Equilibrium Refinements

Mihai Manea

MIT

Sequential Equilibrium

- In multi-stage games where payoffs depend on initial moves by nature, the only subgame is the original game... subgame perfect equilibrium = Nash equilibrium
- Play starting at an information set can be analyzed as a separate subgame if we specify players' beliefs about at which node they are.
- Based on the beliefs, we can test whether continuation strategies form a Nash equilibrium.
- Sequential equilibrium (Kreps and Wilson 1982): way to derive plausible beliefs at every information set.

An Example with Incomplete Information

Spence's (1974) job market signaling game

- The worker knows her ability (productivity) and chooses a level of education.
- Education is more costly for low ability types.
- Firm observes the worker's education, but not her ability.
- The firm decides what wage to offer her.

In the spirit of subgame perfection, the optimal wage should depend on the firm's beliefs about the worker's ability given the observed education.

An equilibrium needs to specify contingent actions and beliefs.

Beliefs should follow Bayes' rule on the equilibrium path.

What about off-path beliefs?

An Example with Imperfect Information

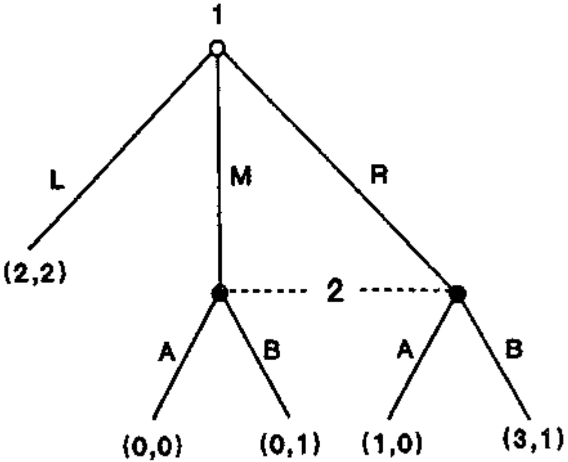


Figure: (L, A) is a subgame perfect equilibrium. Is it plausible that 2 plays A?

Assessments and Sequential Rationality

Focus on extensive-form games of perfect recall with finitely many nodes.

An assessment is a pair (σ, μ)

- σ: (behavior) strategy profile
- $\mu = (\mu(h) \in \Delta(h))_{h \in H}$: system of beliefs

$u_i(\sigma | h, \mu(h))$: i's payoff when play begins at a node in h randomly selected according to $\mu(h)$, and subsequent play specified by σ.

The assessment (σ, μ) is sequentially rational if

$$u_{i(h)}(\sigma_{i(h)}, \sigma_{-i(h)} | h, \mu(h)) \geq u_{i(h)}(\sigma'_{i(h)}, \sigma_{-i(h)} | h, \mu(h))$$

for all information sets h and alternative strategies σ'.

Consistency

Beliefs need to be consistent with strategies.

$\tilde{\sigma}$ is totally mixed if $supp(\tilde{\sigma}_{i(h)}(h)) = A(h)$, i.e., all information sets are reached with positive probability.

Bayes' rule $\rightarrow$ unique system of beliefs $\mu^{\tilde{\sigma}}$ for any totally mixed $\tilde{\sigma}$.

The assessment (σ, μ) is consistent if there exists a sequence of totally mixed strategy profiles $(\sigma^m)_{m \geq 0} \rightarrow \sigma$ s.t. $(\mu^{\sigma^m})_{m \geq 0} \rightarrow \mu$.

Definition 1

A *sequential equilibrium* is an assessment that is sequentially rational and consistent.

