Game Theory

Lecture 8: Refinements

Ludvig Sinander
based on slides by Péter Eső

Hilary Term 2025

University of Oxford

1. Dynamic games

All games can be represented in extensive form (tree + info sets).
Nash equilibrium may fail to eliminate empty threats, so we use:

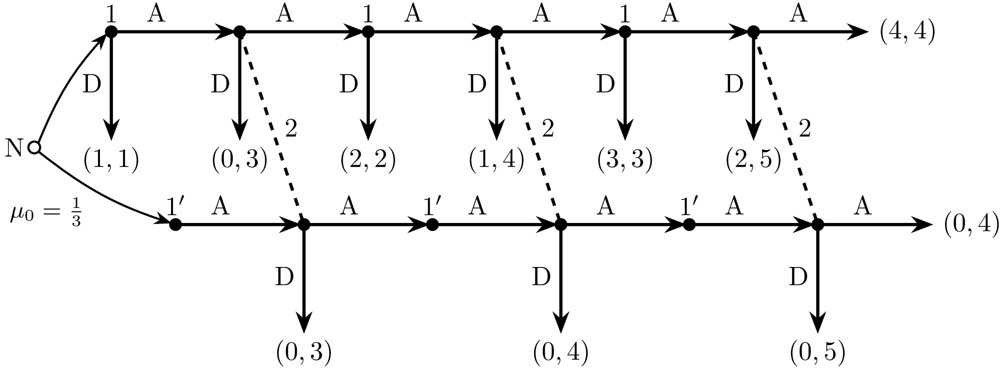
(1) **Subgame-perfect equilibrium:** Strategy profile that induces Nash in every subgame (=continuation from a single node without breaking any information sets), even after a deviation. In finite games with perfect info: same as backward induction.

(2) **Perfect Bayesian equilibrium:** Strategy profile <u>and</u> beliefs. Actions are sequentially rational given the beliefs and anticipated moves. On-path beliefs generated by eq'm strategies via Bayes's rule; off-path beliefs chosen by the modeller to support the PBE.

(3) **Sequential equilibrium:** Off-path beliefs must be the limit of Bayesian beliefs generated by fully-mixed strategies ('trembles'), otherwise same as PBE.

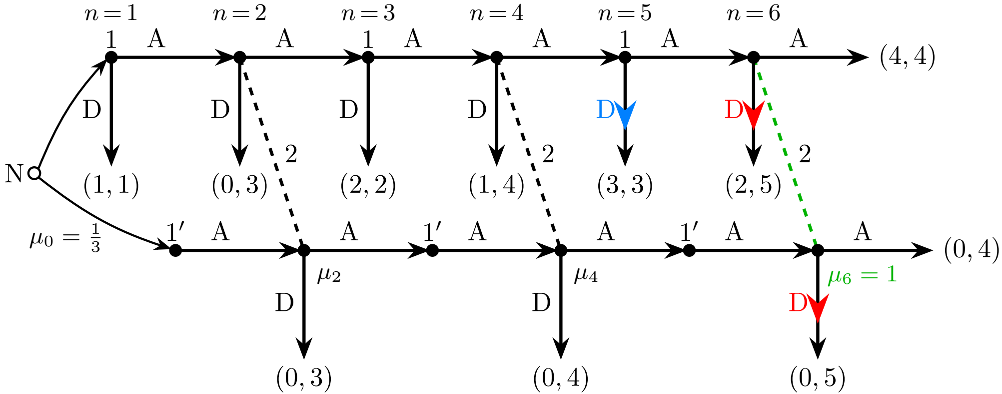
PBE is the most popular equilibrium concept in applied work.

Application from last time: Centipede with 'doubt'



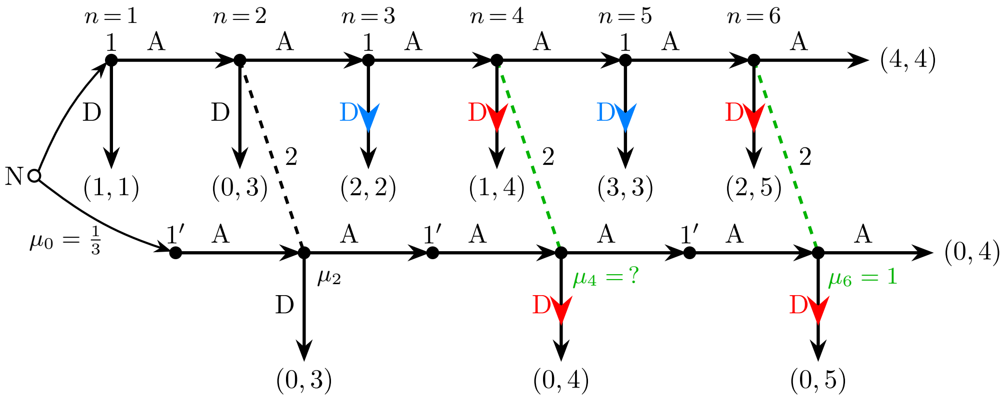
In a shortened version of Rosenthal's (1981) centipede game with 3 passes for each player, P2 (she) believes there is a $\mu_0 = \frac{1}{3}$ chance that P1 (he) is irrational (a 'commitment/crazy type') playing A at every node. P2's prior is commonly known. P1 knows whether he is rational or not.

