From LP to Iterative Approaches

Scalability and Learning in Zero-Sum Games

WarmUp Review

Theorem 3.4.4 (Minimax theorem (von Neumann, 1928)) *In any finite, two-player, zero-sum game, in any Nash equilibrium*[5] *each player receives a payoff that is equal to both his maxmin value and his minmax value.*

Definition 3.4.1 (Maxmin) *The* maxmin strategy *for player i is* $\arg\max_{s_i} \min_{s_{-i}} u_i(s_i, s_{-i})$, *and the* maxmin value *for player i is* $\max_{s_i} \min_{s_{-i}} u_i(s_i, s_{-i})$.

WarmUp Review

0.2	0	-1	+1
0.5	+1	0	-1
0.3	-1	+1	0

Correlated Equilibrium

- Mediator suggests actions to all players before play
- Correlated equilibrium if everyone is incentivized to take suggested action
- NE of stoplight game
  - Two pure strategy NE
  - Mixed NE: Go w/ prob. 1/11
- Correlated equilibrium
  - Could suggest mixture of (Stop,Go) and (Go, Stop)

	Stop	Go
Stop	0, 0	0, 1
Go	1, 0	-10, -10

$$\mathbb{E}_{a\sim D}[u_i(a)] \geq \mathbb{E}_{a\sim D}[u_i(a_i', a_{-i})|a_i]$$

Maxmin Program & Dual

$$
\begin{aligned}
\text{maximize} \quad & U_1^* && && (4.5)\\
\text{subject to} \quad & \sum_{j \in A_1} u_1(a_1^j, a_2^k) \cdot s_1^j \geq U_1^* && \forall k \in A_2 && (4.6)\\
& \sum_{j \in A_1} s_1^j = 1 && && (4.7)\\
& s_1^j \geq 0 && \forall j \in A_1 && (4.8)
\end{aligned}
$$

$$
\begin{aligned}
\text{minimize} \quad & U_1^* && && (4.1)\\
\text{subject to} \quad & \sum_{k \in A_2} u_1(a_1^j, a_2^k) \cdot s_2^k \leq U_1^* && \forall j \in A_1 && (4.2)\\
& \sum_{k \in A_2} s_2^k = 1 && && (4.3)\\
& s_2^k \geq 0 && \forall k \in A_2 && (4.4)
\end{aligned}
$$

Maxmin / LP Properties

- Value of a game: Player 1's maxmin value
  - Well-defined, unique
  - Equilibrium mixed strategies not necessarily unique, though.
- Solvable in polynomial time

Maxmin / LP Properties

- Value of a game: Player 1's maxmin value
  - Well-defined, unique
  - Equilibrium mixed strategies not necessarily unique, though.
- Solvable in polynomial time
  - Q. Polynomial in what?

Maxmin / LP Properties

- Solvable in polynomial time
  - Q. Polynomial in what?
  - A. The size of the LP = number of variables and constraints + encoding length of coefficients
  - Modern interior point methods:
    - Roughly $|LP|^3$

Game Representations

- Normal-Form:
  - Every players utility for every action profile is explicitly listed ("payoff matrix")
- A more practical representation: Extensive-Form (Game Tree)
  - Exponential blow-up when reduced to normal-form

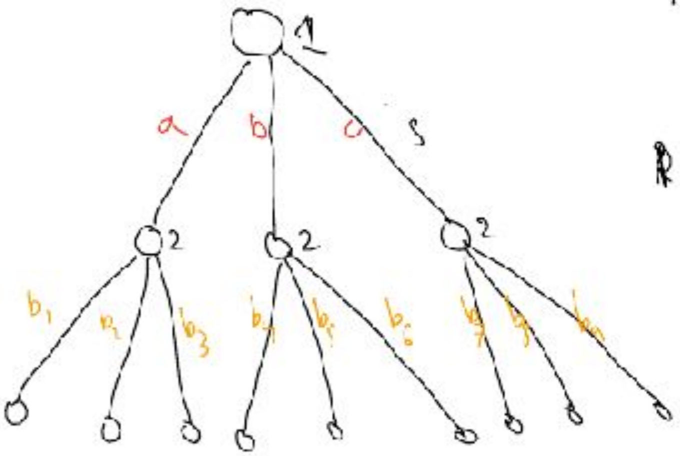
Examples

1) Sequential rock–paper–scissors (toy example)

