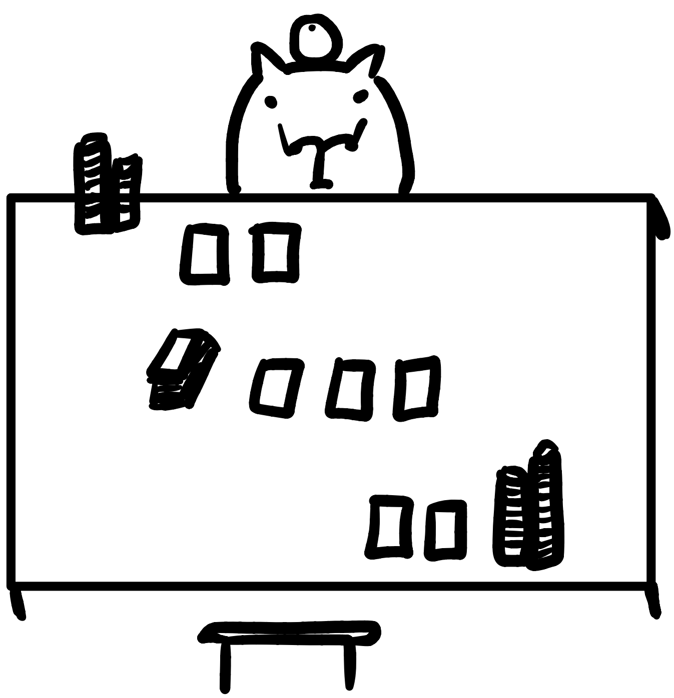
CMU 10-888 F21

NOTES

wanshenl



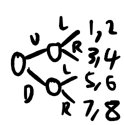
Lecture 09/07 Introduction

- Logistics
  - cs.cmu.edu/~sandholm/cs15-888F21/
  - Tuomas Sandholm, Gabriele Farina
  - 50% project, 40% homework, 10% participation
  - No textbook

- Multi-step imperfect information games
  - Most similar to real world – incomplete information, sequential/simultaneous moves
- Heart of problem
  - 1 agent: expected utility maximizing strategy well-defined
  - Multi agent: best strategy depends on others

- Terminology
  - Agent : player
  - Action / move : choice agent can make at any point in the game
  - Strategy s_i : mapping history (from agent i pov) → actions
  - Strategy set S_i : strategies available to agent i
  - Strategy profile $(s_1, s_2, \dots, s_{|A|})$: one strategy for each agent (though nature ∉ strategy profile)
  - Utility : $u_i = u_i(s_1, s_2, \dots, s_{|A|})$, can include nature for uncertainty
- Agenthood
  - Agent wants to maximize expected utility
  - Utility function u_i of agent i maps outcomes to reals
- Utility functions are scale-invariant
  - Agent i picks $\max_{strategy} \sum_{outcome} p(outcome|strategy)\, u_i(outcome)$
  - If $u_i' = a u_i + b$ for $a > 0$ then agent picks same strategy under u_i' and u_i – note u_i must be finite for comparisons to work
  - Inter-agent utility comparison problematic

- Game representation
  - Extensive/tree form
  - Matrix/normal/strategic form



combinatorial explosion →

	LL	LR	RL	RR
U	1,2	1,2	3,4	3,4
D	5,6	7,8	5,6	7,8

- Dominant strategy equilibrium
  - Best response s_i^* : for all s_i', $u_i(s_i^*, s_{-i}) \geq u_i(s_i', s_{-i})$ — s_{-i} = other player's strategy
  - Dominant strategy s_i^*: s_i^* is a best response for all s_{-i}, i.e., $\forall s_{-i} \forall s_i'\ u_i(s_i^*, s_{-i}) \geq u_i(s_i', s_{-i})$
    - Doesn't always exist
    - Inferior strategies are dominated
  - Dominant strategy equilibrium: strategy profile where each agent has picked dominant strategy
    - Doesn't always exist
    - Requires no counterspeculation
    - Prisoner's Dilemma

	C	D
C	3,3	0,5
D	5,0	1,1

Dominant strategy is (D,D)

Lecture 09/07 cont.

- **Nash equilibrium**
  - A strategy equilibrium is a Nash equilibrium if no player has an incentive to deviate from their strategy given that others do not deviate
  - For every agent i, $u_i(s_i^*, s_{-i}) \geq u_i(s_i', s_{-i})$ for all s_i', i.e., **for fixed s_{-i} $\forall s_i'$ $u_i(s_i^*, s_{-i}) \geq u_i(s_i', s_{-i})$**
    - **Dominant ⇒ Nash, not vice versa**

	boxing	ballet
boxing	2,1	0,0
ballet	0,0	1,2

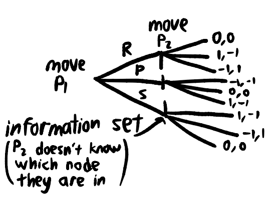
Nash equilibria
(boxing, boxing), (ballet, ballet)

- **Criticisms**
  - Not necessarily unique
    - Refinements (strengthenings) of equilibrium
      - Eliminate weakly dominated strategies
      - Choose Nash w/ highest welfare
      - Subgame perfection
    - Focal points
    - Mediation
    - Communication
    - Convention
    - Learning
  - Does not exist in all games

	1,0	0,1
	0,1	1,0

- **Existence of pure-strategy Nash equilibria**
  - Theorem
    Any **finite** game
    where each action node is **alone** in its information set (at every point, agent whose turn it is to move knows all moves so far)
    is **dominance solvable** by **backward induction** (as long as ties are ruled out)
  - Proof by construction, **multiplayer minimax**

- **Mixed-strategy Nash equilibria**
  - The essence of **being simultaneous is knowledge**
  - Still can draw same tree, just Player 2 doesn't know which state they are in
  - Dashed line for information set



- Bayes-Nash equilibrium: each agent uses best-response strategy and has consistent beliefs
- RPS has symmetric mixed-strategy Nash equilibrium where each player plays each pure strategy with probability 1/3
- In a mixed-strategy equilibrium, each strategy in agent i's mix has equal expected utility

- **Existence and complexity of mixed strategy Nash equilibria**
  - [Nash 50] Every **finite player, finite strategy** game has **at least one Nash equilibria** if we admit mixed-strategy equilibria as well as pure
  - **2-player 0-sum** → **polytime w/ LP**
  - **2-player games** → **PPAD-complete** (even with 0/1 payoffs)
    → **NP-complete** to find even approximately good Nash equilibria
  - **3-player games** → **FIXP-complete**

Lecture 09/07 cont.

- **2-player 0-sum games**
  - **Swappability**: if (x,y) and (x',y') equilibria, so are (x',y) and (x,y')
    - No equilibrium selection problem
  - Equilibrium strategies form **bounded convex polytope**
  - **Any convex combination** of a player's equilibrium strategies is an equilibrium strategy
  - [von Neumann 1928] **Minmax thm.**
    - Let $X \subset \mathbb{R}^n, Y \subset \mathbb{R}^m$ compact convex sets.
    - If $f: X \times Y \to \mathbb{R}$ continuous concave-convex,

      $f(\cdot, y)$ concave for fixed y

      $f(x, \cdot)$ convex for fixed x
    - Then

      $$\max_{x \in X} \min_{y \in Y} f(x,y) = \min_{y \in Y} \max_{x \in X} f(x,y)$$

set $S \subseteq \mathbb{R}$
- **Compact**: every sequence in S has subsequence converging to a point in S
- **convex**: contains line segment between any 2 points in set
  $x_1, x_2 \in S,\ 0 \le \theta \le 1 \Rightarrow \theta x_1 + (1-\theta) x_2 \in S$

function f on interval I,
- **concave**: $\forall x_1, x_2 \in I\ \ \theta \in [0,1]\ \ f(\theta x_1 + (1-\theta) x_2) \ge \theta f(x_1) + (1-\theta) f(x_2)$
- **convex**: $\forall x_1, x_2 \in I\ \ \theta \in [0,1]\ \ f(\theta x_1 + (1-\theta) x_2) \le \theta f(x_1) + (1-\theta) f(x_2)$