Implications of Sequential Rationality

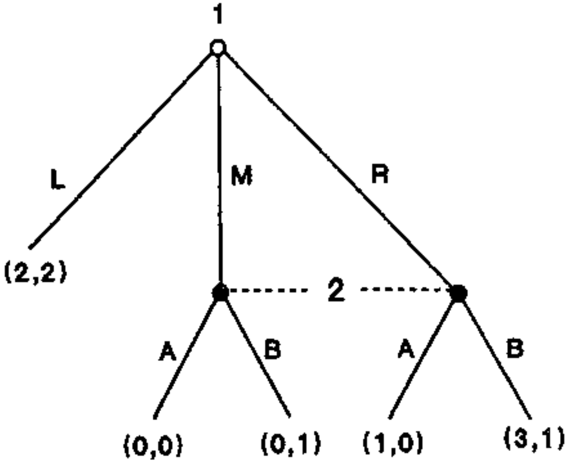


Figure: No belief rationalizes A. 2 plays B, 1 optimally chooses R.

Implications of Consistency

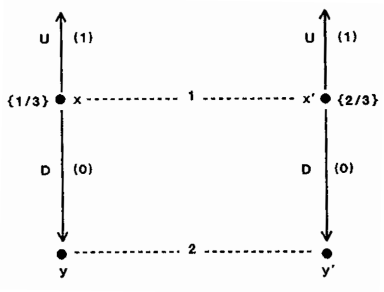


Figure: By consistency, $\mu(y|h_2) = \mu(x|h_1)$, even though D is never played.

Consistency → common beliefs after deviations from equilibrium behavior.

Why should different players have the same theory about something not supposed to happen?

Consistency matches the spirit of equilibrium analysis, which assumes players hold identical beliefs about others' strategies.

Existence of Sequential Equilibrium

Theorem 1

A sequential equilibrium exists for every finite extensive-form game.

Follows from existence of perfect equilibria, prove later.

Continuity

Proposition 1

The sequential equilibrium correspondence has a closed graph with respect to payoffs.

- $(u^k)_{k\geq 0} \to u$: convergent sequence of payoff functions
- (σ^k, μ^k): sequential equilibrium for u_k
- If $(\sigma^k, \mu^k)_{k\geq 0} \to (\sigma, \mu)$, is (σ, μ) a sequential equilibrium for u?
- (σ, μ) is sequentially rational because the expected payoffs conditional on reaching any information set are continuous in payoff functions and beliefs.
- Let $(\sigma^{m,k}, \mu^{m,k})_{m\geq 0} \to (\sigma^k, \mu^k)$ be a convergent sequence of completely mixed strategy profiles and corresponding induced beliefs.
- Find m_k s.t. $\sigma^{m_k,k}$ and $\mu^{m_k,k}$ are within $1/k$ from corresponding components of σ^k and μ^k.
- $(\sigma^k, \mu^k)_{k\geq 0} \to (\sigma, \mu) \Rightarrow (\sigma^{m_k,k}, \mu^{m_k,k}) \to (\sigma, \mu)$, so (σ, μ) is consistent.

Sequential Equilibrium Multiplicity

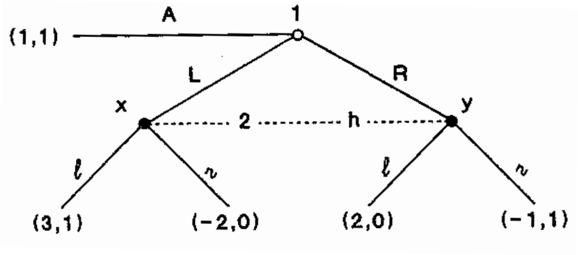
Theorem 2

For generic payoff functions, the set of sequential equilibrium outcome distributions is finite.

Set of sequential equilibrium assesments often infinite

- Infinitely many belief specifications at off-path information sets supporting some equilibrium strategies.
- Set of sequential equilibrium strategies may also be infinite. Off-path information sets may allow for consistent beliefs that make players indifferent between actions... many mixed strategies compatible with sequential rationality.

Example



Courtesy of The MIT Press. Used with permission.