- Say Player 1 secretly chooses R/P/S. Player 2 then chooses R/P/S.
- Extensive form: 12 nodes
- Normal form:
  - P1 pure strategies: just R, P, S.
  - P2 pure strategies: she must specify a move for every possible decision point → $3^3=27$
  - Thus, $|NF| = 27 \times 3$
- Blowup factor: 81 vs. 12

Examples

1) Sequential rock–paper–scissors (toy example)



Player 1's Pure Strategies

a, b, c $\quad$ 3

Player 2's Pure Strategies

(b_1, b_4, b_7), (b_1, b_5, b_7)

(b_1, b_6, b_7), (b_1, b_4, b_8)

...

$3^3 = 27$

Examples

- 2) Simple poker hand (Kuhn poker)
  - 2 players, 3-card deck, each antes 1 chip.
  - Tree size: ~12 decision nodes.
  - Normal form:
    - Each player's pure strategy specifies an action (bet/fold) at every information set.
    - Each has 4 information sets, so ~2^4=16 pure strategies each.
  - NF size: payoff matrix 16 × 16 = 256 entries.
  - Still solvable by hand, but you already see combinatorial growth.

Examples

- 3) Leduc poker (classic benchmark)
  - 2 players, 6-card deck (pairs + suits), structured betting.
  - Game tree: ~936 decision nodes.
  - Normal form:
    - Each player has to specify an action at every possible history they could face.
    - Pure strategies: $\sim 10^{12}$ per player.
  - NF size: payoff matrix of size $\sim 10^{12} \times 10^{12}$
    - Completely impossible to even write down.
    - Yet the extensive form is still just a few pages of rules

Examples

- 4) Poker at scale (No-Limit Hold'em)
  - Game tree size: ~10^{161} infosets.
  - Normal form: astronomically large — far more pure strategies than atoms in the universe.
  - That's why all practical solvers (e.g., Libratus, Pluribus) use regret minimization (CFR variants) in the extensive form.

Normal Form vs. Extensive Form Size

Game Example	Tree / Rules Size (extensive form)	Normal Form Pure Strategies	NF Payoff Matrix Size
Sequential RPS (toy)	Tiny tree (1 move each)	P1: 3, P2: $3^3 = 27$	3×27=81
Kuhn Poker (3 cards)	~12 decision nodes	~16 each	16×16=256
Leduc Poker (benchmark)	~936 decision nodes	~10^{12} each	~10^{24} entries
No-Limit Hold'em (2p)	~10^{161} infosets	astronomically many	infeasible to write

Normal Form vs. Extensive Form Size

Game Example	Tree / Rules Size (extensive form)	Normal Form Pure Strategies	NF Payoff Matrix Size
Sequential RPS (toy)	Tiny tree (1 move each)	P1: 3, P2: 3^3 = 27	3×27=81
Kuhn Poker (3 cards)	~12 decision nodes	~16 each	16×16=25616 \times 16 = 256
Leduc Poker (benchmark)	~936 decision nodes	~10^{12} each	~10^{24} entries
No-Limit Hold'em (2p)	~10^{161} infosets	astronomically many	infeasible to write