  - Great for multi-step imperfect information games
    - Opponent can play non-equilibrium to cause our beliefs to be wrong, but not enough to raise EV (expected value)
  - Solvable in polytime (size of game tree) using LP
    - Game tree may be infeasibly huge, 10^{165} eg

Lecture 09/09

- **Comparison**

Simultaneous Games	Sequential Games
$\Delta^{\lvert A \rvert}$	Q
Convex polytope	Convex polytope
Multi-Bilinearity of expected utility	Multi-Bilinearity of expected utility
Nash equi in 2-player 0-sum is $\max_{x \in \Delta^{\lvert A \rvert}} \min_{y \in \Delta^{\lvert A \rvert}} x^T A y$	Nash is now $\max_{x \in Q_1} \min_{y \in Q_2} x^T A y$
Low # of vertices	Combinatorial # of vertices

- Set of actions A
- Strategy is probability distribution over A

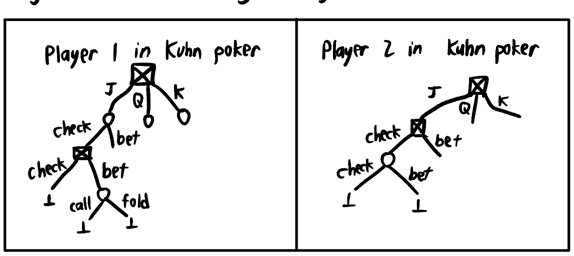
  $x = \begin{pmatrix} x_R \\ x_P \\ x_S \end{pmatrix} \ge 0$ where $x_R + x_P + x_S = 1$

  $x_1 \in \Delta^{\lvert A \rvert}$ $\quad x_1 = (x_{1R}, x_{1P}, x_{1S})$

  $x_2 \in \Delta^{\lvert A \rvert}$ $\quad x_2 = (x_{2R}, x_{2P}, x_{2S})$

- Nash equi in 2-player 0-sum simultaneous
  - P_1 commits to strategy $x \in \Delta^{\lvert A \rvert}$
  - P_2 plays $y^* = \arg\max_y u_2(x,y) = \arg\max_y x^T A_2 y$
  - P_1 can expect to receive utility $g(x) = -u_2(x, y^*) = \min_{y^*} x^T A_1 y^*$

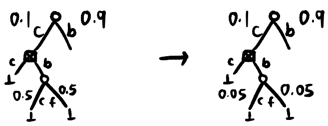
- **Sequential Games**
  - **Decision nodes** $\circ$
  - **Observation nodes** $\boxtimes$



Lecture 09/09 cont.

- **Behavioral strategy** $(0.1, 0.9; 0.5, 0.5; \dots; 1.0, 0.0)$ probabilities on game tree branches

  Calculating expected utility needs product
  - Causes non-convex problems because objective contains products of own variables
- **Sequence form strategy** $(0.1, 0.9; 0.5\times0.1, 0.5\times0.1; \dots; 1.0, 0.0)$
  - Pre-multiply

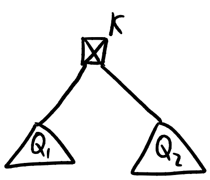


  - Check goes from "sum to 1" to "sum to parent"
  - x is a **valid seq form strategy** iff
    ① $x \geq 0$
    ② $\sum_{a \in A_j} x[ja] = x[p_j] \quad \forall j$ note p_j of a j with no parent is denoted $p_j = \phi$
    ③ $x[\phi] = 1$
  - We denote valid seq form strategies as **Q**

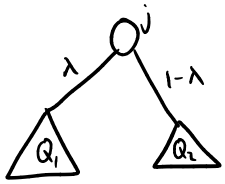
- **Deterministic strategies**
  - $\Pi = Q \cap [0,1]^n$ where $n = 1 + \sum_j |A_j|$ (1 for ϕ)
  - Lemma: **Q = convex hull of Π**
  - Usually $|\Pi|$ exp in size of tree

  = { flip all coins upfront / flip coins as needed }, same thing

- **Inductive Q construction**



$$Q = Q_1 \times Q_2$$



$$Q = \left\{ (\lambda, 1-\lambda; \lambda q_1, (1-\lambda) q_2) : (\lambda, 1-\lambda) \in \Delta^2,\ q_1 \in Q_1,\ q_2 \in Q_2 \right\}$$

$$= \text{convex hull of } \left\{ \begin{pmatrix} 1 \\ 0 \\ Q_1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ Q_2 \end{pmatrix} \right\}$$

— = vertical ;

  - Only 3 operations
    - **Cartesian product**
    - **Convex hull**
    - **"Padding"**

Lecture 09/14

- **Regret**: look at game history, do we regret any action taken
- The player has learnt to play the game when looking back at the history of play, they cannot think of any transformation $\phi: X \to X$ of their strategies that when applied to the whole history of play would have given a strictly better utility to the player } **Hindsight rationality**

Lecture 09/14 cont.

One t = one game, e.g., one poker hand

- Φ-regret minimizer for set X
  - A device where at every time t
  - ① NEXT STRATEGY : output the next strategy $x^t \in X$
  - ② OBSERVE UTILITY(ℓ^t) : $\ell^t : X \to \mathbb{R}$ linear function where x^t scores a utility of $\ell^t(x^t)$
  - Quality metric : Φ-regret

$$R_\Phi^T = \max_{\hat{\phi} \in \Phi} \sum_{t=1}^{T} \left[\ell^t(\hat{\phi}(x)) - \ell^t(x^t) \right]$$

  - Goal
    Have guaranteed $R_\Phi^T = o(T)$ no matter the sequence of utility functions
  - Note that Φ is a set of functions from X to X

- Notable choices for Φ
  - Φ is the set of all mappings $X \to X$
    - "Swap regret minimization"
      - Converges to correlated equilibrium
        in normal form and extensive form, general-sum, multiplayer
  - $\Phi = \{\phi_{a\to b} : a, b \in X\}$ where $\phi_{a \to b}(x) = x$ if $x \neq a$, b if $x = a$
    - "Internal regret minimization"
    - Only one swap
  - Φ = constant functions from X to X
    - $\Phi = \{\phi_{\hat{x}} : \hat{x} \in X\}$ where $\phi_{\hat{x}}(x) = \hat{x}$ $\forall x \in X$
    - "External regret minimization"

- In 2-player 0-sum games (extensive form/normal form)
  then external-regret-minimizing strategies
  converge to Nash equilibrium (in averages)
  i.e., maybe no strategy converges
  but the average of all strategies will

- In general-sum multiplayer games (extensive form/normal form)
  then external-regret-minimizing strategies
  converge to a coarse correlated equilibrium (in empirical frequency)

mediator enforces a certain strategy

- In general-sum multiplayer games
  if all but one player stochastic
  and last player uses regret minimizing strategy
  then last player strategy converges to best response

- [Gordon, Greenwald, Narks 08]
  WANT: Φ-regret minimizer for X, R_Φ
  HAVE: ① an external regret minimizer for Φ, $\mathcal{R}$
  ② for any $\phi \in \Phi$ we have a fixed-point oracle, $X \ni x = \phi(x)$

Lecture 09/14 cont.

