The Folk Theorem

Game Theory · Module 5 — Repeated Games

Node 5.4

Position in Knowledge Graph

System: Game Theory **Structural Domain:** Repeated Games **Node:** 5.4

The folk theorem is the central result of repeated game theory and one of the strangest results in all of game theory. It says, roughly, that for sufficiently patient players, any feasible and individually rational payoff vector of the stage game can be sustained as the average payoff of a subgame-perfect equilibrium of the infinitely repeated game. The name "folk theorem" is literal: the result was in the air, was known to many people, and was not attributed to any single author because it seemed almost obvious once one accepted the machinery of 5.2 and 5.3. The strange thing about the theorem is that it is simultaneously a deep existence result — showing that the SPE payoff set has full dimension, not just a few isolated points — and a devastating indictment of equilibrium as a predictive tool: if every individually rational payoff is sustainable, equilibrium is not selecting, it is permitting. The folk theorem is the source of the "multiple equilibria problem" in repeated games and motivates the refinement program, the learning-dynamics program, and the evolutionary-game-theory program of Modules 7 and 9. It is also, in its constructive form (Fudenberg–Maskin and successors), the basis on which mechanism design (Module 6) can operate: if you want to implement a specific outcome, you need to know what the feasible set of repeated-game equilibria is, and the folk theorem tells you.

Formal Definition

Fix a finite stage game $G = (N, \{S_i\}, \{u_i\})$. Consider the infinitely repeated game G^∞ with discount factor δ.

Feasible payoffs. The set $F \subseteq \mathbb{R}^n$ of feasible payoffs is the convex hull of stage-game payoff vectors:

$$F = \operatorname{conv}\{u(s) : s \in S\}.$$

A payoff $v \in F$ is feasible iff it is a convex combination of stage-game payoff vectors — equivalently, it is the average payoff from some (possibly randomized) sequence of stage profiles.

Individually rational payoffs. Player i's **minmax value** is

$$v_i^{\min} = \min_{s_{-i} \in \prod_{j \neq i} \Delta(S_j)} \max_{s_i \in S_i} u_i(s_i, s_{-i}),$$

the best payoff i can guarantee against the worst (adversarial) coordination of all other players. A payoff $v = (v_1, \ldots, v_n)$ is **individually rational** if $v_i \geq v_i^{\min}$ for every i and **strictly individually rational** if $v_i > v_i^{\min}$ for every i.

The set of feasible and strictly individually rational payoffs is denoted $F^* = \{v \in F : v_i > v_i^{\min} \text{ for all } i\}$.

Folk theorem (Nash version). For every $v \in F^*$, there exists $\delta^* < 1$ such that for all $\delta \in (\delta^*, 1)$, there is a Nash equilibrium of G^∞ with average discounted payoff vector v.

Folk theorem (Fudenberg–Maskin 1986, SPE version). If the set F^* has nonempty interior (i.e., the game has "full dimension"), then for every $v \in F^*$, there exists $\delta^* < 1$ such that for all $\delta \in (\delta^*, 1)$, there is a

subgame-perfect equilibrium of G^∞ with average discounted payoff vector v. The dimensionality condition is required to construct Abreu-style punishments with distinct rewards for distinct players.

Corollary. As $\delta \to 1$, the set of SPE average discounted payoff vectors converges (in the Hausdorff metric) to the closure of F^*.

Intuition

The folk theorem's punchline is: "anything goes." For patient enough players, any cooperative division of the feasible payoff pie can be sustained by the threat of punishment if anyone deviates. There is no equilibrium selection — every individually rational feasible payoff is an equilibrium payoff.

This is both amazing and terrible. Amazing: the theorem shows that rationality alone does not constrain repeated-game behavior beyond the minmax bounds. Cooperation, competition, alternation, exploitation — all of it is rationalizable. Terrible: as a predictive theory of strategic behavior, repeated-game equilibrium is essentially vacuous. If someone asks "what will players do in this repeated Prisoner's Dilemma?" the honest answer is "anything in the feasible individually rational set, depending on which equilibrium they coordinate on." This is not a useful prediction.

The terrible side motivates the response of the field over the past 40 years: equilibrium refinements (which shrink the SPE set by adding consistency constraints), learning dynamics (which select equilibria by studying which ones are limits of learning processes), evolutionary dynamics (which select by which are stable under invasion), and empirical/behavioral game theory (which selects by studying what humans actually do). The folk theorem is the problem; the subsequent literature is the attempted solution.