Sequential equilibrium outcomes: (L, l) and A

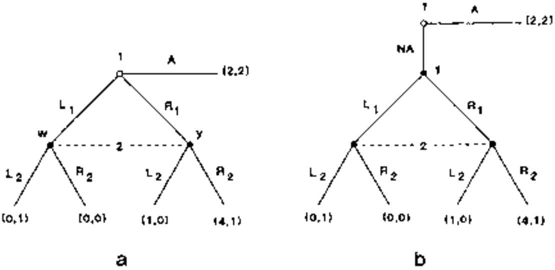
Unique equilibrium leading to (L, l)

Two families of equilibria with outcome $A \ldots 2$ must choose r with positive probability

1. 2 chooses r with probability 1 and believes $\mu(x) \in [0, 1/2]$
2. 2 chooses r with probability in $[2/5, 1]$ and believes $\mu(x) = 1/2$

Kohlberg and Mertens' (1986) Critique

"Strategically neutral" changes in game tree affect equilibria.



Game *a*: (A, L_2) possible in a sequential equilibrium

Game *b*: $((NA, R_1), R_2)$ unique sequential equilibrium strategies. In subgame following NA, R_1 strictly dominates L_1. Then 2 chooses R_2, and 1 best responds with (NA, R_1).

Response

Sensitivity of sequential equilibrium to “irrelevant moves” is not a consequence of consistency, but of sequential rationality... problem present even for subgame perfect equilibria.

Kohlberg and Mertens’ solution: stable equilibria

- Theory of robustness with respect to any profile of small mistakes, solution depending only on the strategic form
- If players make mistakes at every information set, are the two extensive forms equivalent?
  - Game *a*: 1 might play either L_1 or R_1 by mistake intending to choose A.
  - Game *b*: if 1 makes the mistake of not playing A, he is still able to ensure that R_1 is more likely than L_1... why would first mistake be correlated with the second?

Perfect Bayesian Equilibrium

Perfect Bayesian equilibrium (PBE): original solution concept for extensive-form games with imperfect/incomplete information.

Sequential equilibrium now preferred, but worthwhile to know about PBE (used in early/applied research).

PBE similar to sequential equilibrium with fewer restrictions on beliefs

- Strategies: sequentially rational
- Beliefs: derived from Bayes' rule wherever applicable
- Simplest version: no constraints on off-path beliefs

Other Restrictions on Off-Path Beliefs

Fudenberg and Tirole (1991): other restrictions on beliefs in multi-stage games with incomplete information (all implied by consistency)

- If player types are drawn independently by nature, beliefs about different players should remain independent after every history.
- Updating should be "consistent": given a probability-zero history h_t at which strategies call for a positive probability transition to h_{t+1}, beliefs at h_{t+1} should be given by updating beliefs at h_t using Bayes' rule.
- "Not signaling what you don't know": with independent types, beliefs about player i's type at the beginning of period $t+1$ depend only on h_t and i's action at t, not on other players' actions at t.
- Players $i \neq j$ should have the same belief about a third player k even after probability zero histories.

Perfect Equilibrium

	L	R
U	1,1	0,0
D	0,0	0,0

Selten (1975): (trembling-hand) perfect equilibrium

- Both (U, L) and (D, R) are Nash equilibria.
- (D, R) not robust to small mistakes: if 1 thinks that 2 might make a mistake and play L with positive probability, deviate to U.

Definition 2

In a strategic-form game, a profile σ is a *perfect equilibrium* if there is a sequence of *trembles* $(\sigma^m)_{m\geq 0} \to \sigma$, where each σ^m is a totally mixed strategy, such that σ_i is a best reply to σ^m_{-i} for each m and all $i \in N$.

Existence of Perfect Equilibria

Definition 3

σ^ε is an ε-*perfect equilibrium* if $\exists \varepsilon(s_i) \in (0, \varepsilon], \forall i \in N, s_i \in S_i$ s.t. σ^ε is a Nash equilibrium of the game where players are restricted to play mixed strategies in which every pure strategy s_i has probability at least $\varepsilon(s_i)$.

Proposition 2

A strategy profile is a perfect equilibrium iff it is the limit of a sequence of ε-perfect equilibria as $\varepsilon \to 0$.

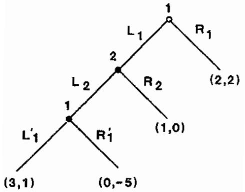
Theorem 3

Every finite strategic-form game has a perfect equilibrium.

Proof.