- [Gordon 08] cont

  Algorithm $\mathcal{R}_\Phi$ at each time t

  ① NEXT STRATEGY

  $\phi^t \leftarrow \mathcal{R}$ NEXT STRATEGY

  return $x^t \in \phi^t(x^t) \in X$

  ② OBSERVE UTILITY (ℓ^t) $\quad \ell^t : X \to \mathbb{R}$ linear

  define $L^t : \phi \to \ell^t(\phi(x^t))$ $\quad L^t : \Phi \to \mathbb{R}$ linear

  $\mathcal{R}_{\text{OBSERVE UTILITY}}(L^t)$

- R_Φ^T of $\mathcal{R}_\Phi$

$$R_\Phi^T = \max_{\hat\phi \in \Phi} \sum_{i=1}^{T} \left[\ell^t(\hat\phi(x^t)) - \ell^t(x^t)\right]$$

$$= \max_{\hat\phi \in \Phi} \sum_{i=1}^{T} \left[\ell^t(\hat\phi(x^t)) - \ell^t(\phi^t(x^t))\right] \quad \text{via fixed point oracle}$$

$$= \max_{\hat\phi \in \Phi} \sum_{i=1}^{T} \left[L^t(\hat\phi) - L^t(\phi^t)\right]$$

= external regret on the set of "strategies" Φ for $\mathcal{R}$

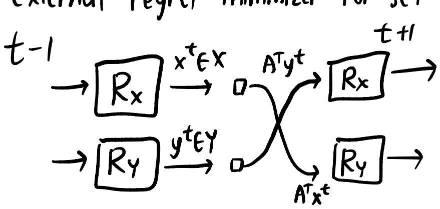
Lecture 09/16

- **Bilinear saddle-point function**

$$\max_{x \in X} \min_{y \in Y} x^T A y$$

  - Models
    - Nash eq in 2-player 0-sum games
    - 0-sum team games
    - Optimal correlated equilibrium
- **Idea: self-play**
  - Have: R_X external regret minimizer for set X

    R_Y external regret minimizer for set Y



- **"Saddle-point gap"**

$$\gamma(\underset{X}{\bar{x}}, \underset{Y}{\bar{y}}) = \underbrace{\left(\max_{\hat{x} \in X} \hat{x}^T A \bar{y} - \bar{x}^T A \bar{y}\right)}_{\geq 0} + \underbrace{\left(\bar{x}^T A \bar{y} - \min_{\hat{y} \in Y} \bar{x}^T A \hat{y}\right)}_{\geq 0}$$

  - γ often called **exploitability**

Lecture 09/16 cont.

- $R_x^T = \max_{\hat{x}\in X} \sum_{t=1}^T (Ay^t)^T(\hat{x}-x^t) = \max_{\hat{x}\in X} \sum_{t=1}^T \hat{x}^T A y^t - \sum_{t=1}^T (x^t)^T A y^t$

- $R_y^T = \max_{\hat{y}\in Y} \sum_{t=1}^T (-A^T x^t)(\hat{y}-y^t) = \max_{\hat{y}\in Y} \sum_{t=1}^T -(x^t)^T A\hat{y} + \sum_{t=1}^T (x^t)^T A y^t$

- $$\begin{aligned} R_x^T + R_y^T &= \max_{\hat{x}\in X} \sum_{t=1}^T \hat{x}^T A y^t + \max_{\hat{y}\in Y} \sum_{t=1}^T -(x^t)^T A \hat{y} \\ &= \max_{\hat{x}\in X} \sum_{t=1}^T \hat{x}^T A y^t - \min_{\hat{y}\in Y} \sum_{t=1}^T (x^t)^T A\hat{y} \\ &= T\Big[\max_{\hat{x}\in X} \hat{x}^T A\Big(\sum_{t=1}^T y^t\Big) - \min_{\hat{y}\in Y} \Big(\sum_{t=1}^T (x^t)\Big)^T A\hat{y}\Big] \\ &= \gamma(\bar{x},\bar{y}) \text{ where } \begin{array}{l} \bar{x} = \frac{1}{T}\sum_{t=1}^T x^t \\ \bar{y} = \frac{1}{T}\sum_{t=1}^T y^t \end{array} \end{aligned}$$

- $\Delta^n = \{(x_1,\dots,x_n) \in \mathbb{R}^n_{\geq 0} : x_1 + \dots + x_n = 1\}$

- Goal: construct an external regret minimizer for Δ^n
  - NEXT STRATEGY output $x^t \in \Delta^n$
  - OBSERVE UTILITY(ℓ^t) $\ell^t : \Delta^n \to \mathbb{R}$ linear

  $$R^T = \max_{\hat{x}\in\Delta^n} \Big\{\sum_{t=1}^T [\ell^t(\hat{x}) - \ell^t(x^t)]\Big\} = O(\sqrt{T})$$

- **Blackwell game**

  (X, Y, u, S)

  - X, Y are closed convex strategy spaces
  - $u : X \times Y \to \mathbb{R}^d$ biaffine utility of the game for player 1
  - S is a subset of $\mathbb{R}^d$, closed and convex, target set

- **Blackwell dynamics**
  1. Player 1 picks an action $x^t \in X$
  2. Player 2 picks an action $y^t \in Y$
  3. Player 1 incurs payoff $u(x^t, y^t) \in \mathbb{R}^d$

- **Blackwell game goal** $\frac{1}{T}\sum_{t=1}^T u(x^t,y^t) \to S$

  $\min_{\hat{s}\in S} \left\|\frac{1}{T}\sum u(x^t,y^t) - \hat{s}\right\|_2 \to 0$ as $T \to \infty$

- $\Gamma := (\overset{X}{\overset{\|}{\Delta^n}}, \mathbb{R}^n, u, \overset{S}{\overset{\|}{\mathbb{R}^n_{\leq 0}}})$

  $u(x^t, \ell^t) := \ell^t - \langle \ell^t, x^t\rangle \cdot \mathbb{1} \in \mathbb{R}^n$ where $\mathbb{1} = \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \in \mathbb{R}^n$

- $R^T = \max_{\hat{x}\in\Delta^n} \sum_{t=1}^T \langle \ell^t, \hat{x}\rangle - \sum_{t=1}^T \langle \ell^t, x^t\rangle$

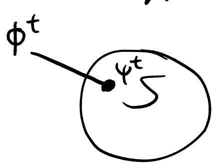
- Lemma $\frac{R^T}{T} \leq \underbrace{\min_{\hat{s}\in\mathbb{R}^n_{\leq 0}} \left\|\hat{s} - \frac{1}{T}\sum_{t=1}^T u(x^t,\ell^t)\right\|_2}_{\text{distance of } \frac{1}{T}\sum u(x^t,\ell^t) \text{ from } S}$

Lecture 09/16 cont.

- $H \subseteq \mathbb{R}^d$ halfspace
  $H = \{ h \in \mathbb{R}^d : a^T h \leq b \}$ for some $a \in \mathbb{R}^d$ $b \in \mathbb{R}$
- Forceable halfspace
  H is forceable if $\exists x^* \in X$ a forcing action such that
  $\forall y \in Y \quad u(x^*, y) \in H$

- Thm by Blackwell
  Blackwell goal can be obtained if every halfspace $H \supseteq S$ is forceable.