The endgame



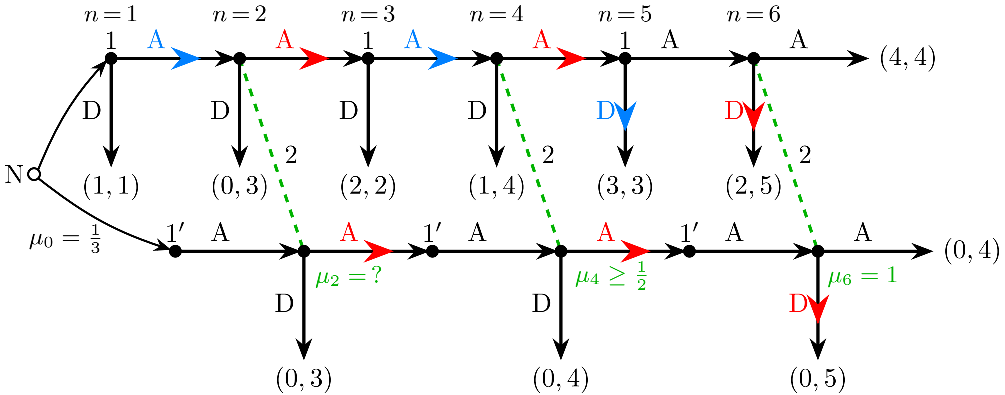
At node (info set) $n = 6$, it is optimal for P2 to play D irrespective of her beliefs. Hence at $n = 5$, rational P1 plays D as well. At $n = 6$, P2 infers that P1 is crazy, if he has played A at $n = 5$. Therefore, in any PBE, P2's belief at node 6 that P1 is crazy is $\mu_6 = 1$.

Why P2 won't play pure action D at $n = 4$ in any PBE



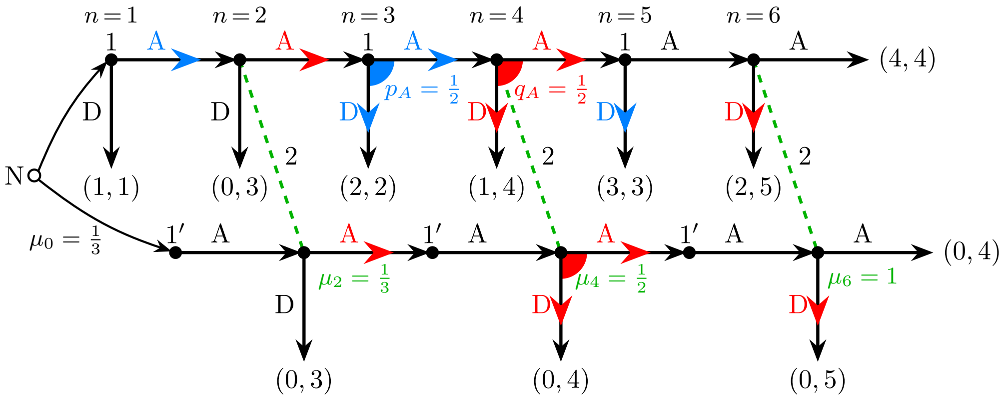
P2 plays pure D at $n = 4$ only if she weakly prefers D to A $\Leftrightarrow 4 \geq \mu_4 5 + (1 - \mu_4)3$ $\Leftrightarrow \mu_4 \leq \frac{1}{2}$. If P2 plays pure D at $n = 4$ then rational P1 plays D at $n = 3$ as well. Then P2 gets to make a move at $n = 4$ only if P1 is crazy, hence $\mu_4 = 1$. Contradiction!