A $1/n$-perfect equilibrium exists by the general Nash equilibrium existence theorem. By compactness, the sequence of $1/n$-perfect equilibria has a convergent subsequence as $n \to \infty$. The limit is a perfect equilibrium. □

Perfection in Strategic Form ⇏ Subgame-Perfection



a. Extensive Form

	L_2	R_2
R_1	2,2	2,2
L_1,L_1'	3,1	1,0
L_1,R_1'	0,−5	1,0

b. Strategic Form

Unique SPE: $(L_1 L_1', L_2)$

(R_1, R_2) is perfect in strategic form, sustained by trembles s.t. after trembling to L_1, player 1 chooses R_1' vs. L_1' with probability ratio $\geq 1/5$. Correlation in trembles at different information sets...unreasonable.

Perfection in Extensive-Form Games

Solution: agent-normal form

- A different player for every information set h.
- "Player" h has the same payoffs as $i(h)$.

Definition 4

A *perfect equilibrium* for an extensive-form game is a perfect equilibrium of its agent-normal form.

Connection to Sequential Equilibrium

Theorem 4

Every perfect equilibrium of a finite extensive-form game is a sequential equilibrium (for some appropriately chosen beliefs).

- σ: perfect equilibrium of the extensive-form game $\Rightarrow \exists(\sigma^m)_{m\geq 0} \to \sigma$ totally mixed strategies in the agent-normal form s.t. σ_h is a best reply to σ^m_{-h} for each m and all information sets h.
- By compactness, $(\mu^{\sigma^m})_{m\geq 0}$ has a convergent subsequence, denote limit by μ.
- By construction, (σ, μ) is consistent.
- σ_h is a best response to $\mu^{\sigma^m}(h)$ and σ^m_{-h} for each m.
- By continuity, σ_h is a best response to $\mu(h)$ and σ_{-h}.
- One-shot deviation principle: (σ, μ) is sequentially rational. □

Properties of Perfect Equilibrium

Kreps and Wilson (1982): every sequential equilibrium is perfect for generic payoffs.

The set of perfect equilibrium outcomes does not have a closed graph.

	L	R
U	1,1	0,0
D	0,0	1/n,1/n

(D, R) is perfect for $n > 0$. In the limit $n \to \infty$, only (U, L) is perfect.

Order-of-limits problem

- As $n \to \infty$, the trembles against which D and R remain best responses become smaller and smaller.
- (D, R) is a reasonable prediction in the limit game if the approximation error in describing payoffs is greater than players' mistakes.

Proper Equilibrium

Myerson (1978): a player is infinitely more likely to tremble to better actions

A player's probability of playing the second-best action is at most ε times the probability of the best, the probability of the third-best action is at most ε times the probability of the second-best...

Definition 5

An ε-*proper equilibrium* is a totally mixed strategy profile σ^ε s.t. if $u_i(s_i, \sigma^\varepsilon_{-i}) < u_i(s'_i, \sigma^\varepsilon_{-i})$, then $\sigma^\varepsilon_i(s_i) \leq \varepsilon \sigma^\varepsilon_i(s'_i)$. A *proper equilibrium* is any limit of ε-proper equilibria as $\varepsilon \to 0$.

Theorem 5

Every finite strategic-form game has a proper equilibrium.

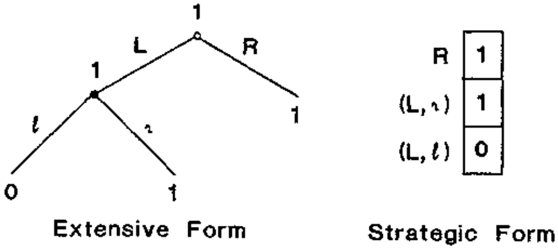
Prove existence of ε-proper equilibria applying Kakutani's fixed point theorem to "mistake hierarchy ε-best response" correspondences, then use compactness to find a limit point.

Properties of Proper Equilibrium

Given an extensive-form game, a proper equilibrium of its strategic form is automatically subgame-perfect (backward induction argument).

Kohlberg and Mertens (1986): a proper equilibrium in a normal-form game is sequential in every extensive-form game having the given normal form.