  At every t, Player 1 plays like this:
  ① Compute $\phi^t = \frac{1}{t-1} \sum_{i=1}^{t-1} u(x^i, y^i)$
  ② Project ϕ^t onto S, call the projection ψ^t
  ③ If $\phi^t \in S$, equivalently $\psi^t = \phi^t$, then play any $x^t \in X$
  ④ Else consider the halfspace H^* tangent to S at ψ^t and contains S, then play any forcing action for H^*
  ⑤ Observe y^t, incur $u(x^t, y^t)$, and repeat.



get your average in the set

- Proof sketch

$$\phi^{t+1} = \frac{t-1}{t} \phi^t + \frac{1}{t} u(x^t, y^t)$$

$$\begin{aligned}
dist(\phi^{t+1}, S)^2 &\leq \| \phi^{t+1} - \psi^t \|_2^2 \\
&= \| \tfrac{t-1}{t} \phi^t + \tfrac{1}{t} u(x^t, y^t) - \psi^t \|_2^2 \quad \text{, note } \psi^t = \tfrac{t-1}{t}\psi^t + \tfrac{1}{t}\psi^t \\
&= \| \tfrac{t-1}{t} (\phi^t - \psi^t) + \tfrac{1}{t} (u(x^t, y^t) - \psi^t) \|_2^2 \\
&= \left(\tfrac{t-1}{t}\right)^2 dist(\phi^t, S)^2 + \tfrac{1}{t^2} \| u(x^t, y^t) - \psi^t \|_2^2 + \tfrac{2}{t^2}(t-1) \cdot \underbrace{\langle \phi^t - \psi^t, u(x^t, y^t) - \psi^t \rangle}_{\text{use alg construction to show } \leq 0 \text{, so term can be ignored}}
\end{aligned}$$

So $dist(\phi^{t+1}, S)^2 \leq dist(\phi^t, S) \cdot \frac{(t-1)^2}{t^2} + \underbrace{\frac{1}{t^2} \| u(x^t, y^t) - \psi^t \|_2^2}_{O(1)}$

$\alpha^t := (t-1)^2 \, dist(\phi^t, S)^2$

$\alpha^{t+1} \leq \alpha^t + O(1) \Rightarrow \alpha^t = O(t)$

$O(t) = (t-1)^2 \, dist(\phi^t, S)^2$

$\Rightarrow dist(\phi^t, S) = O\left(\frac{1}{\sqrt{t}}\right)$

- Blackwell for Γ game earlier
  ① $\phi^t = \frac{1}{t-1} \sum_{\tau=1}^{t-1} u(x^\tau, y^\tau) = \frac{1}{t-1} \sum_{\tau=1}^{t-1} \ell^\tau - \langle \ell^\tau, x^\tau \rangle \cdot \mathbb{1}$
  ② $S = \mathbb{R}^n_{\leq 0}$, $\psi^t = [\phi^t]^-$
  ③ If $\phi^t \in S$ do anything
  ④ If $\psi^t \neq \phi^t$ $\quad H^t = \{ z \in \mathbb{R}^n : (\phi^t - \psi^t)^T z \leq 0 \}$
  $(\phi^t - [\phi^t]^-)^T z \leq 0$
  $([\phi^t]^+)^T z \leq 0$

Lecture 09/16 cont.

- Is it true that $\forall \phi^t$, there exists x^* st $\forall \ell \in \mathbb{R}^n$
  $u(x^*, \ell) \in H^+$

$$\Leftrightarrow ([\phi^t]^+)^T (\ell - \langle \ell, x^* \rangle \cdot 1) \le 0$$
$$\ell^T [\phi^t]^+ - (\ell^T x^*)([\phi^t]^+)^T 1 \le 0$$
$$\ell^T \frac{[\phi^t]^+}{[\phi^t]^+ \cdot 1} - \ell^T x^* \le 0$$
$$x^* = \frac{[\phi^t]^+}{[\phi^t]^+ \cdot 1} = \Delta^n$$

Lecture 09/21

- Algorithm
  - At every t,
    1. $\phi^t \leftarrow \frac{1}{t-1} \sum_{\tau=1}^{t-1} u(x^\tau, \ell^\tau)$
    2. $\psi^t \leftarrow [\phi^t]^-$
    3. If $\psi^t \neq \phi^t$ then play $\frac{[\phi^t]^+}{1^T [\phi^t]^+} \in \Delta^n$
       else play any point in Δ^n
    4. Observe ℓ^t and iterate
- $r^t = (t-1)\phi^t$

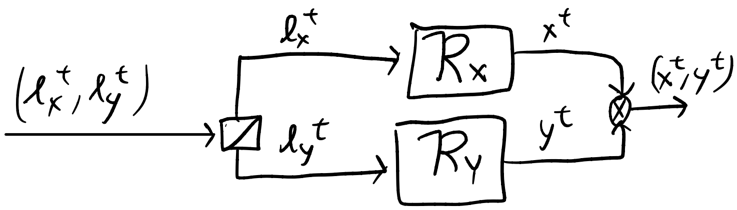
- Regret Matching
  - $r^0 \leftarrow 0$, $x^0 \leftarrow \frac{1}{n} \in \Delta^n$
  - function NEXT STRATEGY( )
    - $\theta^t \leftarrow [r^t]^+$
    - if $\theta^t \neq 0$ then output $\frac{\theta^t}{1^T \theta^t} \in \Delta^n$
    - else output $\frac{1}{n} 1 \in \Delta^n$
  - function OBSERVE UTILITY(ℓ^t)
    - $r^{t+1} \leftarrow r^t + \ell^t - \langle \ell^t, x^t \rangle 1$
    - In practice, faster to $[\cdot]^+$ the updates

- Regret circuit for Cartesian Product
  - Setting
    - Sets X and Y
    - Regret minimizers $\mathcal{R}_X, \mathcal{R}_Y$
  - Goal
    - Regret minimizer for $X \times Y = \{(x,y) : x \in X, y \in Y\}$

  - function NEXT STRATEGY
    - $x^t \leftarrow \mathcal{R}_X$.NEXT STRATEGY
    - $y^t \leftarrow \mathcal{R}_Y$.NEXT STRATEGY
    - output (x^t, y^t)
  - function OBSERVE UTILITY(ℓ^t) where $\ell^t = (\ell^t_X, \ell^t_Y)$
    - $\mathcal{R}_X$.OBSERVE UTILITY(ℓ^t_X)
    - $\mathcal{R}_Y$.OBSERVE UTILITY(ℓ^t_Y)

Lecture 09/21 cont.

$$
\begin{aligned}
R^T &= \max_{(\hat{x},\hat{y})} \sum_{t=1}^T (\ell_x^t, \ell_y^t)^T (\hat{x},\hat{y}) - (\ell_x^t, \ell_y^t)^T (x^t, y^t) \\
&= \max_{(\hat{x},\hat{y})} \sum_{t=1}^T \ell_x^{t^T} \hat{x} + \ell_y^{t^T} \hat{y} - \ell_x^{t^T} x^t - \ell_y^{t^T} y^t \\
&= \max_{(\hat{x},\hat{y})} \left\{ \left(\sum_{t=1}^T \ell_x^{t^T} \hat{x} - \ell_x^{t^T} x^t \right) + \left(\sum_{t=1}^T \ell_y^{t^T} \hat{y} - \ell_y^{t^T} y^t \right) \right\} \\
&= \left(\max_{\hat{x}\in X} \sum_{t=1}^T \ell_x^{t^T} \hat{x} - \ell_x^{t^T} x^t \right) + \left(\max_{\hat{y}\in Y} \sum_{t=1}^T \ell_y^{t^T} \hat{y} - \ell_y^{t^T} y^t \right) \\
&= R_X^T + R_Y^T
\end{aligned}
$$



Regret circuit for Convex Hulls

- Setting
  - Sets X and Y
  - Regret minimizers $\mathcal{R}_X, \mathcal{R}_Y$
- Goal
  - Regret minimizer for $co(X,Y) = \{\lambda_1 x + \lambda_2 y : x\in X, y \in Y, \lambda \in \Delta^2\} \in \mathbb{R}^n$
- Need
  - Any regret minimizer $\mathcal{R}_\Delta$ for Δ

- function NEXT STRATEGY
  - $x^t \leftarrow \mathcal{R}_X$.NEXT STRATEGY
  - $y^t \leftarrow \mathcal{R}_Y$.NEXT STRATEGY
  - $\lambda^t \leftarrow \mathcal{R}_\Delta$.NEXT STRATEGY
  - output $\lambda_1^t x^t + \lambda_2^t y^t$