There is another reading of the folk theorem that is more positive. The theorem shows that **repeated-game institutional design is unconstrained by theoretical impossibility**. If you want to implement a specific outcome — say, a non-equilibrium cooperative profile that Pareto-dominates the stage NE — the folk theorem says this is possible in principle, given enough patience. The question becomes practical: how to reach the desired equilibrium rather than one of the many others. The folk theorem guarantees feasibility; everything after is selection and implementation. This reading is the foundation of mechanism design (Module 6) as applied to repeated interactions: the designer chooses a repeated-game equilibrium from the feasible set and designs institutions to select it.

The constructive proofs of the folk theorem, due to Fudenberg–Maskin (1986) for the SPE version, are also instructive. They show that any target payoff can be reached by a "public randomization" over stage profiles combined with a stick-and-carrot punishment for deviators. The mechanism is: all players agree on a sequence of stage profiles (possibly correlated or randomized) that averages to the target payoff; any deviation triggers an Abreu-style punishment phase that minimizes the deviator's continuation payoff; the phase eventually returns to the cooperation sequence. For δ close enough to 1, this satisfies SPE.

Structural Role

5.4 is the capstone theoretical result of Module 5. Its role is:

- **Settling the question of what is possible** in repeated games: everything feasible and individually rational.
- **Motivating refinements** by making it clear that SPE alone is not enough.
- **Providing the theoretical basis** for mechanism design in repeated environments (Module 6).
- **Connecting to learning dynamics** (Module 9) by defining the target of analysis: which equilibrium, among the folk-theorem set, will players actually learn or converge to?

The folk theorem also has deep connections outside Module 5:

- **Module 7 (evolutionary games)** studies which repeated-game equilibria are evolutionarily stable, cutting the folk-theorem set down to a smaller selected set.
- **Module 8 (network games)** studies how local interaction structure affects the sustainability of folk-theorem-style cooperation.
- **Module 9 (learning)** asks which folk-theorem equilibria are limits of decentralized learning dynamics; typically only certain correlated or coarse equilibria are robust.
- **Module 10 (strategic systems integration)** treats the folk theorem as the boundary condition: any SPE in the folk-theorem set is a candidate for an institutional or political equilibrium.

Mathematical Development

The Nash-version folk theorem: sketch of proof

Claim. For $v \in F^*$, there is a Nash equilibrium of G^∞ with average discounted payoff v for δ sufficiently close to 1.

Construction. Since v is feasible, there is a sequence of stage profiles $(s^1, s^2, \ldots)$ — possibly with public randomization — whose average payoff (as $T \to \infty$) is v. Design the strategy:

- On path: play the sequence $(s^1, s^2, \ldots)$.
- Off path: if any player deviates at any period, all *other* players immediately switch to the minmax strategy against the deviator, playing forever.

Incentive compatibility (Nash). A deviation by i gives at most a one-shot gain bounded by $2M$, where M is the maximum stage payoff. The cost is the switch from v_i to $v_i^{\min}$ forever after. The total cost is $\delta/(1-\delta) \cdot (v_i - v_i^{\min})$. For v strictly individually rational ($v_i > v_i^{\min}$), the cost exceeds the gain for δ sufficiently close to 1. Nash equilibrium holds.

What this proof does not give. It does not give SPE. The threat "switch to minmax forever" may not be credible: the minmaxing players are giving up their own payoffs to punish i, and they may want to deviate from the punishment themselves. The Nash folk theorem is not SPE folk theorem.

The SPE-version folk theorem (Fudenberg–Maskin 1986)

The constructive proof uses Abreu's stick-and-carrot:

- **Carrot phase:** the target cooperative sequence, yielding average payoff v.
- **Stick phase for deviator i:** a profile that gives i their minmax value for K periods, where K is chosen large enough to deter deviation.
- **Return to carrot:** after the stick, return to the cooperative sequence. However, the return is not to the same cooperative payoff — it is to a slightly higher payoff for the punishers to compensate them for the cost of punishment. This is the "full dimension" requirement.

