Part I

Foundations

Chapter 2

Nash Equilibrium and Subgame Perfection

2.1 Nash Equilibrium and Nash's Theorem

Definition 2.1: Nash Equilibrium

A strategy profile $(\bar{\sigma}_1, \bar{\sigma}_2, \cdots, \bar{\sigma}_n)$ (where $\bar{\sigma}_i \in \Delta(S_i)$) is a Nash equilibrium of the game $G = (S_i, u_i)_{i=1}^n$ if for any player i,

$$u_i(\bar{\sigma}_i, \bar{\sigma}_{-i}) \geq u_i(s_i, \bar{\sigma}_{-i}), \quad \forall s_i \in S_i.$$

Remark.

Equivalently we can define Nash equilibrium by

$$u_i(\bar{\sigma}_i, \bar{\sigma}_{-i}) \geq u_i(\sigma_i, \bar{\sigma}_{-i}), \quad \forall \sigma_i \in \Delta(S_i).$$

But it suffices to only check pure strategies s_i because mixed strategies are just probability distributions over pure strategies, and the expected utility of a mixed strategy is a convex combination of the utilities of the pure strategies.

Theorem 2.2: Nash

Every finite game has a Nash equilibrium in mixed strategies.

Proof for Theorem

Here we apply the Kakutani's Fixed Point Theorem.

Theorem 2.3: Kakutani's Fixed Point Theorem

Suppose $F : X \rightrightarrows X$ (i.e., $F(x) \subseteq X$), and

- X is compact and convex;
- $F(x)$ is convex for all $x \in X$;
- F has a closed graph, i.e., the set $\{(x, y)|y \in F(x)\}$ is closed.

Then, there exists $x^* \in X$ such that $x^* \in F(x^*)$.

Let $X_i \equiv \Delta(S_i)$, and $X = X_1 \times X_2 \times \cdots \times X_n = \prod_{i=1}^n \Delta(S_i)$. Define the best response correspondence $B_i : \prod_{j \neq i} \Delta(S_j) \rightrightarrows \Delta(S_i)$ by

$$\sigma_i \in B_i(\sigma_{-i}) \iff u_i(\sigma_i, \sigma_{-i}) \geq u_i(\sigma_i', \sigma_{-i}), \quad \forall \sigma_i' \in \Delta(S_i).$$

And we define $B : X \rightrightarrows X$ by

$$B(\sigma) = \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_n \end{pmatrix} (\sigma).$$

Three properties of B need to be verified. First, $B(\sigma)$ is convex for every σ: if $\sigma_i^a, \sigma_i^b \in B_i(\sigma_{-i})$, both attain the maximum payoff, and any convex combination $\lambda\sigma_i^a + (1-\lambda)\sigma_i^b$ does too by linearity of u_i in player i's mixture. Second, X is compact and convex as the product of simplices. Third, the graph of B is closed: if $\sigma^k \to \bar{\sigma}$ and $\sigma^k \in B(\sigma^k)$, continuity of u_i implies $\bar{\sigma} \in B(\bar{\sigma})$. Kakutani's theorem then delivers a fixed point $\bar{\sigma} \in B(\bar{\sigma})$, which is a Nash equilibrium. ∎

Remark.

- "Mixed strategies" is important for Nash's Theorem. Finite games may not have NE in pure strategies.
- If $\bar{\sigma}$ is a NE, and $\bar{\sigma}_i(s_i) > 0$, then s_i must be a best response to $\bar{\sigma}_{-i}$, and thus $u_i(s_i, \bar{\sigma}_{-i}) = u_i(\bar{\sigma}_i, \bar{\sigma}_{-i})$. In other words, if a pure strategy is played with positive probability in a mixed strategy NE (i.e., in the *support* of the mixed strategy $\bar{\sigma}$), then it must yield the same expected payoff as the mixed strategy itself.
- Then why bother mixing? Because each player mixes to make other players mix. If a player chooses a pure strategy, then other players will have no incentive to mix, and thus the first player will have no incentive to mix either. Therefore, mixing can be a strategic move to induce other players to mix, which can lead to a more favorable outcome for the player.

Remark (Alternative Proof via Brouwer (Geanakoplos, 2003)).

Nash's theorem can also be proved using the simpler **Brouwer Fixed Point Theorem** (which only requires a continuous *function*, not a correspondence). The trick is to replace the best-response correspondence $B(\sigma)$, which is in general multi-valued, with a continuous *single-valued* map whose fixed points coincide with Nash equilibria.

