.

- Agents myopically maximize.

- Perturb payoff functions with payoff shocks so that the best responses are a continuous function.

Defn. Strategy profile σ is *stable [or strongly stable]* if the sequence of intended strategies, converges to σ with positive probability [or with probability 1].

Lemma 2: If a behavior converges on a positive probability set of histories, beliefs converge on that set too.

Generalizes past results in the statistics literature to a setting where data are endogenously generated by their own actions and so are not exchangeable.

Proof takes some work. Then the main result is easier:

Theorem 2: If σ is stable under an optimal policy profile for a perturbed game, then it is a Berk–Nash equilibrium of the perturbed game.

Converse requires one more relaxation of the best responses ("asymptotically optimal")

Also that σ is weakly identified: for each player, all parameters that minimize the KL divergence at σ generate the same distribution of observations.

Theorem 3: If σ is a Berk–Nash equilibrium of a perturbed game that is weakly identified given σ, then there is a profile of priors with full support and an asymptotically optimal policy profile φ such that σ is strongly stable under φ.

Note: if agents aren't myopic, then their beliefs may not converge even though beliefs would converge with myopic agents and would also converge if the prior was correctly specified (Fudenberg, Romanyuk, Strack 2017).

Other recent papers on learning with mis-specified priors or models: Acemoglu, Chernozhukov, Yildiz [2016], Bohren-Hauser [2017], Heidhues,Koszegi, Strack [2017?], Liang [2017].

Omitted Material 3: ***Imitation Processes***:

Non-Bayeisan learning rules for strategic form games, where agents observe something about play or payoffs of others.

Like REL, these processes don't require agents to know payoff matrix, and are stochastic even w/o mutations.

Examples:

1) *Imitation + aspiration* Binmore-Samuelson [1997]: Realized payoff is game payoff + noise term. Each period an agent is picked at random to reevaluate his choice; sticks with his strategy if realized payoff exceeds an exogenous aspiration level and otherwise imitates at random.

2) *Imitation+payoff comparison* Benäim-Weibull [2003]: When an agent reevaluates, he picks another randomly chose nagent. If that agent's strategy is different, imitate with a probability that is increasing in the payoff difference.

3) *Imitate the most successful:* Agent observes the payoff of every other agent with i.i.d. noise, then picks strategy with highest observed payoff. If the noise terms have the appropriate extreme-value distribution, the probability that strategy i is chosen when there are x_i agents playing i and x_j playing j is $\frac{x_i \exp(\sigma u_i)}{x_i \exp(\sigma u_i) + x_j \exp(\sigma u_j)}$:

more popular strategies are more likely to be imitated.

4) *Frequency-dependent Moran process:*

When an agent playing i reconsiders, copies another agent playing j with probability $\frac{x_i u_i}{x_i u_i + x_j u_j}$. A lot like the previous process but respond to payoffs and not exponential of payoffs- so can move from one process to the other by transforming the payoff functions.

One interpretation: imitation + boundedly rational "popularity weighting". *Idea:* popularity weighting can be a proxy for memory, and lead to better performance; as in Ellison-Fudenberg [1993], [1995]. But the (frequency-independent) version of this process comes from population genetics.

These processes don't "slow down" so not analyzed with stochastic approximation. Instead the limit of the ergodic distribution is characterized using the tools of finite-state Markov chains, see Fudenberg and Imhof [2006].