Why P2 won't play pure action A at $n = 4$ in any PBE



P2 plays A at $n = 4$ only if $\mu_4 \geq \frac{1}{2}$. If she does then rational P1 also plays A at $n = 3$. Both types of P1 play A at $n = 3$, hence Bayes rule implies $\mu_2 \equiv \mu_4 \geq \frac{1}{2}$, and so P2 picks A at $n = 2$ as well. Then rational P1 plays A at $n = 1$, and so $\mu_2 = \mu_0 = \frac{1}{3}$. Contradiction!

Verify the following PBE (there is no other)



At $n = 1$, rational P1 plays A, hence $\mu_2 = \mu_0 = \frac{1}{3}$. At $n = 2$, P2 also plays A.
At $n = 3$, rational P1 mixes 50–50%, this makes P2 update her beliefs to $\mu_4 = \frac{1}{2}$.
At $n = 4$, P2 also mixes, with 50–50% as well. Beyond that point both players play D.

Steps for checking the solution

In any PBE, we must make P2's beliefs at $n = 4$ equal $\mu_4 = \frac{1}{2}$, so that she can mix.

To that end, let rational P1 play A with probability $p_A = \frac{1}{2}$ at $n = 3$.

Then Bayes's rule yields $\mu_4 = \frac{\mu_2}{\mu_2 + (1-\mu_2)p_A} = \frac{\frac{1}{3}}{\frac{1}{3}\times 1 + \frac{2}{3}\times\frac{1}{2}} = \frac{1}{2}$, as desired.

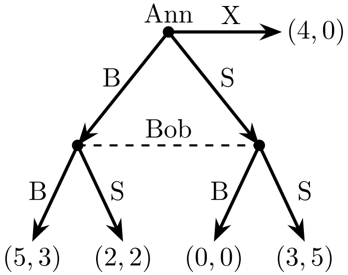
Final step: At $n = 2$, with beliefs $\mu_2 = \frac{1}{3}$, P2 prefers A, because

$$\frac{2}{3}\left[\frac{1}{2}\cdot 2 + \frac{1}{2}\cdot\left(\frac{1}{2}\cdot 4 + \frac{1}{2}\cdot 3\right)\right] + \frac{1}{3}\left(\frac{1}{2}\cdot 4 + \frac{1}{2}\cdot 5\right) = 3.\dot{3} > 3.$$

Interpretation of the PBE/SE of the game: Initially, rational P1 plays A to make P2 believe he is likely to be crazy. P2 rationally goes along, but the players 'burst the bubble' before the end. More detail in course notes and especially Kreps's textbook, chapter 14.6 (pp. 536–43)

2. Why we need ‘forward induction’: BoS with option X

Consider BoS where Ann has an attractive outside option:

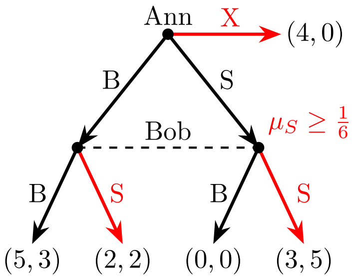


	B	S
B	$\underline{5}$, $\underline{3}$	2, 2
S	0, 0	3, $\underline{5}$
X	4, $\underline{0}$	$\underline{4}$, $\underline{0}$

Intuitively, Ann’s outside option, with a payoff that beats coordinating on S, should help her achieve (B, B). Bob must know Ann won’t ‘come in’ unless she expected to get more than 4.

But (X, S) is Nash/SPE—and also PBE and SE (see the next two slides).

How (X, S) can be sustained in PBE



	B	S
B	<u>5</u>, <u>3</u>	2, 2
S	0, 0	3, <u>5</u>
X	4, <u>0</u>	<u>4</u>, <u>0</u>

If Bob believes $\mu_S \geq 1/6$ then he plays S. Anticipating that, Ann plays X.
Hence (X, S) is PBE with $\mu_S \geq 1/6$.

Of course, (B, B) is also PBE with $\mu_S = 0$.

The next slide shows how to support (X, S) in sequential equilibrium.

(X, S) in SE and why it is 'counter-intuitive'

For $m = 1, 2, \ldots$, let $\sigma_A^m = \frac{1}{m^2}\mathrm{B} + \frac{1}{m}\mathrm{S} + (1 - \frac{1}{m^2} - \frac{1}{m})\,\mathrm{X}$, and let $\sigma_B^m = \frac{1}{m}\mathrm{B} + (1 - \frac{1}{m})\mathrm{S}$.