However, not necessarily a trembling-hand perfect equilibrium in the agent-normal form of every such game.



Courtesy of The MIT Press. Used with permission.

(L, r) is proper in the strategic form but not perfect in the extensive form.

Forward Induction

Equilibrium: off-path observations interpreted as errors

Forward induction: players should believe in the rationality of their opponents even after observing deviations.

- When a player deviates from equilibrium strategies, the opponent should believe that the player expects follow up play that makes the deviation reasonable.
- The deviation is informative about the player's type or, in general extensive form games, about his future play.

Forward induction not an equilibrium concept: in equilibrium, all players expect specified strategies to be exactly followed

An attempt to describe strategic uncertainty. . . no single, rigorous definition

Example

1 chooses between O, which generates payoffs $(2, 2)$, or I, which leads to

	T	W
T	0,0	3,1
W	1,3	0,0

SPE: (OW, T). Reasonable?

- If 1 plays I, this suggests he does not intend to follow up with W: O yields a payoff of 2, while W leads to a payoff of at most 1 for player 1.
- Player 2, anticipating that 1 will play T, should play W.
- If 1 can convince 2 to play W, he gets the higher payoff from (T, W).

Forward Induction and Strict Dominance

Reduced normal form

	T	W
O	2,2	2,2
IT	0,0	3,1
IW	1,3	0,0

(O, T) is a perfect (in fact, proper) equilibrium.

If we rule out IW because it is s. dominated by O, then the only perfect equilibrium is (IT, W).

An equilibrium concept that is not robust to deletion of s. dominated strategies is troubling...

Kohlberg and Mertens (1986): stable equilibria

Requirements

- Iterated dominance: every stable set must contain a stable set of any game obtained by deleting a s. dominated strategy
- Admissibility: no mixed strategy in a stable set assigns positive probability to a weakly dominated strategy
- Invariance to extensive-form representation: stable sets depend only on the strategic form

Stability: necessarily set-valued

	L	R
U	3,2	2,2
M	1,1	0,0
D	0,0	1,1

Both M and D are strictly dominated. Depending on which one is eliminated first, L or R becomes weakly dominated... both (U, L) and (U, R) must be contained in the solution.

Stable Equilibria

Definition 6

A closed set S of Nash equilibria in a finite strategic-form game is *strategically stable* if it is minimal among sets with the property that for every $\eta > 0$, there exists $\varepsilon > 0$ s.t. for all choices of $\varepsilon(s_i) \in (0, \varepsilon)$, the game where each player i is constrained to play every s_i with probability at least $\varepsilon(s_i)$ has a Nash equilibrium which is within η of some equilibrium in S.

Minimality is necessary: by upper hemi-continuity, the set of all Nash equilibria would be strategically stable (no refinement).

Difference with trembling-hand perfection: convergence to an equilibrium in S for *any* sequence of perturbations.

Every equilibrium in a stable set has to be a perfect equilibrium... implied by minimality. Perfection in the normal form, not in the agent-normal form.

There exist stable sets that do not contain any sequential equilibrium.

Properties of Stable Equilibria

Theorem 6

There exists a stable set that is contained in a connected component of the set of Nash equilibria. Generically, each component of the set of Nash equilibria leads to a single distribution over outcomes, so there exists a stable set that induces a unique outcome distribution. A stable set contains a stable set of any game obtained by eliminating a weakly dominated strategy and also of any game obtained by deleting a strategy that is not a best response to any of the opponents' strategy profiles in the set (NWBR).

NWBR—robustness to forward induction: knowing that a player would not use a particular strategy is consistent with the equilibrium theories from the stable set

Forward Induction in Signaling Games

- NWBR useful to show that some equilibrium components are not stable
- Cho and Kreps (1987): equilibrium refinement for signaling games weaker than stability—the intuitive criterion—iterated applications of NWBR
- Kohlberg and Mertens (1986) motivate their stability concept by mathematical properties and robustness with respect to trembles a la Selten's perfect equilibrium.
- Cho and Kreps provide a behavioral foundation based on refining the set of plausible beliefs in the spirit of Kreps and Wilson's sequential equilibrium.