Full dimension. The requirement is that the set F^* has non-empty interior in $\mathbb{R}^n$, equivalently the feasible and strictly individually rational payoff set is "fat." This ensures that for any target v and any player i, there are nearby alternative payoff vectors that give strictly more to everyone except i — the necessary "reward" for the punishers.

Credibility. The punishment profile gives each non-i player a payoff that, when combined with the future reward, exceeds what they would get by deviating from the punishment. The reward covers the short-term cost of minmaxing i, making the punishment phase itself a Nash of its subgame.

Theorem statement (Fudenberg–Maskin 1986). If G has the full-dimension property, then for every $v \in F^*$, there exists $\delta^* < 1$ such that for all $\delta > \delta^*$, there is an SPE of G^∞ with average discounted payoff v.

Without full dimension. For two-player games, the full-dimension requirement can sometimes be dropped, but it genuinely matters in three or more players. There are games with two players where the folk theorem still holds via other mechanisms (mutual minmax), and games with three players where the folk theorem requires the full-dimension condition and fails without it.

Minmax values and the individually rational set

Computing the minmax value $v_i^{\min}$ requires solving a min-max problem over the opponent's mixed strategies:

$$v_i^{\min} = \min_{\sigma_{-i}} \max_{s_i} u_i(s_i, \sigma_{-i}).$$

For two-player zero-sum games, the minmax value coincides with i's Nash equilibrium payoff (von Neumann's minmax theorem). For general-sum games, the minmax value is typically lower than i's Nash payoff — the opponents can, by coordinating adversarially, push i below their Nash payoff.

Example (Prisoner's Dilemma). Opponent plays D; your best response is D; minmax value is 1. This equals the stage Nash payoff here because the Nash strategy D for the opponent is exactly the minmaxing strategy against you.

Example (Battle of the Sexes). Stage game payoffs $(2, 1), (0, 0), (0, 0), (1, 2)$ on the off-diagonal and diagonal respectively. Player 1's minmax value: what is the lowest payoff the opponent can force on player 1? If opponent mixes 50-50, player 1's best response is indifferent, and expected payoff is $0.5 \cdot 2 + 0.5 \cdot 1 = 1.5$ or $0.5 \cdot 0 + 0.5 \cdot 0 = 0$ depending on player 1's strategy. Player 1's best response gives 1.5. The minmaxing mix is 50-50 and player 1's minmax value is 2/3 (the specific mix can be computed by solving the minmax problem exactly). Here minmax is below Nash, because Nash is a coordination, not an adversarial pin-down.

Extensions and nearby results

- **Imperfect monitoring (Fudenberg–Levine–Maskin 1994).** When players do not observe each other's actions directly but only noisy signals, the folk theorem still holds under identifiability conditions on the signal structure. This is the foundation of the repeated-game theory of cartels and collusion under imperfect monitoring.
- **Nash-reversion folk theorem (Friedman 1971).** An earlier, weaker version of the folk theorem using grim trigger to stage NE as the punishment. Works without the full-dimension condition but yields a smaller equilibrium set.
- **Finite-horizon folk theorem (Benoit–Krishna 1985).** For finite repeated games with multiple stage equilibria, as $T \to \infty$, the set of SPE payoffs converges to the folk-theorem set. Even without infinite horizon, "long enough" horizon plus multiplicity suffices.
- **Continuous-time repeated games.** Sannikov (2007) and others developed a continuous-time version of the folk theorem with imperfect monitoring, with surprising differences from the discrete-time case.

Worked Example

Prisoner's Dilemma folk theorem

Stage game: $(3, 3), (0, 5), (5, 0), (1, 1)$ as before.

Feasible payoffs. Convex hull of $(3, 3), (0, 5), (5, 0), (1, 1)$. This is a quadrilateral in $\mathbb{R}^2$ with vertices at these four points.

Minmax values. For each player, opponent plays D; best response is D; minmax value is 1. So $v_1^{\min} = v_2^{\min} = 1$.

Feasible and individually rational set. The subset of the quadrilateral with $v_1 > 1$ and $v_2 > 1$. This is a triangle with vertices (approximately) $(3, 3), (5, 1^+), (1^+, 5)$.

What the folk theorem says. For δ close to 1, every point in this triangle is an SPE average payoff. Cooperation $(3, 3)$, asymmetric $(4, 2)$ or $(2, 4)$, exploitation $(5, 1.01)$ with player 1 getting nearly the defect payoff — all are sustainable.