Using Bayes's rule, Bob's belief that Ann has played S conditional on Ann having played either B or S 'by mistake' is

$$\mu_S^m = \Pr[\mathrm{S}|\text{not X}] = \frac{1/m}{1/m^2 + 1/m} = \frac{m}{m+1}.$$

As $m \to \infty$: $\sigma_A^m \to \mathrm{X}$, $\sigma_B^m \to \mathrm{S}$, $\mu_S^m \to 1$, hence (X, S) with $\mu_S = 1$ is sequential equilibrium.

But: Ann guarantees 4 by playing X. By not playing X she must be expecting more!
Only B (followed by B) can yield more for her. Hence, intuitively, shouldn't we let $\mu_S = 0$?

This is a 'forward induction' argument that goes beyond PBE/SE.

3. Beer–Quiche game (I.-K. Cho & D. Kreps, 1987)

A pub-goer (P1) orders either Beer or Quiche. A bully (P2) sees this and either fights him or not. P2 wants to fight if P1 is wimpy but not if P1 is strong. Only P1 knows his type; the prior prob's are 20% wimpy, 80% strong. Wimpy P1 likes Quiche, strong P1 likes Beer; neither type of P1 wants a fight.

Players: Nature, P1, P2.

Histories & actions:

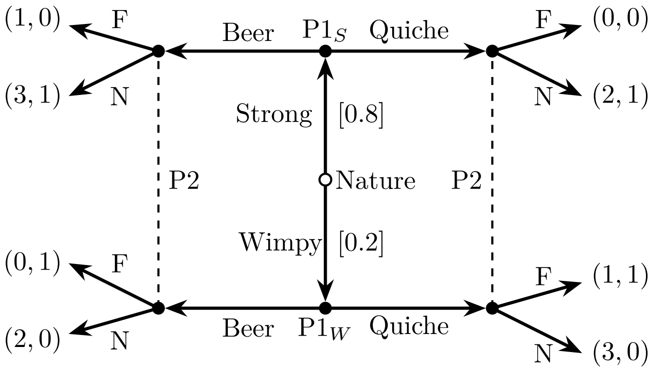
0. Nature picks P1's type, wimpy ($\mathrm{P1}_W$, 20%) or strong ($\mathrm{P1}_S$, 80%).
1. P1 observes his type, chooses either Beer or Quiche.
2. P2 observes P1's action but not his type, picks Fight (F) or Not (N).

Payoffs (starting from 0 baseline):

– P1 gets +1 for consuming favourite item, +2 for avoiding a fight.

– P2 gets +1 for fighting wimpy or <u>not</u> fighting strong P1.

Beer–Quiche in extensive form



Starting from 0 baseline payoffs, P1 gets +1 for consuming favourite item and +2 for avoiding a fight, whereas P2 gets +1 for fighting $\mathrm{P1}_W$ or <u>not</u> fighting $\mathrm{P1}_S$.

See the game's strategic form with interim payoffs on the next slide (skippable).

*Beer–Quiche in strategic form; interim payoffs

Use B and b for Beer by $P1_S$ and $P1_W$, also Q and q for Quiche by $P1_S$ and $P1_W$; distinguish between F (after Beer) and $\mathcal{F}$ (after Quiche) by P2, same for N and $\mathcal{N}$.

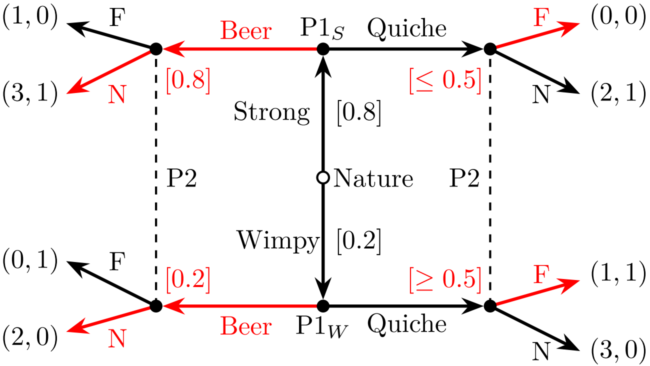
P1 and P2 each have four strategies. Payoffs written in each cell in the order $P1_S$, $P1_W$, P2.

	$F\mathcal{F}$	$F\mathcal{N}$	$N\mathcal{F}$	$N\mathcal{N}$
Bb	$\underline{1}$, 0, 0.2	0, 1, 0.2	$\underline{3}$, $\underline{2}$, $\underline{0.8}$	$\underline{3}$, 2, $\underline{0.8}$
Bq	$\underline{1}$, $\underline{1}$, 0.2	1, $\underline{3}$, 0.0	$\underline{3}$, 1, $\underline{1.0}$	$\underline{3}$, $\underline{3}$, 0.8
Qb	0, 0, 0.2	$\underline{2}$, 0, $\underline{1.0}$	0, $\underline{2}$, 0.0	2, 2, 0.8
Qq	0, $\underline{1}$, 0.2	$\underline{2}$, $\underline{3}$, $\underline{0.8}$	0, 1, 0.2	2, $\underline{3}$, $\underline{0.8}$

Two pure-strategy Bayesian Nash equilibria: (Bb, $N\mathcal{F}$) and (Qq, $F\mathcal{N}$).