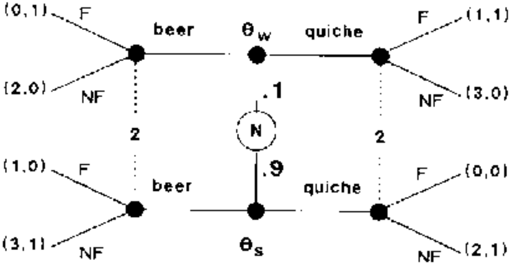
Signaling Games

- Two players: sender S and receiver R
- T: set of types for S
- $p(t)$: probability of type $t \in T$
- S privately observes his type t, then sends a message $m \in M(t)$
- $T(m) = \{t \mid m \in M(t)\}$: types that can send message m
- R observes m and chooses an action $a \in A(m)$
- Payoffs $u_S(t, m, a)$ and $u_R(t, m, a)$

Intuitive Criterion Idea

- Behavioral explanation of one aspect of NWBR: robustness to replacing the equilibrium path by its expected payoff.
- Presumes that players are certain about play on the equilibrium path, but there is uncertainty off the path.
- If we begin with a stable set and delete a strategy in which type t sends message m using NWBR, the reduced game should have a stable set contained in the original set.
- Surviving equilibria should assign probability 0 to type t following m.

The Beer-Quiche Game



Courtesy of The MIT Press. Used with permission.

- Player 1 is wimpy (w) or surly (s), with probabilities .1 and .9; $T = \{w, s\}$.
- 1 orders breakfast: $M = M(t) = \{beer, quiche\}, \forall t \in T$.
- Player 2 decides whether to fight: $A(m) = \{F, NF\}, \forall m \in M$.
- 1 gets utility 1 from having his favorite breakfast—beer if surly, quiche if wimp—but a disutility of 2 from fighting.
- When 1 is w, 2's payoff is 1 if he fights and 0 otherwise; when 1 is s, payoffs are reversed.

Sequential Equilibria

All sequential equilibria involve pooling

- Compare $\sigma_2(F|beer)$ and $\sigma_2(F|quiche)$
- Breakfast leading to a smaller probability of fighting must be selected with probability 1 in equilibrium by player 1 type who likes it...

Classes of sequential equilibria

1. Both types of player 1 drink beer.
2. Both types of player 1 eat quiche.

Player 2 does not fight in equilibrium. Player 2 must fight with probability at least $1/2$ when observing the out-of-equilibrium breakfast... supported by any belief for player 2 placing probability at least $1/2$ on w following the out-of-equilibrium breakfast.

Intuitive Criterion in Beer-Quiche

Quiche equilibrium unreasonable… NWBR violated

- Unreasonable for wimp to deviate to beer: no matter how 2 reacts, wimp cannot get more than 2, and he is already getting 3.
- Seeing beer, 2 should conclude that 1 is surly and not fight, which would induce surly type to deviate.

Forward-induction argument does not rule out the beer equilibrium

- In the beer equilibrium, it is unreasonable for surly type to deviate to quiche, while reasonable for wimp.
- 2's belief that 1 is wimpy if he orders quiche is reasonable.

Irrational Strategies for the Receiver

What if 2 can also pay a milion dollars to 1?

- It would be reasonable for both types to deviate.
- But 2 would never want to pay a million dollars.
- Assume 1 cannot expect 2 to play a irrational strategy.

Intuitive Criterion

For any $T' \subseteq T$ and any message m,

$$BR(T', m) = \cup_{\mu \mid \mu(T')=1} BR(\mu, m)$$

for strategies that R could rationally play after m and if he is certain that $t \in T'$.

Consider a sequential equilibrium

- $u_S^*(t)$: equilibrium payoff to type t
- $\tilde{T}(m) = \{t \mid u_S^*(t) > \max_{a \in BR(T(m),m)} u_S(t, m, a)\}$: types that do better in equilibrium than they could possibly do by sending m, no matter how R reacts, as long as R is rational.