For each player i, define $\phi_i : \Delta \to \Delta_i$ by

$$\phi_i(\sigma) \;=\; \arg\max_{\sigma_i' \in \Delta_i} \Big\{ u_i(\sigma_i', \sigma_{-i}) - \|\sigma_i' - \sigma_i\|_1^2 \Big\}.$$

The maximand is the sum of an affine (linear in σ_i') term and a strictly concave term, so it is strictly concave; hence $\phi_i(\sigma)$ is a singleton and, by the Theorem of the Maximum, ϕ_i is continuous in σ. The product map $\phi : \Delta \to \Delta$ is therefore continuous on the compact convex domain Δ, and Brouwer delivers a fixed point $\bar\sigma = \phi(\bar\sigma)$.

To see that $\bar\sigma$ is a Nash equilibrium, suppose for contradiction that some player i has a strictly profitable deviation σ_i' with $u_i(\sigma_i', \bar\sigma_{-i}) - u_i(\bar\sigma) = D > 0$. By linearity of u_i in σ_i, the convex combination $\varepsilon\sigma_i' + (1-\varepsilon)\bar\sigma_i$ raises i's payoff by exactly εD, while the penalty $\|\varepsilon\sigma_i' + (1-\varepsilon)\bar\sigma_i - \bar\sigma_i\|_1^2 = \varepsilon^2\|\sigma_i' - \bar\sigma_i\|_1^2$ is of order ε^2. For small ε, the linear gain dominates the quadratic loss, so the perturbed strategy strictly beats $\bar\sigma_i$ in the maximand defining ϕ_i—contradicting $\phi_i(\bar\sigma) = \bar\sigma_i$.

The Kakutani-based proof is the textbook standard, but the Brouwer route is conceptually cleaner: it makes explicit that the gap between "best-response correspondence has a fixed point" (which can be multi-valued and hence requires Kakutani) and "Nash equilibrium exists" (which only needs the existence of *some* consistent profile) can be bridged by a tiny regularization. The same trick reappears in computational game theory, where strictly concave perturbations are used to make best responses unique and approximate equilibria via gradient methods.

2.2 Iterated Dominance and Rationalizability

Chapter 1 introduced strictly dominated strategies and the iterated elimination of dominated strategies (IESDS), and informally identified the surviving set with *rationalizable* strategies. We now make that identification precise—the equivalence is a non-trivial theorem and its proof clarifies what each side of the definition is doing.

Definition 2.4: Iteratively Undominated Set (IUD)

A product set $X = X_1 \times \cdots \times X_n \subseteq S$ is **iteratively undominated** if there exists a sequence $\{X^t\}_{t=0}^T$ of product sets such that

(i) $X^0 = S$ and $X^T = X$;

(ii) $X^{t+1} \subseteq X^t$ for every t;

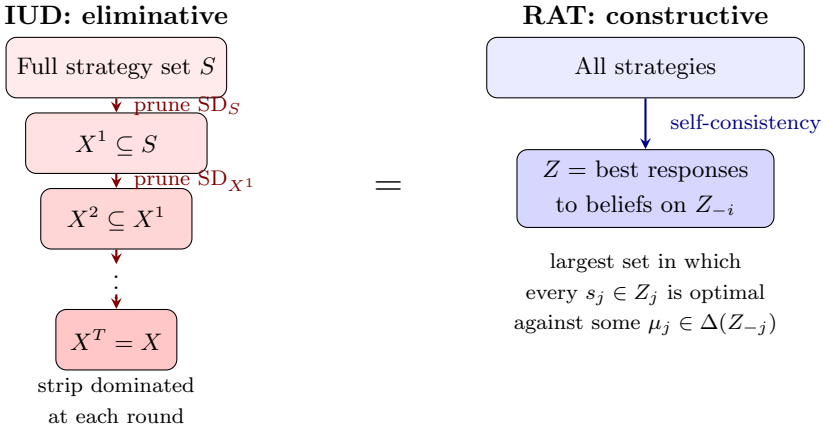
(iii) every $s_j \in X_j^t \setminus X_j^{t+1}$ is strictly dominated in the restricted game $(X_j^t, u_j)_{j=1}^n$;

(iv) no $s_j \in X_j^T$ is strictly dominated in the restricted game $(X_j^T, u_j)_{j=1}^n$.