Construction for $(4, 2)$. This payoff can be reached by playing (C, C) half the time and (D, C) half the time, averaging $(4, 2)$. In the repeated game, alternate: period 1 play (C, C), period 2 play (D, C), period 3 (C, C), period 4 (D, C), and so on. This gives approximately $(4, 2)$ as average discounted payoff.

How is player 2 incentivized to keep playing C in odd-numbered periods, when they are getting "exploited"? The answer is the punishment threat: if they deviate, grim trigger to (D, D) forever gives them 1 (minmax). Since $2 > 1$, for δ close enough to 1 the threat sustains the asymmetric cooperation. The exact critical δ is below 1 and depends on the deviation gain, which for player 2 in period 1 is $5 - 3 = 2$ (deviating to D against C). Cost of switching to (D, D) forever: $2 - 1 = 1$ per period. Critical $\delta \geq 2/3$.

Try $\delta = 0.8$: future cost is $0.8 \cdot 1/(1 - 0.8) = 4 > 2$. ✓ Sustainable.

For a more asymmetric division like $(4.5, 1.5)$, player 2's individually-rational margin is only 0.5, and the deviation gain from C against C is still 2 (the same). Critical $\delta \geq 2/(0.5/(1 - 0.8))$ — working this out, the cost needs to exceed the gain, so $\delta \cdot 0.5/(1 - \delta) \geq 2$, giving $\delta \geq 4/4.5 \approx 0.89$. So more patient players are needed for more asymmetric equilibria. As the target approaches the individual rationality boundary, the required δ approaches 1.

Poker Anchor

Concept mapping

Formal object	Poker instantiation
Feasible payoff set	All possible long-run win rate distributions at a given table
Minmax value	The worst long-run win rate opponents can force on you if they coordinate (you can always guarantee at least this much by playing defensively)
Individually rational set	Win rate distributions where each player does better than their minmax
Folk theorem	Any individually rational win rate distribution is sustainable in the long run with sufficiently patient players and informal agreements
Critical discount factor	Minimum session/relationship length needed to sustain a particular cooperative division

Concrete situation: soft home game with skill asymmetry

Six players sit in a weekly home game. Two are strong regulars; four are recreationals who are slightly losing. By applying maximum skill every hand, the regulars could push the recreationals to a much larger loss. But the regulars know that if they win too much too fast, the recreationals stop showing up and the game breaks. So there is an implicit folk-theorem equilibrium: the regulars play slightly sub-maximally, the recreationals lose slightly, and everyone keeps showing up.

The feasible payoff set here is all possible allocations of the weekly pool. The minmax for a recreational is their loss when regulars play maximally. The minmax for a regular is zero (they cannot be forced to lose in a soft game — they can always fold everything). The individually rational set is roughly: regulars winning between zero and their "maximum-skill" win rate, and recreationals losing between their "maximum-exploited" loss and zero.

Within this set, the actual equilibrium is somewhere in the middle — regulars winning at maybe 60% of their maximum rate, recreationals losing at maybe 60% of their maximum rate. The specific point is selected by social dynamics: what pace of losing is "acceptable" to the recreationals without them leaving. This is a folk-theorem equilibrium of the (implicit) repeated game between the regulars and recreationals.

Punishment structure. If one regular started playing maximally, the others could "punish" them by excluding them from the game or targeting them socially. This is an Abreu-style stick-and-carrot: short-term exclusion followed by return to cooperation. The credibility of this punishment is what keeps any individual regular from defecting.

Concrete situation: the whole poker ecosystem

Consider poker itself as a giant folk-theorem equilibrium. Strong players could, in principle, drive weaker players out of the game by winning maximally. But they do not, because if they did, the game would shrink (no more weak players to win against). So there is an implicit ecosystem-wide norm: win enough to make a living, but not so much that the games dry up. The specific rate at which strong players win from weak players is a selected point in the folk-theorem set of the repeated ecosystem game.

The institutional structures — rake, tournament structures, stakes distribution, site selection — are all mechanisms for implementing a particular selected equilibrium in the folk-theorem set. Sites that want to maximize long-run revenue do not want winners to win too fast; they adjust rake and rewards to push the ecosystem toward an equilibrium that keeps recreationals in the game. This is mechanism design applied to the folk-theorem feasible set.