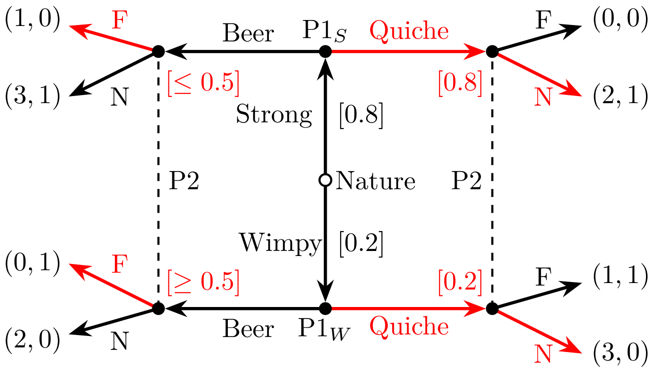
Can these be supported in PBE/SE with appropriate beliefs?

‘Pooling’ PBE where both types of P1 drink beer



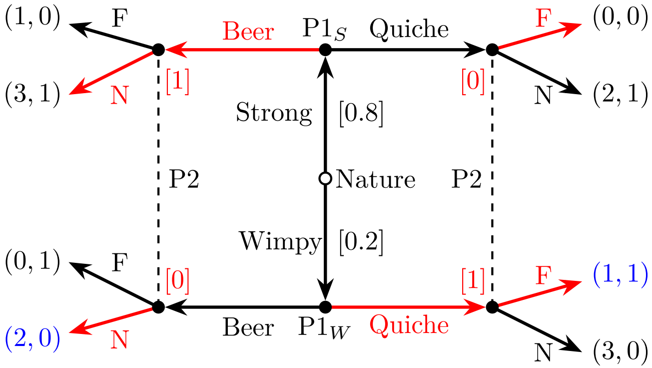
If both types of P1 choose Beer then P2 believes $\Pr[\mathrm{P1}_S|\mathrm{Beer}] = 0.8$ by Bayes’s rule.
Hence P2 does not fight after Beer. We are free to chose beliefs after Quiche.
Set $\Pr[\mathrm{P1}_S|\mathrm{Quiche}] \leq 0.5$ in order to make F after Quiche rational for P2.

‘Pooling’ PBE where both types of P1 have quiche



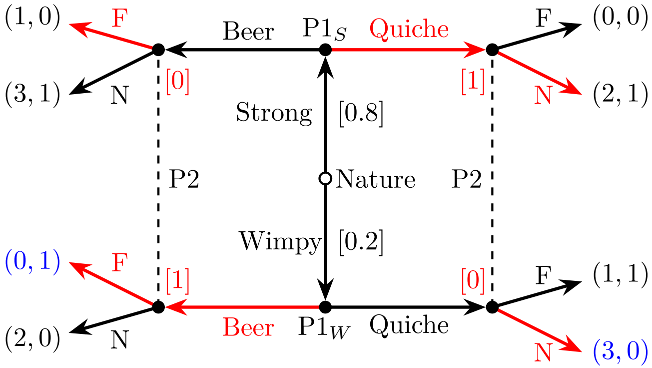
If both types of P1 choose Quiche then P2 believes $\Pr[\mathrm{P1}_S|\text{Quiche}] = 0.8$ by Bayes’s rule.
Hence P2 does not fight after Quiche. We are free to chose beliefs after Beer.
Set $\Pr[\mathrm{P1}_S|\text{Beer}] \leq 0.5$ in order to make F after Beer rational for P2.