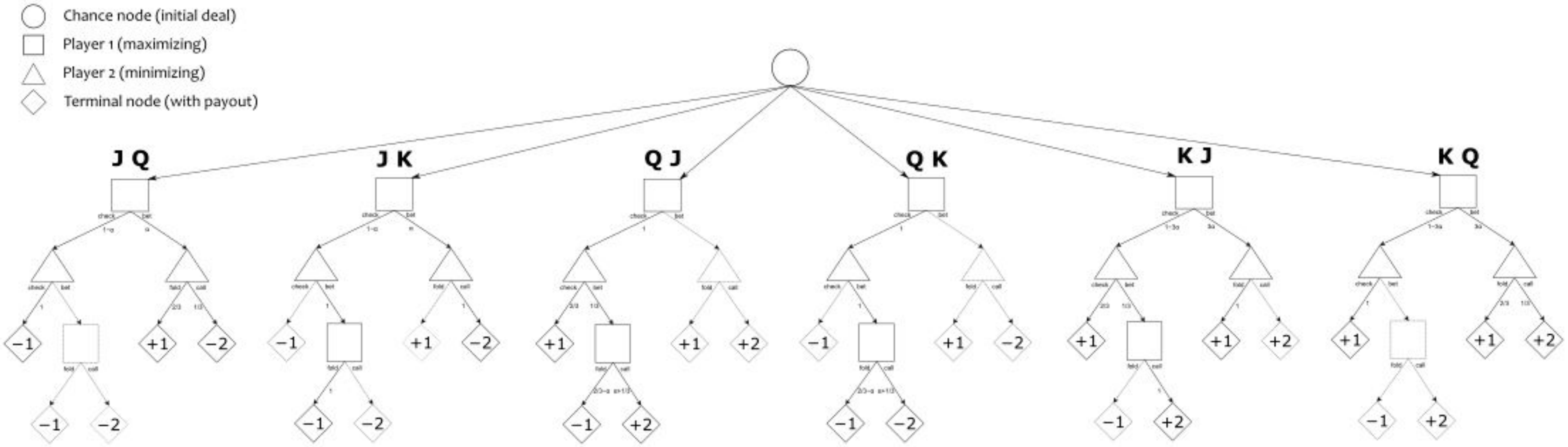
Even writing down the LP is infeasible for large games

Kuhn Poker

Rules

- Deck = {J, Q, K}. Each player antes 1 chip.
- Cards are dealt: each gets 1, one card unused.
- Betting: single round, possible actions = {check, bet}.
- If both check → showdown, high card wins pot (2).
- If one bets: opponent can fold (bettor wins pot) or call (showdown, pot = 4).

Kuhn Poker



Chance node (initial deal)
Player 1 (maximizing)
Player 2 (minimizing)
Terminal node (with payout)
J Q
J K
Q J
Q K
K J
K Q

Kuhn Poker

Pure Strategies

- Player 1: must specify action at two info sets:
  - Holding J/Q/K when first to act.
  - Holding J/Q/K facing a bet from P2 after checking.
- So: 2 actions per infoset × 3 cards = 2^3=8 possibilities for each decision point, times 2 → total ~16 pure strategies.
- Player 2: must specify action at two info sets:
  - With J/Q/K after P1 checks.
  - With J/Q/K after P1 bets.
- Same logic → ~16 pure strategies.
- Result: normal form payoff matrix = 6×16=256

- Compare: the tree only had ~12 actual decision nodes.

Iterative Approaches

Instead of solving the whole LP, let strategies evolve through repeated play.

Each round:

- Observe opponent's past play (or empirical distribution).
- Choose/update your own strategy.

Only need local payoff computations, not the whole matrix.

Self Play

- Both players learn best response to opponent's latest strategy
- Does not converge to a Nash equilibrium even in small games
- Will continue to cycle in games without pure strategy NE

Self Play

- Both players learn best response to opponent’s latest strategy
- Does not converge to a Nash equilibrium even in small games
- Will continue to cycle in games without pure strategy NE

$$x_i^{T+1} = \arg\max_{x_i \in \Delta} u_i(x_i, x_{-i}^T)$$

Self Play

- Step 1
  - P1 plays Rock.
  - Best response for P2 is Paper (since Paper beats Rock).
- Step 2
  - P2 plays Paper.
  - Best response for P1 is Scissors (beats Paper).
- Step 3
  - P1 plays Scissors.
  - Best response for P2 is Rock (beats Scissors).
- Step 4
  - P2 plays Rock.
  - Best response for P1 is Paper (beats Rock).

...and the cycle continues:
Rock → Paper → Scissors → Rock →...