- function OBSERVE UTILITY(ℓ^t) where $\ell^t = (\ell_x^t, \ell_y^t)$
  - $\mathcal{R}_X$.OBSERVE UTILITY(ℓ^t)
  - $\mathcal{R}_Y$.OBSERVE UTILITY(ℓ^t)
  - $\mathcal{R}_\Delta$.OBSERVE UTILITY(ℓ_Δ^t)
    - where $\ell_\Delta^t : (\lambda_1, \lambda_2) \to \lambda_1 \cdot \ell^t(x^t) + \lambda_2 \cdot \ell^t(y^t)$

$$
\begin{aligned}
R^T &= \max_{\hat{z}\in co(X,Y)} \sum_{t=1}^T \left[\ell^{t^T} \hat{z} - \ell^{t^T}(\lambda_1^t x^t + \lambda_2^t y^t) \right] \\
&= \max_{\substack{\hat{\lambda}\in\Delta^2 \\ \hat{x}\in X \\ \hat{y}\in Y}} \sum_{t=1}^T \left[\ell^{t^T}(\hat{\lambda}_1 \hat{x} + \hat{\lambda}_2 \hat{y}) - \ell^{t^T}(\lambda_1^t x^t + \lambda_2^t y^t) \right] \\
&= \max_{\substack{\hat{\lambda}\in\Delta^2 \\ \hat{x}\in X \\ \hat{y}\in Y}} \left\{ \hat{\lambda}_1 \sum_{t=1}^T \ell^{t^T} \hat{x} + \hat{\lambda}_2 \sum_{t=1}^T \ell^{t^T} \hat{y} \right\} - \sum_{t=1}^T \ell^{t^T}(\lambda_1^t x^t + \lambda_2^t y^t) \\
&= \max_{\hat{\lambda}\in\Delta^2} \left\{ \hat{\lambda}_1 \left(\max_{\hat{x}\in X} \sum_{t=1}^T \ell^{t^T}\hat{x} \right) + \hat{\lambda}_2 \left(\max_{\hat{y}\in Y} \sum_{t=1}^T \ell^{t^T}\hat{y} \right) \right\} - \sum_{t=1}^T \ell^{t^T}(\lambda_1^t x^t + \lambda_2^t y^t) \\
&= \max_{\hat{\lambda}\in\Delta^2} \left\{ \hat{\lambda}_1 \left(R_X^T + \sum_{t=1}^T \ell^{t^T} x^t \right) + \hat{\lambda}_2 \left(R_Y^T + \sum_{t=1}^T \ell^{t^T} y^t \right) \right\} - \sum_{t=1}^T \ell^{t^T}(\lambda_1^t x^t + \lambda_2^t y^t) \\
&\le \max_{\hat{\lambda}\in\Delta^2} \left\{ \hat{\lambda}_1 \left(\sum_{t=1}^T \ell^{t^T} x^t \right) + \hat{\lambda}_2 \left(\sum_{t=1}^T \ell^{t^T} y^t \right) \right\} - \sum_{t=1}^T \ell^{t^T}(\lambda_1^t x^t + \lambda_2^t x^t) + \max\{R_X^T, R_Y^T\} \\
&= \max_{\hat{\lambda}\in\Delta^2} \underbrace{\left\{ \sum_{t=1}^T \hat{\lambda}_1 \ell^{t^T} x^t + \hat{\lambda}_2 \ell^{t^T} y^t - \lambda_1^t \ell^{t^T} x^t - \lambda_2^t \ell^{t^T} y^t \right\}}_{\begin{pmatrix} \ell^{t^T} x^t \\ \ell^{t^T} y^t \end{pmatrix}^T \begin{pmatrix} \hat{\lambda}_1 - \lambda_1^t \\ \hat{\lambda}_2 - \lambda_2^t \end{pmatrix}} + \max\{R_X^T, R_Y^T\} \\
&= R_\Delta^T + \max\{R_X^T, R_Y^T\}
\end{aligned}
$$

Lecture 09/21 cont.

- Recall inductive Q construction
  - Can essentially ignore padding in convex hull

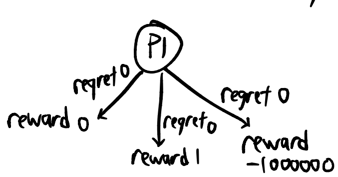
- **CFR**
  - Need R_j for every decision point j, a regret minimizer for $\Delta^{|A_j|}$
  - fn NEXT STRATEGY()
    $b^t \leftarrow$ Construct behavioral strategy that picks actions A_j at each j with probability R_j.NEXT STRATEGY()
    output sequence form representation x^t of b^t
  - fn OBSERVE UTILITY(ℓ^t)
    At each j, construct the $|A_j|$ dimensional utility vector
    $$\ell_j^t[a] = \ell^t[ja] + \sum_{j'a' \geq ja} x^t[j'a'] \cdot \ell^t[j'a']$$

Lecture 09/23 Speeding up CFR

- This lecture covers SOTA ~2019
- Convergence of CFR
  - Cumulative $O(\sqrt{T})$
  - Average $O(\frac{\sqrt{T}}{T}) = O(\frac{1}{\sqrt{T}})$

- Technique: **Alternation**
  - Normal CFR: update agent 1 and agent 2 based on opponent strategy in $t-1$
  - Now: update agent 1 based on agent 2 in $t-1$,
    update agent 2 based on agent 1 in t
  - Converges faster in practice, still provable $O(\sqrt{T})$ cumulative regret [Burch JAIR19]
  - Motivation: update each agent based on newest strategy of each opponent

- Technique: **Re-weighting**
  - [Brown, Sandholm, AAAI19]
  - Motivation
    - CFR+ was fastest, but has limitations



Iteration
1
probabilities 1/3, 1/3, 1/3
expected reward is −333333
update regret as action ev − achieved ev
↳ regrets become 333333, 333334, −666667 (↑ floored to 0)

2
probabilities 1/2, 1/2, 0
expected reward ≃ 0.5
update regret becomes 333332.5, 333334.5

    - Causes CFR+ to take 471407 iterations to learn to pick middle with 100% probability
  - Solution
    - Discount early bad iterations' regrets and average strategy by weighting iteration t by t
    - Called Linear CFR
    - Takes 970 iterations to learn to pick middle action
    - Worst case convergence bound only increases by $2/\sqrt{3}$

Lecture 09/23 cont.

- **Theorem.** For any sequence of nondecreasing weights,
  1. Suppose T iterations of RM+ in 2-player 0-sum games
  2. Then weighted avg strategy profile, where iteration t is weighted proportional to $w_t > 0$ and $w_i \le w_j$ for all $i < j$, is a

$$\frac{w_T}{\sum_{t=1}^{T} w_t} \Delta |I| \sqrt{|A|} \sqrt{T}\text{-Nash equilibrium}$$

- At least for now, smart reweighting doesn't seem to pay off
  - Too much to store
  - Time could be spent on just doing more CFR

- **Linear CFR+**
  - In theory, yes
  - In practice, does very poorly

- **Discounted CFR**, a less aggressive combination (DCFR)
  - On each iteration,
    - multiply positive regrets by $\frac{t^\alpha}{t^\alpha+1}$
    - multiply negative regrets by $\frac{t^\beta}{t^\beta+1}$
    - Weight contributions to average strategy by $\left(\frac{t}{t+1}\right)^\gamma$
  - For $\alpha=1.5, \beta=0, \gamma=2$, consistently outperforms CFR+ in practice
    - $\beta = \infty$ = no discounting = vanilla CFR
    - $\beta = 1$ = linear CFR
    - $\beta = -\infty$ = CFR+
  - Worst case convergence bound only a small constant worse than CFR

- CFR+ also works better when assigning iteration t a weight of t^2 than t, empirically
- The relative ranking is mostly the same across games for the empirical graph shown

- **Monte Carlo Linear CFR**
  - CFR+, DCFR do poorly with sampling
  - Linear CFR does quite well with sampling

- DCFR > CFR+, was SOTA in large imperfect information games
- Linear TODO COPY

- Technique: **Dynamic Pruning**
  - Why not permanent pruning like $\alpha\beta$ pruning in perfect information games?
    - Game tree can change!