Why no 'separating' PBE where $S \mapsto$ Beer and $W \mapsto$ Quiche



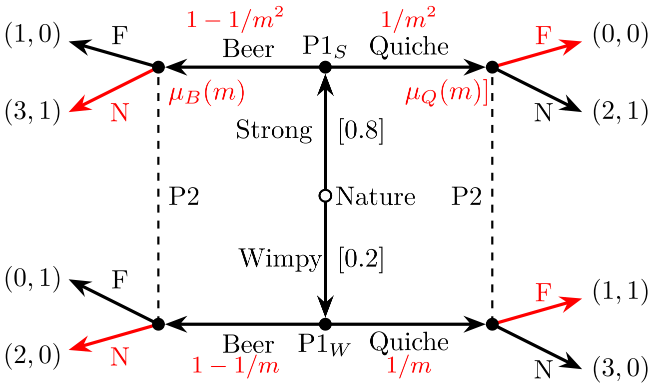
Suppose that $\mathrm{P1}_S$ chooses Beer and $\mathrm{P1}_W$ chooses Quiche in some PBE.
This pins down P2's beliefs and best responses (shown in red in the figure).
Then $\mathrm{P1}_W$ has a strict incentive to deviate to Beer (to avoid F), a contradiction.

Why no ‘separating’ PBE where $S \mapsto$ Quiche and $W \mapsto$ Beer



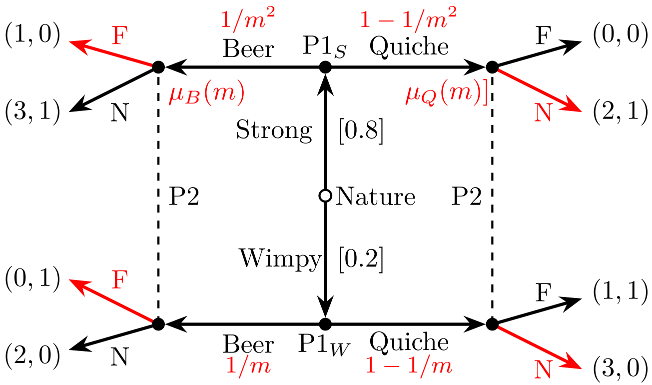
Suppose $\mathrm{P1}_S$ chooses Quiche and $\mathrm{P1}_W$ chooses Beer in some PBE.
This pins down P2’s beliefs and best responses (shown in red in the figure).
Then $\mathrm{P1}_W$ has a strict incentive to deviate to Quiche (to avoid F), a contradiction.

Sequential equilibrium construction for pooling on Beer



For $m = 1, 2, \ldots$, let $P1_S$ play Beer with probability $1 - 1/m^2$ and Quiche with $1/m^2$; let $P1_W$ play Beer and Quiche with probabilities $1 - 1/m$ and $1/m$, respectively. As $m \to \infty$, $\mu_B(m) \to 0.8$ and $\mu_Q(m) \to 0$. (Feel free to check this at home.)

Sequential equilibrium construction for pooling on Quiche



For $m = 1, 2, \ldots$, let P1$_S$ play Quiche with probability $1 - 1/m^2$ and Beer with $1/m^2$;
let P1$_W$ play Quiche and Beer with probabilities $1 - 1/m$ and $1/m$, respectively.
As $m \to \infty$, $\mu_Q(m) \to 0.8$ and $\mu_B(m) \to 0$. (Feel free to check this at home.)

Why pooling on Quiche is counter-intuitive

Note that in this equilibrium $P1_S$ gets 2, $P1_W$ gets 3 units of payoff; if P1 deviates to Beer then P2 believes he is more likely to be wimpy ($P1_W$) than strong ($P1_S$), hence fights him.

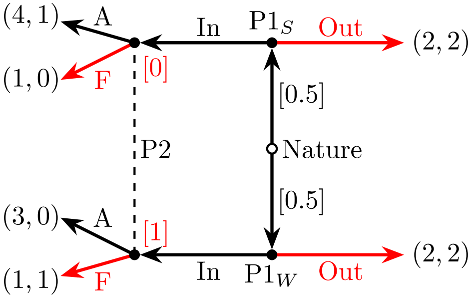
$P1_W$ cannot possibly gain by deviating to Beer: his equilibrium payoff is as high as it can be. Then why does P2 put at least 50% belief on this type upon observing a deviation to Beer?

This kind of reasoning is called **forward induction** (idea & term due to E. Kohlberg). How to formalise? One simple criterion:

Intuitive Criterion (I.-K. Cho & D. Kreps, 1987)**:** Upon observing a deviation, we must not put positive probability (beliefs) on types whose equilibrium payoff exceeds all the possible payoffs that they could get from the deviation, assuming the opponent's reaction to the deviation is rational for some beliefs.

Result (Cho–Kreps '87)**:** In the Beer–Quiche game, the PBE/SE outcome selected by the Intuitive Criterion is pooling on Beer.

4. Beyond PBE/SE & Intuitive Criterion



Pooling on Out, with $\Pr[P1_S|\text{In}] = 0$ and F by P2, is PBE/SE satisfying the Intuitive Criterion, because both types of P1 could potentially gain from deviation to In (if P2 played A not F).