0.2	0	-1	+1
0.5	+1	0	-1
0.3	-1	+1	0

$$x_i^{T+1} = \arg\max_{x_i \in \Delta} u_i(x_i, x_{-i}^T)$$

Fictitious Play (Follow the Leader

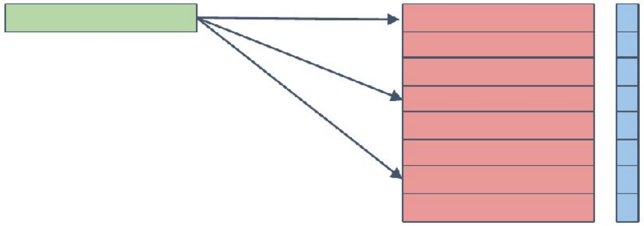
- Both players learn best response to opponent's **average strategy**

$$x_i^{T+1} = \arg\max_{x_i \in \Delta} u_i(x_i, x_{-i}^T)$$

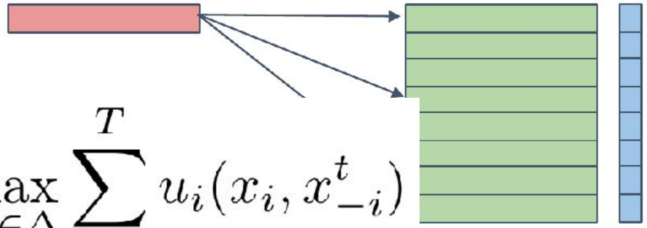
Fictitious Play (Follow the Leader

- Both players learn best response to opponent's **average strategy**
- Average strategy converges to a Nash equilibrium (Robinson, 1951)

Player 1 Best Responds to Player 2's Average Policy



Player 2 Best Responds to Player 1's Average Policy



$$x_i^{T+1} = \cancel{\arg\max_{x_i \in \Delta} u_i(x_i, x_{-i}^T)} \quad \longrightarrow \quad x_i^{T+1} = \arg\max_{x_i \in \Delta} \sum_{t=1}^{T} u_i(x_i, x_{-i}^t)$$

Fictitious Play (FTL) Example

- Example 1: Coordination Game

	Left	Right
Left	8, 8	0, 0
Right	0, 0	8, 8

Fictitious Play (FTL) Example

- Example 2: RPS

0.2	0	-1	+1
0.5	+1	0	-1
0.3	-1	+1	0

A Problem. A Regret?

- What if I'm playing a repeated game against someone who knows I am playing fictitious play?
- Then they would know exactly what my next move will be and could choose a best response every time
- Can we find iterative algorithms that will not be too bad even when the opponent knows the algorithm?
- No-regret algorithms do exactly this
  - And achieve faster convergence than FP as well!

Regret

for $t = 1, \dots, T$:

- Agent chooses an *action distribution* $x^t \in X := \Delta^n$
- Environment chooses a *utility vector* $u^t \in [0,1]^n$
- Agent observes u^t and gets utility $\langle u^t, x^t \rangle$

Δ^n = set of distributions on n things
$= \{x \in \mathbb{R}^n : x \geq 0, \sum x_i = 1\}$

- Regret = how much better you could have done if you had always played the best fixed action in hindsight.
- Goal: If regret grows sublinearly, average regret goes to 0.
- Intuition: “I don’t look back wishing I’d stuck to one action all along.”

$$R^T := \max_{\hat{x} \in X} \boxed{\left\{ \sum_{t=1}^{T} \langle u^t, \hat{x} \rangle \right\}} - \boxed{\sum_{t=1}^{T} \langle u^t, x^t \rangle}$$

Maximum utility that was achievable by the **best fixed** action in hindsight

Utility that was actually accumulated

No Regret ⇒ Equilibrium

- Theorem (informal): In 2-player zero-sum games, if both players use no-regret algorithms, the empirical distributions of play converge to Nash equilibrium.

- Key consequence:
  - You don’t need to solve the LP.
  - Just play the game repeatedly with a no-regret rule.

Why This Matters

- LP is exact, but infeasible at scale.
- Iterative self-play is unstable.
- **No-regret learning is the scalable, principled fix.**
- This is the foundation of modern game-solving (e.g., CFR in poker).