Definition 2.5: Rationalizable Set (RAT)

A product set $Z = Z_1 \times \cdots \times Z_n \subseteq S$ is **rationalizable** if (i) every $s_j \in Z_j$ is a best response to some belief $\mu_j \in \Delta(Z_{-j})$ whose support lies in Z_{-j}, and (ii) Z is the maximal such set: any other set Z' satisfying (i) is contained in Z.

The two definitions describe rationality from opposite directions. IUD is *eliminative*: start with everything and prune dominated strategies, justifying each removal by reference to the previous round. RAT is *constructive*: insist that every surviving strategy be optimal under *some* belief over surviving opponent strategies, and take the largest such set. The next theorem shows they pick out the same object.



Theorem 2.6: IUD = RAT

Let X be the iteratively undominated set and Z the rationalizable set of a finite game G. Then $X = Z$.

Proof for Theorem

We use the lemma from Chapter 1 (Theorem 1.5) that $s_i \in S_i$ is strictly dominated (possibly by a mixed strategy) if and only if it is never a best response (NBR) to any belief in $\Delta(S_{-i})$. The proof proceeds by mutual inclusion.

$Z \subseteq X$. We show by induction that $Z_j \subseteq X_j^t$ for every t. The base case $t = 0$ is immediate since $X_j^0 = S_j$. For the inductive step, suppose $Z_j \subseteq X_j^t$. Take any $s_j \in Z_j$. By rationalizability, s_j is a best response to some belief μ_j supported on $Z_{-j} \subseteq X_{-j}^t$. By the SD $\iff$ NBR lemma applied to the restricted game (X^t, u), s_j is therefore not strictly dominated in (X^t, u). Hence $s_j \in X_j^{t+1}$, completing the induction. In the limit, $Z \subseteq X^T = X$.

$X \subseteq Z$. By the termination condition (iv), no $s_j \in X_j^T = X_j$ is strictly dominated in (X, u). By the SD $\iff$ NBR lemma, every $s_j \in X_j$ is a best response from X_j to some belief μ_j on X_{-j}. To upgrade "best response from X_j" to "best response from S_j" (the requirement in the definition of rationalizability), suppose for contradiction that some $s_j \in X_j$ is best-responded-to from X_j but *not* from S_j: there exists $s_j' \in S_j \setminus X_j$ that strictly improves on s_j against μ_j. Since s_j' is not in X_j, it was eliminated at some

stage t of the iterated procedure—meaning s'_j was strictly dominated in (X^t, u). But this contradicts that s'_j is a strict best response to $\mu_j \in \Delta(X^t_{-j})$ (by the SD $\iff$ NBR lemma in the restricted game). Hence X satisfies condition (i) of rationalizability. Since Z is by definition the maximal such set, $X \subseteq Z$. ■

Remark (Why the Equivalence Is Substantive).

IUD reaches the rationalizable set "from above," RAT defines it "intrinsically." Their equality is the formal statement that *rationality and common knowledge of rationality* imply nothing more than what iterated elimination of strictly dominated strategies extracts. In particular: there is no mileage to be gained by allowing players to entertain richer hierarchies of beliefs beyond what IESDS already captures. Conversely, IESDS is not a brute-force algorithm divorced from epistemic foundations—each round of elimination corresponds exactly to one more level of mutual knowledge of rationality. The order-of-elimination invariance for strict dominance, observed informally in Chapter 1, is now a corollary: every elimination order converges to the unique RAT set.

2.3 Potential Games

One of the central questions in game theory is whether a game has a *pure strategy Nash equilibrium* (PSNE). Some games, like matching penny, have no PSNE. However, certain classes of games are guaranteed to have PSNE. Two important classes are:

- Potential games
- Supermodular games

In this section we will focus on potential games, which are a broad class of games that include many interesting examples, such as congestion games, coordination games, and certain types of auctions. Potential games have a special structure that allows us to analyze them using a potential function, which captures the incentives of all players in a single function.

2.3.1 Definition

Definition 2.7: Potential Game

A finite game $G = (S_i, u_i)_{i=1}^n$ is called a *potential game* if there exists a function $P : S \to \mathbb{R}$ (where $S = S_1 \times S_2 \times \cdots \times S_n$) such that for every player i and every pair of strategies $s_i, s'_i \in S_i$:

$$u_i(s_i, s_{-i}) - u_i(s'_i, s_{-i}) = P(s_i, s_{-i}) - P(s'_i, s_{-i}), \quad s_{-i} \in S_{-i}.$$

The function P is called the *potential function*.

Example (Congestion Game).