- [Lanctot ICML09] **Partial pruning**
  - If opponent's probability of reaching there is 0, safe to prune

Lecture 09/23 cont.

- [Brown, Sandholm NeurIPS15] **Interval Regret Based Pruning**
  - Also prune paths that agent reaches with 0 probability
  - Must be temporary!
    - Action $a \in A(I)$ such that $\sigma^t(I,a) = 0$
    - Known a will not be played with positive probability until far future iteration t'
      - In RM, $R^t(I,a) << 0$
      - To find t', project conservatively or check dynamically
  - So we can procrastinate in deciding what happens before a on iterations $t, t+1, \ldots, t'-1$
  - Upon reaching t', instead of $t'-t$ iterations over $D(I,a)$, just one iteration playing the average of the opponent's strategies in those missed iterations, and declare we played that strategy on all those missed iterations
  - All other players can partial prune a out

- **Total Regret Based Pruning**

- Check slides for the rest

strategy = $(x_1, x_2; x_3, x_4, \ldots)$

Space → quantization?

iterations?

accuracy?

not true

any more?

Purification

poker competition

20mB limit

dropping

tartanian8 / baby

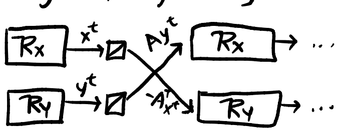
Lecture 09/28

- Recall **Regret Minimizer** from before
  - NEXT STRATEGY outputs $x^t \in X \subseteq \mathbb{R}^n$
  - OBSERVE UTILITY(ℓ^t) $\ell^t \in \mathbb{R}^n$
  - $R^T = \max_{\hat{x} \in X} \sum_{t=1}^T (\ell^t)^T \hat{x} - (\ell^t)^T x^t$
  - Goal: $R^T = o(T)$

- **Predictive Regret Minimizer**
  - NEXT STRATEGY$(m^t \in \mathbb{R}^n)$ outputs x^t taking into account a **prediction m^t for the next utility ℓ^t**
  - OBSERVE UTILITY(ℓ^t) $\ell^t \in \mathbb{R}^n$
  - $R^T = \max_{\hat{x} \in X} \sum_{t=1}^T (\ell^t)^T \hat{x} - (\ell^t)^T x^t$

- **RVU bound** [Syrgkanis 15]
  - **Regret bounded by variations in utility**
  - $R^T \le \alpha + \beta \sum_{t=1}^T \|\ell^t - m^t\|_*^2 - \gamma \sum_{t=2}^T \|x^t - x^{t-1}\|^2$
    - $\|v\|_* = \text{dual norm} = \max_x \frac{\|v^T x\|}{\|x\|}$
      - $\|\cdot\|_1 \to \|\cdot\|_{1*} = \|\cdot\|_\infty$
      - $\|\cdot\|_2 \to \|\cdot\|_{2*} = \|\cdot\|_2$
      - $\|\cdot\|_p \to \|\cdot\|_{p*} = \|\cdot\|_q$ where $\frac{1}{q} + \frac{1}{p} = 1$

- **Accelerated self play**
  - Setting: $\max_{x \in X} \min_{y \in Y} x^T A y$



  - R_x and R_y satisfy RVU bound
  - $\bar{x} = \frac{1}{T} \sum_{t=1}^T x^t$
  - $\bar{y} = \frac{1}{T} \sum_{t=1}^T y^t$
  - $m_x^t = Ay^{t-1}$
  - $m_y^t = -A^T x^{t-1}$
  - Fact. $\gamma(\bar{x}^T, \bar{y}^T) \le \frac{1}{T}(R_x^T + R_y^T)$

    ↑saddle point gap

- $$\begin{aligned}
\gamma(\bar{x}^T, \bar{y}^T) &\le \frac{1}{T}(R_x^T + R_y^T) \\
&\le \frac{1}{T}\Big(2\alpha + \beta \sum_{t=1}^T \|\ell_x^t - m_x^t\|_*^2 - \gamma \sum_{t=2}^T \|x^t - x^{t-1}\|^2 \\
&\qquad + \beta \sum_{t=1}^T \|\ell_y^t - m_y^t\|_*^2 - \gamma \sum_{t=2}^T \|y^t - y^{t-1}\|^2\Big) \\
&\le \frac{1}{T}\Big(2\alpha + \beta \sum_{t=1}^T \|A^T(x^t - x^{t-1})\|_*^2 - \gamma \sum_{t=2}^T \|x^t - x^{t-1}\|^2 \\
&\qquad + \beta \sum_{t=1}^T \|A(y^t - y^{t-1})\|_*^2 - \gamma \sum_{t=2}^T \|y^t - y^{t-1}\|^2\Big)
\end{aligned}$$

  note $\|Mz\|_* \le \|M\|_{op} \|z\|$

  $$\begin{aligned}
&\le \frac{1}{T}\Big[2\alpha + \beta \|A\|_{op}^2 \sum_{t=2}^T \|x^t - x^{t-1}\|^2 - \gamma \sum_{t=2}^T \|x^t - x^{t-1}\|^2 \\
&\qquad + \beta \|A\|_{op}^2 \sum_{t=2}^T \|y^t - y^{t-1}\|^2 - \gamma \sum_{t=2}^T \|y^t - y^{t-1}\|^2 \\
&\qquad + \beta \|A\|_{op}^2 \|y^1\|^2 + \beta \|A\|_{op}^2 \|x^1\|^2\Big) \\
&\le \frac{O(1)}{T}
\end{aligned}$$

Lecture 09/28

Predictive FTRL follow the regularized leader

- $x \in \mathbb{R}^n$ convex, compact set
- $\varphi : X \to \mathbb{R}$ 1-strongly convex
- $\eta > 0$ stepsize

- fn INITIALIZE(): $L^0 \leftarrow 0 \in \mathbb{R}^n$ where at every time T, $L^T = \sum_{t=1}^T \ell^t$
- fn NEXT STRATEGY(m^t):
  return $\operatorname{argmax}_{\hat{x} \in X} (L^{t-1} + m^t)\hat{x} - \frac{1}{\eta}\varphi(\hat{x})$
- fn OBSERVE UTILITY(ℓ^t): $L^t \leftarrow L^{t-1} + \ell^t$

Predictive OMD online mirror descent

- $X \subseteq \mathbb{R}^n$ convex and compact set
- $\varphi : X \to \mathbb{R}$ 1-strongly convex
- $\eta > 0$ stepsize

- fn INITIALIZE(): $z^0 \leftarrow$ any $\hat{z} \in X : \nabla \varphi(\hat{z}) = 0$
- fn NEXT STRATEGY(m^t):
  return $\operatorname{argmax}_{\hat{x} \in X} \left((m^t)^T \hat{x} - \frac{1}{\eta} D_\varphi(\hat{x} \| z^{t-1}) \right)$
- fn OBSERVE UTILITY(ℓ^t):
  $z^t \leftarrow \operatorname{argmax}_{\hat{z} \in X} \left((\ell^t)^T \hat{z} - \frac{1}{\eta} D_\varphi(\hat{z} \| z^{t-1}) \right)$