The equilibrium fails the intuitive criterion if $\exists t' \in T, m \in M(t')$ s.t.

$$u_S^*(t') < \min_{a \in BR(T(m) \setminus \tilde{T}(m), m)} u_S(t', m, a).$$

Discussion

The equilibrium fails the intuitive criterion if some sender type is getting less than any payoff he could possibly get by playing m, assuming he could convince the sender that he is not in $\tilde{T}(m)$ because m does not make sense for any of those types.

In the beer-quiche example, the quiche equilibrium fails this criterion.

Iterated intuitive criterion

- Use the intuitive criterion as above to rule out pairs (t, m).
- Rule out actions of R, by requiring that R best respond to a belief about types that have not yet been eliminated given the message.
- Possibly rule out more pairs (t, m) given surviving strategies...

Banks and Sobel (1987)

- A type t' is *infinitely more likely* to choose the out-of-equilibrium message m than t if the set of possible best responses of R that make t' strictly prefer to deviate to m is a strict superset of the responses that make t weakly prefer to deviate.
- Conditional on observing m, R should put probability 0 on type t.
- D1: analogue of intuitive criterion under this elimination procedure
- D2: allow t' to vary across different best responses of S, requiring only that every best response that weakly induces t to deviate would strictly induce *some* t' to deviate
- *Universal divinity*: iteration of D2

Forward Induction and Iterated Weak Dominance

Iterated strict dominance and rationalizability narrow down the set of predictions without pinning down strategies perfectly.

Iterated weak dominance (IWD) captures some of the force of backward and forward induction without assuming that players coordinate on a certain equilibrium.

In games with perfect information, IWD implies backward induction: any suboptimal strategy at a penultimate node is weakly dominated. . .

Beer-quiche Game and IWD

Solve beer-quiche game applying IWD to ex-ante normal form.

1. $(\textit{beer if } w, \textit{quiche if } s)$ s. dominated by .1 *beer* $+$ .9 *quiche* for both w and s
   - For any strategy of 2, same total probability that 1 is fought by 2.
   - Player 1 has his favorite breakfast with probability 0 under first strategy and positive probability under second.
2. Fighting after beer weakly dominated by not fighting after beer
   - Only surviving strategies leading to beer: *beer* for both w and s and $(\textit{quiche if } w, \textit{beer if } s)$.
   - Best response to either strategy is not fighting (probability of $s \geq .9$).
3. Surly chooses beer in any surviving equilibrium, which generates his highest possible payoff of 3.

Stability and IWD

IWD captures part of the forward induction notion implicit in stability.

Stable components contain stable sets of games obtained by removing a weakly dominated action.

Kohlberg and Mertens' motivating example

	T	W
O	2,2	2,2
IT	0,0	3,1
IW	1,3	0,0

(IT, W): unique outcome under both IWD and stability

Ben-Porath and Dekel (1992)

Variation of the battle of the sexes: before game starts, player 1 has the option to "burn money."

If 1 decides not to burn money, play standard battle of the sexes

	L	R
U	5,1	0,0
D	0,0	1,5

If 1 decides to burn two units of utility, the game is

	L	R
U	3,1	-2,0
D	-2,0	-1,5

IWD: 1 can ensure his favorite equilibrium without burning

The mere option of burning money selects player 1's favorite equilibrium.

Burning Money and IWD

1. Burning followed by D s. dominated by not burning and D
2. Any strategy playing R after burning weakly dominated by L after burning
   - same outcome if 1 does not burn
   - after burning, L is better than R against the only surviving strategy, U
3. Not burning and D s. dominated by burning and U
   - burning and U yields a payoff of 3 for player 1 under surviving strategies (2 plays L after burning)
   - not burning and D gives player 1 at most 1
4. R after not burning weakly dominated by L after not burning
   - same outcome if 1 burns
   - after not burning, L is a best response to U

Only surviving outcome: no burning and (U, L).

Ben-Porath and Dekel: in any game where a player has a unique best outcome that is a strict Nash equilibrium and can signal with a sufficiently fine grid of burning stakes, she gets her best outcome under IWD.