Consider a road network where n drivers choose routes. Let s_i be the route chosen by driver i, $n_r(s)$ the number of drivers on route r under profile s, and $c(r,k)$ the congestion cost incurred by each user of route r when k drivers use it. Driver i's payoff is

$$u_i(s_i, s_{-i}) = -c\big(s_i,\ n_{s_i}(s)\big).$$

Define the potential function

$$P(s) \ = \ -\sum_{r} \sum_{k=1}^{n_r(s)} c(r,k).$$

When driver i switches from route r to route r', the change in P is exactly $c(r', n_{r'}(s)+1) - c(r, n_r(s))$, which equals the change in driver i's own payoff (up to sign). Hence P is a potential function and the game admits a PSNE—the route assignment that minimizes total congestion cost.

Remark.

- Potential games can be thought of as games where all players have identical payoffs.
- Potential games arise naturally when:
  - Players care about a common objective (like aggregate welfare)
  - Individual incentives are aligned with some global measure
  - Strategic interactions are "symmetric" in the sense of cross-partial derivatives
- The potential function P captures the payoff differences for individual players when they deviate from one strategy to another. The key insight is that a player's incentive to deviate depends only on how the potential function changes, not on the baseline payoff levels.
- In the continuous case with smooth payoff functions, a game is a potential game if and only if:

  $$\frac{\partial P}{\partial s_i} = \frac{\partial u_i}{\partial s_i} \quad \text{for all players } i$$

  This leads to the following necessary condition, for all pairs of players i and j:

  $$\frac{\partial^2 u_i}{\partial s_j \partial s_k} = \frac{\partial^2 u_j}{\partial s_k \partial s_j} \quad \text{whenever these derivatives exist.}$$

2.3.2 Existence of PSNE in Potential Games

Theorem 2.8: Existence of PSNE in Potential Games

Every potential game has at least one pure strategy Nash equilibrium.

Proof for Theorem

A PSNE of game G is a maximizer of the potential function P. Since S is finite (in the discrete case), P attains its maximum. At such a maximum point s^*, if player i deviates to any other strategy $s_i \neq s_i^*$, then:

$$u_i(s_i^*, s_{-i}^*) - u_i(s_i, s_{-i}^*) = P(s_i^*, s_{-i}^*) - P(s_i, s_{-i}^*) \geq 0$$

Thus no player wants to deviate, so s^* is a PSNE. ∎

Remark.

Different information structures (simultaneous move vs. sequential move) can lead to different equilibria in general games. However, in potential games, the set of pure strategy Nash equilibria is often robust across different information structures. This is because the potential function captures the incentives of all players in a way that is independent of the timing of moves. As a result, the same strategy profiles that maximize the potential function will be equilibria regardless of whether players move simultaneously or sequentially.

2.4 Supermodular Games

While potential games ensure the existence of PSNE through a global potential function, supermodular games approach the problem differently: they rely on the structure of strategic complementarities between players' actions.

The key insight is *strategic complementarity*: if one player's action increases, it becomes more attractive for other players to increase their actions as well. This creates a natural coordination mechanism that leads to equilibrium. Unlike potential games, which require a global objective function, supermodular games only require local complementarity conditions on payoffs.

Examples of strategic complementarity abound:

- In a price competition setting, if a competitor raises prices, your prices become more attractive to consumers, so you want to raise your price too.
- In technology adoption, if more people adopt a technology, the benefits of adoption increase for you (network effects).
- In public goods games, if others contribute more, your optimal contribution may also increase.

2.4.1 Definitions

We need lattice-theoretic concepts to formalize strategic complementarity.

Definition 2.9: Meet & Join

Let x and y be two vectors of $\mathbb{R}^n$. The *meet* of x and y, denoted by $x \wedge y$, is the coordinate-wise minimum:

$$x \wedge y = \begin{pmatrix} \min(x_1, y_1) \\ \min(x_2, y_2) \\ \vdots \\ \min(x_n, y_n) \end{pmatrix}$$

The *join* of x and y, denoted by $x \vee y$, is the coordinate-wise maximum:

$$x \vee y = \begin{pmatrix} \max(x_1, y_1) \\ \max(x_2, y_2) \\ \vdots \\ \max(x_n, y_n) \end{pmatrix}$$

Remark.

Meet and join decompose a pair of strategy profiles into their "lower bound" and "upper bound" in each dimension. Later we will see, if a payoff function is supermodular, then being at both extremes (the largest and smallest actions) together is at least as good as being at the two middle ground positions.