What this understanding buys you

Every sustained poker norm is an equilibrium selection. When you see a pattern in poker — recreational-friendly site design, implicit caps on aggression in social games, the existence of loose-passive home games — you are looking at a selected point in a folk-theorem equilibrium set. The selection is done by social and institutional forces, not by game-theoretic uniqueness.

You cannot predict poker outcomes from equilibrium alone. The folk theorem's punchline applies: in repeated poker relationships, any individually rational division of winnings is a possible equilibrium. The specific division depends on the selection mechanism (social norms, site policies, reputation, personality). Equilibrium theory tells you what is feasible; it does not tell you what will happen.

Institutional design is central. If you want a specific pattern of play (say, balanced cooperation among regulars, or sustainable losses from recreationals), you need to design the institution to select that equilibrium. Changing the rake, changing the rewards structure, changing the social rules — these are all equilibrium-selection moves, and they are the practical implementation of the folk theorem.

Cross-Links

- **5.2 Infinitely repeated games** — the framework.
- **5.3 Trigger strategies** — the construction.
- **5.5 Reputation** — an alternative route to cooperation that sidesteps the folk-theorem selection problem.
- **5.6 Cooperation in repeated Prisoner's Dilemma** — the canonical case study.
- **6.5 Implementation theory** — how to select a specific equilibrium from a multiple-equilibrium set.
- **9.5 Convergence to equilibrium** — which folk-theorem equilibria are limits of learning.
- **10.1 Games embedded in institutions** — institutional selection among folk-theorem equilibria.

Structural Compression

Invariant: For sufficiently patient players in an infinitely repeated game, the set of subgame-perfect equilibrium average discounted payoffs coincides with the set of feasible and (strictly) individually rational payoffs of the stage game, provided a full-dimensionality condition holds; every such payoff is sustainable by an Abreu stick-and-carrot construction.

Mechanism: Feasibility is given by the convex hull of stage payoffs; individual rationality is given by the minmax value; credibility is achieved by follow-up rewards to punishers (the "carrot"); patience is quantified by the discount factor, with larger δ supporting more asymmetric and tighter equilibria.

Cross-System Echo: The folk theorem is the game-theoretic analogue of "controllability" in dynamical systems: for a fully actuated system with enough time, any trajectory in a large feasible region can be reached by some control input. The repeated game is "fully actuated" in the sense that players have access to every stage profile through public randomization, and the "enough time" is the large-δ condition. The analogy is structural: controllability is about reachability; the folk theorem is about sustainability as equilibrium. The similarity will be used in Module 10 when we discuss adaptive institutional design.

Exercises

1. For the stage game G with payoff vectors $(4,4), (0,6), (6,0), (2,2)$ on the profiles $(C,C), (C,D), (D,C), (D,D)$, compute the feasible payoff set, the minmax values, and the individually rational set. Sketch them in $\mathbb{R}^2$.
2. For the same stage game, find the critical discount factor to sustain the payoff vector $(5,3)$ as an SPE via grim trigger using the stage NE as punishment.
3. Show that the Nash-reversion folk theorem (Friedman 1971) gives a smaller equilibrium set than the Fudenberg–Maskin SPE folk theorem for games where the stage NE is Pareto-dominated by the minmax.
4. For the Battle of the Sexes game, compute the minmax values and identify the individually rational feasible set. Construct an SPE of the infinitely repeated game that gives a payoff vector on the interior of this set.
5. A 3-player game has stage payoffs for profiles that are complicated — compute the folk-theorem feasible set and verify the full-dimension condition. If full dimension fails, exhibit a payoff vector in the individually rational set that is not an SPE average.
6. Prove that as $\delta \to 1$, the set of SPE average payoffs converges (in the Hausdorff metric) to the closure of F^*. Use the compactness of F^* and the Abreu construction.
7. **Argue, structurally,** why the folk theorem is both a positive and a negative result. Give two concrete examples (one poker, one non-poker) of folk-theorem multiplicity in real-world strategic situations, and explain how the actual equilibrium is selected from the folk-theorem set in each.

Game Theory · Module 5 — Repeated Games · Node 5.4
Concepts: nash-equilibrium, repeated-interaction, convexity
Prerequisites: 5.2, 5.3
Enables: 5.5, 5.6, 6.5