- $D_\varphi(a \| c) = \varphi(a) - \varphi(c) - \langle \nabla \varphi(c), a - c \rangle$
  - If $\varphi = \frac{1}{2}\|\cdot\|_2^2$
    then
$$\begin{aligned} D_\varphi(a \| c) &= \tfrac{1}{2}\|a\|_2^2 - \tfrac{1}{2}\|c\|_2^2 - c^T(a-c) \\ &= \tfrac{1}{2}\|a\|_2^2 - \tfrac{1}{2}\|c\|_2^2 - c^T A + c^T c \\ &= \tfrac{1}{2}\|a\|_2^2 + \tfrac{1}{2}\|c\|_2^2 - c^T A \\ &= \tfrac{1}{2}\|c - a\|_2^2 \end{aligned}$$

- $$\begin{aligned} &\operatorname{argmax}_{\hat{x} \in X} \; g^T \hat{x} - \tfrac{1}{\eta} D_\varphi(\hat{x} \| c) \\ &= \operatorname{argmax}_{\hat{x} \in X} \; g^T \hat{x} - \tfrac{1}{2\eta}\|\hat{x} - c\|_2^2 \\ &= \operatorname{argmax}_{\hat{x} \in X} \; \eta g^T \hat{x} - \tfrac{1}{2}\|\hat{x}\|_2^2 + c^T \hat{x} \\ &= \operatorname{argmax}_{\hat{x} \in X} \; -\tfrac{1}{2}\|\hat{x} - (\eta g + c)\|_2^2 \\ &= \operatorname{Proj}_X(\eta g + c) \end{aligned}$$

- **Thm.** Let Ω be the range of φ over X
$$\Omega = \max_{x, x' \in X} \varphi(x) - \varphi(x')$$

Then at all times T and for all $\eta > 0$
$$R^T \le \frac{\Omega}{\eta} + \eta \sum_{t=1}^T \|\ell^t - m^t\|_*^2 - \frac{1}{c\eta} \sum_{t=2}^T \|x^t - x^{t-1}\|^2$$

where $c = \begin{cases} 4 & \text{for FTRL} \\ 8 & \text{for OMD} \end{cases}$

and $\|\cdot\|$ is the norm for which φ is 1-strongly convex

Lecture 09/30

- For FTRL, $\text{argmax}_{\hat{x}\in X} \; a^T\hat{x} - \frac{1}{\eta}\varphi(\hat{x}) \;\; = \text{argmax}_{\hat{x}\in X} \; \eta a^T\hat{x} - \varphi(\hat{x}) = \nabla_{\varphi^*}(\eta a)$
- For OMD, $\text{argmax}_{\hat{x}\in X} \; a^T\hat{x} - \frac{1}{\eta}D_\varphi(\hat{x}\|c)$

- $\text{argmax}_{\hat{x}\in X} \; a^T\hat{x} - \frac{1}{\eta}D_\varphi D(\hat{x}\|c)$
  $= \text{argmax}_{\hat{x}\in X} \; a^T\hat{x} - \frac{1}{\eta}\varphi(\hat{x}) + \frac{1}{\eta}\varphi(c) + \frac{1}{\eta}(\nabla\varphi(c))^T(\hat{x}-c)$
  $= \text{argmax}_{\hat{x}\in X} \; a^T\hat{x} - \frac{1}{\eta}\varphi(\hat{x}) + \frac{1}{\eta}(\nabla\varphi(c))^T\hat{x}$
  $= \text{argmax}_{\hat{x}\in X} \; (a + \frac{1}{\eta}(\nabla\varphi(c))^T)\hat{x} - \frac{1}{\eta}\varphi(\hat{x})$
  $= \nabla_{\varphi^*}(a + \frac{1}{\eta}(\nabla\varphi(c)))$

- Def. $\varphi : X \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is "nice"
  if the following quantities can be computed in $O(n)$ time
  ① $\nabla\varphi(c) \;\; \forall c \in X$
  ② $\nabla_{\varphi^*}(a) = \text{argmax}_{\hat{x}\in X} \; a^T x - \varphi(\hat{x})$

- For $x \in \Delta^n$ a "nice" regularizer is, among others,
  $\varphi(x) = \Sigma x_i \log x_i$
  $\frac{\partial}{\partial x_i}\varphi(x) = 1 + \log x_i$

  ②: $\text{argmax}_{x\in\mathbb{R}^n} \; \Sigma_{i=1}^n a_i x_i - \Sigma_{i=1}^n x_i \log x_i, \;\; \Sigma_{i=1}^n x_i = 1, \;\; x_i \geq 0 \;\forall i$
  By Lagrange multiplier theorem,
  $L(x,\alpha) = \Sigma a_i x_i - \Sigma x_i \log x_i - \alpha(\Sigma x_i - 1)$
  $\nabla_x L(x,\alpha) = 0$ when
  $\frac{\partial}{\partial x_i} L(x_i,\alpha) = a_i - 1 - \log x_i - \alpha = 0$
  $\Rightarrow \log x_i = a_i - 1 - \alpha$
  $\Rightarrow x_i = \exp(a_i - 1 - \alpha)$
  $\Rightarrow x_i^* = \frac{\exp(a_i)}{\Sigma \exp(a_i)}$
  So $\nabla_{\varphi^*}(a) = \text{softmax}(a)$
  $= \left(\frac{\exp(a_i)}{\Sigma\exp(a_i)}\right)_i$

- What about $\varphi(x) = \frac{1}{2}\|x\|_2^2$?
  $\nabla\varphi(x) = x$
  But then $\text{argmax}_{\hat{x}\in\Delta^n} \; a^T x - \frac{1}{2}\|x\|_2^2$ is "hard" to solve

- What about sequence form polytopes
  $$\varphi(x) = \sum_{j\in J}\sum_{a\in A_j} w_{ja}\cdot x[ja]\log x[ja]$$
  where w_{ja}'s are chosen recursively according to
  $$w_{ja} = \gamma_j - \sum_{\substack{j'\in J \\ p_{j'}=ja}} \gamma_{j'}$$
  $$\gamma_j = 1 + \max_{a\in A_j}\left\{\sum_{\substack{j'\in J \\ p_{j'}=ja}} \gamma_{j'}\right\} \geq 1$$

  - $\varphi(x)$ is nice and 1-strongly convex wrt ℓ_2

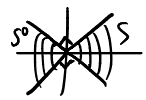
Lecture 09/30 cont.

- **Predictive Blackwell game**
  - Before X plays, they receive prediction $v^t \in \mathbb{R}^d$ of the next utility
  - Consider only S cone

- Take $\mathcal{R}$ a regret minimizer over set $S^\circ = \{y \in \mathbb{R}^n : \langle y, x\rangle \le 0 \ \forall x \in S\} \subset \mathbb{B}_2$

  fn NextStrategy(v^t)
  $\theta^t \leftarrow \mathcal{R}$.NextStrategy($v^t$)
  output a forcing action for $H^t = \{x \in \mathbb{R}^n : \langle x, \theta^t\rangle \le 0\}$

  fn ReceivePayoff($u(x^t, y^t)$)
  $\mathcal{R}$.ObserveUtility($u(x^t, y^t)$)



- Ignoring v^t for now,

$$\min_{\hat{s} \in S} \|\hat{s} - x\|_2 = \max_{\hat{s} \in S^\circ \cap \mathbb{B}_2} \frac{\langle \hat{s}, x\rangle}{\|\hat{s}\|} = \max_{\hat{s} \in S^\circ \cap \mathbb{B}_2} \langle \hat{s}, x\rangle$$

$$\begin{aligned} d_s\left(\tfrac{1}{T}\textstyle\sum u(x^t, y^t)\right) &= \min_{\hat{s} \in S} \left\|\hat{s} - \tfrac{1}{T}\textstyle\sum u(x^t, y^t)\right\| \\ &= \max_{\hat{s} \in S^\circ} \frac{\langle \hat{s}, \frac{1}{T}\sum u(x^t, y^t)\rangle}{\|\hat{s}\|} \\ &= \max_{\hat{s} \in S^\circ} \frac{1}{\|\hat{s}\|} \frac{1}{T} \textstyle\sum_{t=1}^T \langle \hat{s}, \ell^t\rangle \\ &= \max_{\hat{s} \in S^\circ} \frac{1}{T\|\hat{s}\|} \left[R^T + \underbrace{\textstyle\sum_{t=1}^T \langle \theta^t, \ell^t\rangle}_{\le 0 \ \forall t}\right] \end{aligned}$$

$$R^T = \max_{\hat{s} \in S^\circ} \sum_{t=1}^T \langle \hat{s}, \ell^t\rangle - \sum_{t=1}^T \langle \theta^t, \ell^t\rangle$$