Definition 2.10: Supermodular Function

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *supermodular* if for all $x, y \in \mathbb{R}^n$:

$$f(x \wedge y) + f(x \vee y) \geq f(x) + f(y)$$

This is called the *supermodularity inequality*.

Intuitively, the supermodularity inequality says: having some coordinates large and others small is better than having all coordinates at intermediate levels.

In a game setting, we say player i's payoff is supermodular if:

$$u_i(s) + u_i(s') \leq u_i(s \wedge s') + u_i(s \vee s') \quad \forall s, s' \in S$$

This means: player i's payoff is supermodular in their own strategy relative to opponents' strategies.

Remark.

When f is twice continuously differentiable, supermodularity is equivalent to:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \geq 0 \quad \text{for all } i \neq j.$$

This means: if you increase x_i, the marginal benefit of increasing x_j also increases (*strategic complementarity*).

2.4.2 Existence of PSNE in Supermodular Games

The next two lemmas establish that best response correspondences have monotonicity properties under supermodularity, which enables us to use fixed point theorems.

Lemma 2.11: Monotonicity of Best Responses

Suppose $u : [0,1]^p \times [0,1]^q \to \mathbb{R}$ is supermodular. Suppose $y', y'' \in [0,1]^q$ with $y'' \geq y'$. Let $x' \in [0,1]^p$ be a maximizer of $u(\cdot, y')$, and $x'' \in [0,1]^p$ be a maximizer of $u(\cdot, y'')$. Then $x' \wedge x''$ is a maximizer of $u(\cdot, y')$, and $x' \vee x''$ is a maximizer of $u(\cdot, y'')$.

Proof for Lemma

By construction, since x' maximizes $u(\cdot, y')$ and x'' maximizes $u(\cdot, y'')$:

$$u(x', y') \geq u(x' \wedge x'', y')$$

Also note that $u(x' \wedge x'', y') = u(x' \wedge x'', y' \wedge y'')$ since $y' \wedge y'' = y'$ (as $y' \leq y''$). Similarly:

$$u(x'', y'') \geq u(x' \vee x'', y'') = u(x' \vee x'', y' \vee y'')$$

By supermodularity:

$$u(x', y') + u(x'', y'') \geq u(x' \wedge x'', y' \wedge y'') + u(x' \vee x'', y' \vee y'')$$

Suppose $u(x', y') > u(x' \wedge x'', y' \wedge y'')$. Then:

$$u(x', y') + u(x'', y'') > u(x' \wedge x'', y' \wedge y'') + u(x' \vee x'', y' \vee y'')$$

which violates supermodularity.

Therefore, we must have:

$$u(x', y') = u(x' \wedge x'', y' \wedge y'')$$
$$u(x'', y'') = u(x' \vee x'', y' \vee y'')$$

This means both $x' \wedge x''$ and $x' \vee x''$ are maximizers at the respective points. ∎

Lemma 2.12: Monotonicity of Largest Maximizers

Suppose $u : [0,1]^p \times [0,1]^q \to \mathbb{R}$ is supermodular. Define:

$$h(y) = \max\{x : x \text{ maximizes } u(\cdot, y)\}$$

(the largest maximizer). Then h is non-decreasing: if $y'' \geq y'$, then $h(y'') \geq h(y')$.

Proof for Lemma

Suppose $y'' \geq y'$. Let $x' = h(y')$ and $x'' = h(y'')$ be the largest maximizers at y' and y'' respectively. By the previous lemma, $x' \wedge x''$ maximizes $u(\cdot, y')$ and $x' \vee x''$ maximizes $u(\cdot, y'')$.

Since $x' \vee x''$ is a maximizer at y'' and x'' is the largest maximizer at y'', we have $x'' \geq x' \vee x''$. By definition, $x' \vee x'' \geq x'$. So $x'' \geq x'$. Therefore, $h(y'') = x'' \geq x' = h(y')$.

Theorem 2.13: Existence of PSNE in Supermodular Games

Every finite supermodular game has at least one pure strategy Nash equilibrium.

Proof for Theorem

This is a direct result of Tarski's Fixed Point Theorem.

Theorem 2.14: Tarski's Fixed Point Theorem

If $f : L \to L$ is a non-decreasing function on a complete lattice L, then f has a fixed point.