This equilibrium 'feels wrong' as P1$_S$ gains from deviating to In for a larger set of mixed replies of P2 than P1$_W$ does. Also a 'forward induction' argument, but not captured by the Intuitive Criterion. Formalised as *Divinity Criterion* (J. Banks & J. Sobel, 1987).

Robustness to small mistakes (‘trembles’)

The common question addressed by refinements of out-of-equilibrium beliefs is what to infer from a deviation.

Overarching idea: players interpret deviations as *accidental mistakes.* (Explicit in SE.)

Suppose each player i must play every available action a with a probability of at least $\varepsilon_{i,a} > 0$; all players best respond subject to these minimal mistake probabilities, or ‘trembles’

Trembling-hand perfect equilibrium (TPE): there exists some vanishing sequence of thresholds $\varepsilon_{i,a}$ along which Nash play subject to such trembles converges to the TPE

- proposed by R. Selten in 1975; PBE/SE coincide with TPE in most games

Strategic stability (E. Kohlberg & J.-F. Mertens, 1986): for every vanishing sequence of mistake probabilities, optimal play subject to such trembles converges to the proposed stable strategies

An equilibrium is stable if it is robust to all types of mistakes. Issue: existence. . .

Take-away on extensive-form refinements

Some PBE / SE may be unattractive due to counter-intuitive out-of-equilibrium beliefs.

Refinements of PBE/SE was a prominent research topic in the 1990s.

- 'forward induction' belief criteria: Intuitive Criterion, Divinity (D1, D2, Universal), iterated deletion of weakly dominated strategies, extensive-form rationalisability …
- robustness to 'trembles': TPE, proper equilibrium, persistent equilibrium, stability, …

The 'holy grail' of equilibrium concepts satisfying all ideal requirements was never found…

In applications to signalling games, forward-induction criteria are used (which particular one varies by context).

In this course (on exam) you need to be able to use **PBE+Intuitive Criterion**.

5. Signalling: Brief history of thought

Thorstein Veblen (1899), book titled *The Theory of the Leisure Class.*

– satirical take on consumerism; coined the term ‘conspicuous consumption’ (waste to display status), described ‘Veblen goods’ (goods with demand increasing in price)

Potlatch: Custom of indigenous peoples of the Pacific Northwest.

– feast to display (and distribute?) wealth

– banned by US & Canada in 1885, ‘worse than useless custom’

Model of labour market signalling by Michael Spence in 1973, Nobel Prize in 2001.

Signalling in evolutionary biology: Amotz Zahavi’s *handicap principle* in 1975.

The idea

In situations with adverse selection, actions speak louder than words.

Credible signal (recall from Micro): An action that is cheap enough for the ‘good’ but too expensive for the ‘bad’ type to take.

Real-life examples of signalling ...

- fitness by wasting energy (gazelle’s stotting, peacock’s tail)
- criminality by getting tattoos, using drugs
- productivity by obtaining education
- low risk of accident by agreeing to high excess (called ‘deductible’ in USA)
- safe borrower (=likely to repay loan) by offering a collateral
- firm’s future profitability by buying back shares.

Review Spence's (1973) signalling game

The game, familiar from 2nd-year Microeconomics:

(1) Nature picks worker's type $\theta \in \{L, H\}$, $H > L \geq 0$ with $\Pr(\theta = H) = \lambda$

(2) worker observes θ and chooses education level $e \geq 0$

(3) (at least two) firms observe e but not θ and set wage w

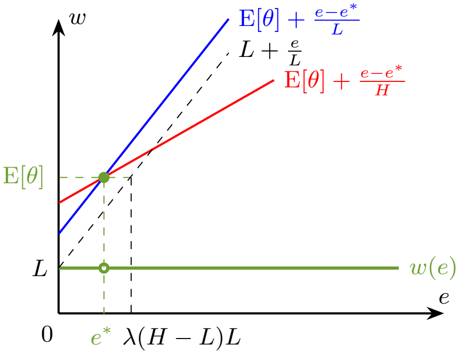
(4) worker accepts at most one offer.

Worker's payoff: $w - e/\theta$; zero outside option.
Firm's payoff is $\theta - w$ if it employs the worker, 0 if it does not.

PBE? $\{e_L, e_H, w(e), \mu(e) \text{ for all } e \geq 0\}$, where $\mu(e) = \Pr[\theta = H|e]$.