Why Tarski holds (sketch). *Let L be a complete lattice with top $\top$ and bottom $\bot$. The set $A = \{x \in L : x \leq f(x)\}$ contains $\bot$ (since $\bot \leq f(\bot)$) and is therefore non-empty. Let $\bar{x} = \sup A$, which exists by completeness. Monotonicity gives $f(x) \leq f(\bar{x})$ for every $x \in A$, and combining with $x \leq f(x)$ yields $x \leq f(\bar{x})$, so $f(\bar{x})$ is an upper bound of A and hence $\bar{x} \leq f(\bar{x})$. Applying f once more and using monotonicity, $f(\bar{x}) \leq f(f(\bar{x}))$, so $f(\bar{x}) \in A$ and $f(\bar{x}) \leq \bar{x}$. Together: $f(\bar{x}) = \bar{x}$. The argument is constructive on finite lattices: iterate f on $\bot$ until the sequence stabilizes.*

For each player i, define $b_i(s_{-i})$ as the largest maximizer of $u_i(\cdot, s_{-i})$. By the lemmas above, b_i is non-decreasing in s_{-i}.

Define the best response mapping $b : S \to S$ by:

$$b(s) = (b_1(s_{-1}), b_2(s_{-2}), \ldots, b_n(s_{-n}))$$

Since b is component-wise non-decreasing and S is a finite lattice, by *Tarski's Fixed Point Theorem*, b has a fixed point $\bar{s}$, i.e., $b(\bar{s}) = \bar{s}$. This fixed point is a Nash equilibrium.

Example (Bertrand Duopoly with Differentiated Products).

Consider two firms competing in prices. Let $D_i(p_1, p_2)$ be firm i's demand, which is decreasing in its own price p_i and increasing in the competitor's price p_j. With constant marginal cost c, firm i's profit is:

$$\pi_i(p_1, p_2) = (p_i - c)D_i(p_1, p_2)$$

The key observation is:

$$\frac{\partial^2 \pi_i}{\partial p_i \partial p_j} = \frac{\partial}{\partial p_j}\left[\frac{\partial \pi_i}{\partial p_i}\right] > 0$$

This is because when the competitor's price increases, the marginal profit from raising your own price increases (since demand becomes more favorable). Thus, firm i's profit is supermodular in (p_1, p_2), and the duopoly game has a PSNE.

2.5 Subgame Perfect Nash Equilibrium

Nash equilibrium constrains only on-path play. In extensive-form games—where players move sequentially and can condition on what they have observed—this is too weak: many Nash equilibria are sustained by threats that the threatener would never actually carry out. Subgame perfection refines NE by demanding optimality in every *subgame*, ruling out such non-credible threats.

2.5.1 Definitions and Existence

Definition 2.15: Subgame

Given an extensive form T, a subgame T' is some node $r \in T$ and all its succeeding nodes such that if an information set I intersects T', then all nodes in I are also in T'.

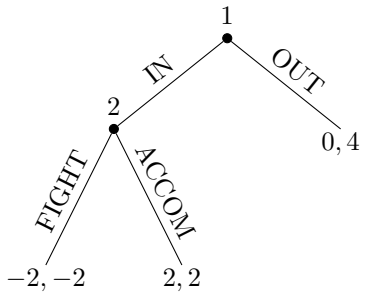
Definition 2.16: Subgame Perfect Nash Equilibrium

σ is a *subgame perfect equilibrium* if

- σ is a Nash equilibrium.
- σ induces a Nash equilibrium in every subgame $T' \subseteq T$.

Example (Entry Deterrence Game).

Consider the following game:



From its normal form (payoff matrix) below, we know that there are two Nash equilibria: (OUT, FIGHT) and (IN, ACCOM).

	FIGHT	ACCOM
OUT	$0, 5$	$0, 5$
IN	$-2, -2$	$2, 2$

However, only (IN, ACCOM) is a subgame perfect equilibrium, because if Player 1

chooses IN, then Player 2's best response is to choose ACCOM (so to choose FIGHT is not a credible threat). Therefore, the unique subgame perfect equilibrium is (IN, ACCOM).

Suppose now the game is repeatedly played for $T = 20$ times. At $t = 1$, Player 1 chooses IN or OUT. If Player 1 chooses OUT, the game ends. If Player 1 chooses IN, then in any period $t \geq 1$, the Player 1 can choose to stay (IN) or exit (OUT). If OUT is ever chosen by Player 1, then OUT for all the periods onwards.

Clearly, using backward induction, we know that the unique subgame perfect equilibrium is for Player 1 to choose IN and Player 2 to choose ACCOM in all periods.

However, if we add one additional constraint: Player 1 can only fight for at most 15 periods, then the analysis changes.

Claim

The unique SPE with the constraint is that

- Player 1 never enters (IN), and
- Player 2 chooses FIGHT if Player 1 chooses IN.

Proof for Claim.

Suppose that Player 1 has entered and been fought for the previous 14 periods. In period $t = 15$, if Player 1 chooses IN, Player 2 compares the payoffs of FIGHT and ACCOM:

- If Player 2 chooses FIGHT: Player 1's capacity is exhausted, forcing them to choose OUT from $t = 16$ to $t = 20$. Player 2's total remaining payoff is $-2(\text{at } t = 15) + 5 \times 4(\text{monopoly profit}) = 18$.
- If Player 2 chooses ACCOM: Player 1 stays in the market. Player 2's total remaining payoff is $2(\text{at } t = 15) + 5 \times 2(\text{duopoly profit}) = 12$.

Since $18 > 12$, FIGHT becomes a *credible threat* at $t = 15$. Anticipating this, Player 1 will choose OUT at $t = 15$ to avoid the -2 payoff.

Consider period $t = 14$. Player 1 anticipates that choosing IN will result in Player 2 choosing FIGHT (to trigger the exit at $t = 15$ and secure future monopoly profits). Choosing IN at $t = 14$ followed by exit at $t = 15$ yields a lower payoff than choosing OUT immediately at $t = 14$.

Following this logic backward to $t = 1$, consequently, the only rational choice for Player 1 is to never enter (OUT), and for Player 2 to threaten to FIGHT at any entry node.

Theorem 2.17

Every finite game in extensive form has at least one subgame perfect equilibrium.

Proof for Theorem

The proof is to find all subgames that do not have any proper subgames. In each such subgame, find a Nash equilibrium of the subgame, and replace each node by expected payoffs from any Nash equilibrium of the subgame. ■

Remark.

The notion of subgame perfect equilibrium is to generalize backward induction to games with imperfect information.

2.5.2 The Backward Induction Algorithm

The proof of the existence theorem is constructive: it gives an explicit *algorithm* for computing an SPE in any finite extensive-form game. The procedure—**backward induction**—is the workhorse of finite-horizon analysis.

Backward Induction Algorithm

Let T be a finite extensive-form game.

1. **Identify minimal subgames.** A subgame is *minimal* if it contains no proper subgame. Equivalently, every information set inside it is a singleton in which only one player moves before reaching a terminal node.
2. **Solve each minimal subgame.** Find a Nash equilibrium of the minimal subgame (in finite games this exists, by Nash's theorem applied to the subgame's normal form).
3. **Substitute continuation payoffs.** Replace every minimal subgame by a single terminal node carrying the equilibrium payoff vector. This yields a smaller extensive-form game T' with one less "layer."
4. **Recurse.** Apply Steps 1–3 to T'. Terminate when only the root remains; the strategies recorded along the way constitute an SPE.

Remark (Why Backward Induction Produces an SPE).

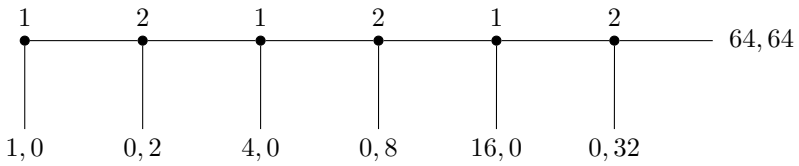
By construction, the strategies prescribed in each minimal subgame are a Nash equilibrium of that subgame. After substitution, the next iteration treats those subgames as terminal nodes with fixed payoffs, so any Nash equilibrium of the reduced game T' remains a Nash equilibrium when expanded back. Iterating, the prescribed strategies form a Nash equilibrium in every subgame—which is exactly the SPE condition. The argument is the game-theoretic analogue of **dynamic programming**: a globally optimal plan is built by stitching together locally optimal continuations.

Remark (Verification via the One-Deviation Property).

A practical corollary: to check that a candidate strategy profile σ^* is an SPE, it suffices to verify, at every history (or every information set in the perfect-information case), that no player can profitably deviate by changing her action at that single node and reverting to σ^* thereafter. This is the **one-shot deviation principle**, a finite-horizon analogue of the result we will state and prove for infinitely repeated games in the chapter on dynamic games. The principle reduces SPE verification from a global check across all alternative strategies to a local, period-by-period inequality at each node—precisely what the entry-deterrence example above (with Player 1's 15-period capacity constraint) implicitly relied on.