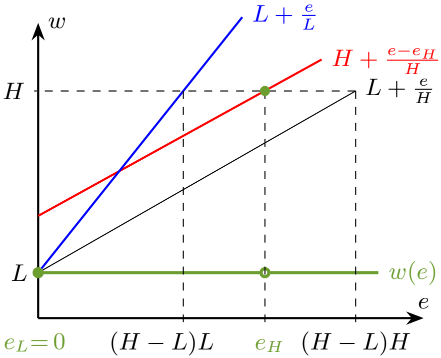
For revision: Find all perfect Bayesian equilibria of this game.
Then apply the Intuitive Criterion to select a unique outcome.

Example of a pooling equilibrium



Both types set $e_L = e_H = e^*$. Hence $\mu(e^*) = \lambda$, $w(e^*) = \lambda H + (1-\lambda)L$. For any $e' \neq e^*$ (out of equilibrium education level) set $\mu(e') = 0$; therefore $w(e') = L$ (low wage).
Pooling action must not be too high, $e^* \leq \lambda(H-L)L$, otherwise L deviates to $e' = 0$.

Example of a separating equilibrium

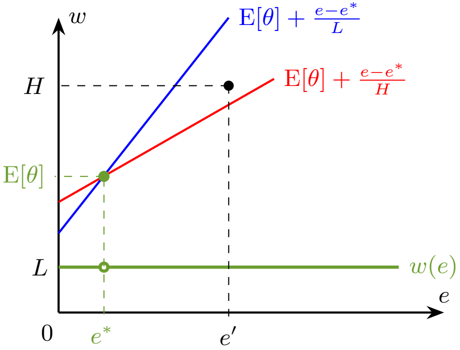


In any separating PBE, $e_L = 0$ (why signal low type) and $e_H > 0$, hence $\mu(0) = 0$, $\mu(e_H) = 1$.

Out-of-equilibrium beliefs: set $\mu(e') = 0$ for all $e' \neq e_H$ as well.

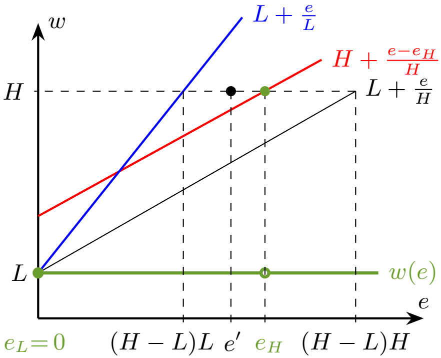
High type must select sufficiently high e_H to separate: $(H - L)L \leq e_H \leq (H - L)H$.

Intuitive Criterion rules out pooling...



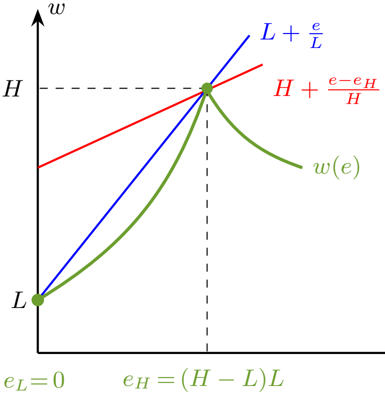
Only type H could gain from deviating to a sufficiently high e' (if perceived as high type, contrary to what off-path beliefs in the pooling PBE specify) as shown in the figure. Hence, by the Intuitive Criterion $\mu(e') = 1$, so $w(e') = H$, breaking the equilibrium.

...and all non-minimal-cost separating equilibria



Only type H could gain from deviating to $(H-L)L < e' < e_H$ (if perceived as coming from the high type which is not what we assumed in the PBE) as shown in the figure.
Hence, by the Intuitive Criterion $\mu(e'){=}1$, so $w(e') = H$, breaking the equilibrium.

The minimal-cost separating outcome survives!



$w(e_L) = L$ and $w(e_H) = H$ by Bayes's rule; $w(e)$ is drawn to be continuous just for fun.
H type distorts e just enough so that L type won't imitate. Robust!

Take-away on Spencian signalling

Structure of a Spencian signalling game:

– Sender wants the receiver think he is of high type, H.

– Single-crossing: Greater action is relatively cheaper for H.

With two types, perfect Bayesian equilibrium + Intuitive Criterion yield the **minimal-cost separating outcome**. (Also called the 'Riley outcome' after John Riley.)

Same outcome as that of competitive screening (2nd year Micro!).

Two misconceptions about signalling:

(1) Signal has to be costly. (Also works if e is *more beneficial* for H than L.)

(2) Wasteful, must be banned. (True that it's always excessive, but it may resolve a missing market problem due to adverse selection—see Akerlof's 'market for lemons'.)