The backward-induction algorithm settles existence and provides a constructive solution method, but it does *not* guarantee uniqueness: if a minimal subgame has multiple Nash equilibria, choosing different ones leads to different SPEs. This multiplicity is what powers the finite-horizon Folk Theorem (in the chapter on dynamic games): when a stage game has several payoff-distinct NE, the choice of which NE to play in late periods can be used as a reward or punishment to sustain non-NE play in early periods.

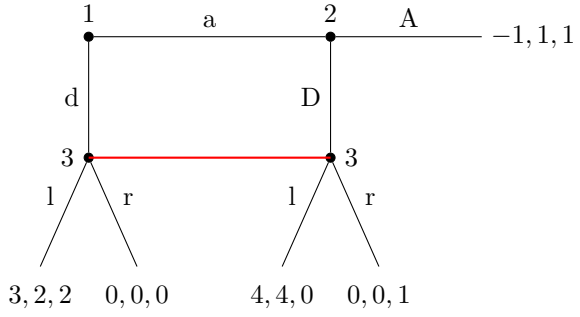
Example (Centipede Game).



The unique subgame perfect equilibrium is for each player to exit immediately at their first decision node, yielding payoff $(1, 0)$. However, the backward induction outcome $(1, 0)$ is Pareto dominated by the outcome $(64, 64)$ at the end. Both players would be better off if the game continued to the end, but the equilibrium logic prevents this.

2.5.3 Trembling-Hand Perfect Equilibrium

Example (Selten's Horse).



There are two pure strategy Nash equilibria: (d, A, l) and (a, A, r). Since this game has no proper subgames, both equilibria are subgame perfect.

However, the SPE (d, A, l) appears problematic: if Player 1 accidentally chooses a instead of d, Player 2 would prefer to choose D rather than A. This suggests the equilibrium is not robust to small mistakes.

Recall that σ is a Nash equilibrium if and only if for each player i and any pure strategy s_i': $u_i(\sigma_i, \sigma_{-i}) \geq u_i(s_i', \sigma_{-i})$, and for all the pure strategies that are in the support of σ_i, the expected payoff is the same: $u_i(s_i, \sigma_{-i}) = u_i(\sigma_i, \sigma_{-i})$ for all s_i such that $\sigma_i(s_i) > 0$. Hence, if for some pure strategy s_i', we have $u_i(\sigma_i, \sigma_{-i}) > u_i(s_i', \sigma_{-i})$, then $\sigma_i(s_i') = 0$.

To formalize the intuition from the previous example of Selten's Horse, suppose all players' choices are subject to small trembles (mistakes).

Definition 2.18: ε-Equilibrium

A mixed strategy profile σ^ε is an *ε-equilibrium* (for small $\varepsilon > 0$) if:

- Every pure strategy has positive probability: $\sigma_i^\varepsilon(s_i) > 0$ for all i and s_i.
- Strategies that are not a best response are played with low probability: if $u_i(s_i', \sigma_{-i}^\varepsilon) > u_i(s_i, \sigma_{-i}^\varepsilon)$, then $\sigma_i^\varepsilon(s_i') < \varepsilon$.

Definition 2.19: Trembling Hand Perfect Equilibrium

A strategy profile σ is a *trembling hand perfect equilibrium* (THPE) if it is the limit of a sequence of ε-equilibria σ^ε as $\varepsilon \to 0$.

Remark (Chapter Summary).

This chapter built the central solution concept of the book. Nash's theorem (Theorem 2.1) guarantees that every finite game has at least one mixed-strategy equilibrium; the proof, via Kakutani or—more directly—Brouwer, identifies an equilibrium with a fixed point of a best-response correspondence. We then mapped out two families of refinements and existence results. *Iterated dominance and rationalizability* (IUD = RAT) describe what rationality plus common knowledge of rationality alone can deliver, before equilibrium fixed-point reasoning enters. *Potential and supermodular games* carve out structural classes in which pure-strategy equilibria are guaranteed and have nice comparative-statics properties (Topkis-Milgrom-Roberts). For dynamic games, *subgame perfection* (Definition 2.5.1) sharpens Nash equilibrium by requiring optimality at every history; *trembling-hand perfection* sharpens it again by ruling out equilibria sustained by zero-probability mistakes. Each refinement narrows the prediction at the cost of stronger assumptions about behavior off the equilibrium path—a tradeoff that recurs in every later chapter.

Part II

Bargaining

Part III

Auctions and Mechanism Design

Part IV

Matching

Part V

Information and Dynamic Games

Part VI

Problem Sets and Solutions

Part VII

Exams and Solutions