By $\cap \mathbb{B}_2$, eliminate $\|\hat{s}\|$ in denominator

$\ldots \le \frac{1}{T} R^T$

Lecture 10/05

- A halfspace H that contains the target set S is forceable if $\exists x^* \in X \ \forall y \in Y \ u(x^*, y) \in H$
  We call x^* a forcing action for H

$$u(x, \ell) = \ell - \langle x, \ell\rangle \cdot 1$$

$$\frac{R^T}{T} \le \min_{\hat{s} \in \mathbb{R}^n_{\le 0}} \left\|\hat{s} - \tfrac{1}{T}\textstyle\sum u(\ell^t, x^t)\right\|_2$$

- Regret Minimizer for $\mathbb{R}^n_{\ge 0}$, e.g. OMD, FTRL $\longrightarrow$ Abernethy's Algorithm

Abernethy's Algorithm $\downarrow$ Blackwell Game $\Gamma = (\Delta^n, \mathbb{R}^n, u(\cdot,\cdot), \mathbb{R}^n_{\le 0})$

Blackwell's Algorithms $\longrightarrow$ Blackwell Game $\Gamma = (\Delta^n, \mathbb{R}^n, u(\cdot,\cdot), \mathbb{R}^n_{\le 0})$ $\longrightarrow$ Regret Minimizer on Δ^n (regret matching)

Different paths to RM/RM+

Regret matching
1. Blackwell alg
2. FTRL with $\frac{1}{2}\|\cdot\|_2^2$

Predictive Regret Matching
1. PFTRL with $\frac{1}{2}\|\cdot\|_2^2$

Regret Matching Plus
1. OMD with $\frac{1}{2}\|\cdot\|_2^2$

Predictive Regret Matching Plus $\leftarrow$ SOTA for non-poker
1. POMD with $\frac{1}{2}\|\cdot\|_2^2$

(poker is odd in general, e.g. there are really bad actions)

predictive generally better except in poker where x same

Lecture 10/05 cont.

- **(P) FTRL**

  fn NextStrategy(m^t)

  return $\underset{\hat{x}\in X}{\arg\max}\left\{\left(L^{t-1}+m^t\right)^T\hat{x}-\frac{1}{\eta}\varphi(\hat{x})\right\}$

  fn ObserveUtility(ℓ^t)

  $L^t \leftarrow L^{t-1}+\ell^t$

- **(P) OMD**

  fn NextStrategy(m^t)

  return $\underset{\hat{x}\in X}{\arg\max}\left\{(m^t)^T\hat{x}-\frac{1}{\eta}D_\varphi(\hat{x}\,\|\,z^{t-1})\right\}$

  fn ObserveUtility(ℓ^t)

  $z^t \leftarrow \underset{\hat{z}\in X}{\arg\max}\left\{(\ell^t)^T\hat{z}-\frac{1}{\eta}D_\varphi(\hat{x}\,\|\,z^{t-1})\right\}$

- **Abernethy's algorithm**
  - fn NextStrategy()

    $\theta^t \leftarrow \mathcal{R}$.NextStrategy()

    return forcing action x^t for $H^t=\{x: \langle x,\theta^t\rangle \le 0\}$
  - fn ObserveBlackwellPayoff ($u(x^t,y^t)$)

    $\mathcal{R}$.ObserveUtility($u(x^t,y^t)$)

- For today, $\mathcal{R}$ regret minimizer for S° instead of $S^\circ\cap\mathbb{B}_2$ and it is either FTRL or OMD no matter η or φ

- Fact. Define $R^T(\hat{x})=\sum(\ell^t)^T\hat{x}-\sum(\ell^t)^T x^t$

  For POMD and PFTRL: $\forall \hat{x}\in X \quad R^T(\hat{x})\le \frac{\varphi(\hat{x})}{\eta}+\eta\sum\|\ell^t-m^t\|_*^2$

- $$\begin{aligned}\min_{\hat{s}\in S}\left\|\hat{s}-\frac{1}{T}\sum u(x^t,y^t)\right\|_2 &= \max_{\hat{s}\in S^\circ\cap\mathbb{B}_2}\left\langle \frac{1}{T}\sum u(x^t,y^t),\ \hat{s}\right\rangle\\ &= \max_{\hat{s}\in S^\circ\cap\mathbb{B}_2}\left\langle \frac{1}{T}\sum \ell^t,\ \hat{s}\right\rangle\\ &= \frac{1}{T}\left[\max_{\hat{s}\in S^\circ\cap\mathbb{B}_2}\sum\langle\ell^t,\hat{s}\rangle-\sum\langle\ell^t,\theta^t\rangle\right]+\underbrace{\frac{1}{T}\sum\langle\ell^t,\theta^t\rangle}_{\le 0}\\ &= \frac{1}{T}\max_{\hat{s}\in S^\circ\cap\mathbb{B}_2}R^T(\hat{s})\\ &\to 0\end{aligned}$$

Lecture 10/05 cont.

- Take Blackwell's game Γ.
  Use Abernethy's alg to solve Γ,
  $\mathcal{R} = \text{FTRL}$, $\varphi = \frac{1}{2}\|\cdot\|_2^2$, domain $\mathbb{R}^n_{\geq 0}$.

fn NextStrategy()

$$\theta^t \leftarrow \underset{\hat{x} \in \mathbb{R}^n_{\geq 0}}{\text{argmax}} \left\{ \eta (L^{t-1})^T \hat{x} - \frac{1}{2\eta} \|\hat{x}\|_2^2 \right\}$$

$$= \underset{\hat{x} \in \mathbb{R}^n_{\geq 0}}{\text{argmax}} \left\{ -\frac{1}{2} \|\hat{x} - L^{t-1}\eta\|_2^2 \right\}$$

$$= \underset{\hat{x} \in \mathbb{R}^n_{\geq 0}}{\text{argmin}} \left\{ \|\hat{x} - \eta L^{t-1}\|_2^2 \right\}$$

$$= \text{proj}_{\mathbb{R}^n_{\geq 0}} (\eta L^{t-1})$$

$$= [\eta L^{t-1}]^+ \in \mathbb{R}^n_{\geq 0}$$

$$x^t \leftarrow \frac{\theta^t}{\mathbb{1}^T \theta^t}$$

return x^t

- fn ObserveBlackwellPayoff$(u(x^t, y^t) \in \mathbb{R}^n)$

$$L^t \leftarrow L^{t-1} + u(x^t, y^t)$$

- Thm. predictive regret matching guarantees regret

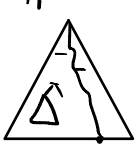
$$R^T \leq \max_{\hat{s} \in \mathbb{R}^n_{\geq 0} \cap \mathcal{B}_2} \frac{\varphi(\hat{s})}{\eta} + \eta \sum \|u(x^t, y^t) - v^t\|_*^2 \quad \forall \eta > 0$$

$$\leq \frac{1}{2\eta} + \eta \sum \|u(x^t, y^t) - v^t\|_*^2$$

$$\leq \sqrt{2 \sum \|u(x^t, y^t) - v^t\|_*^2}$$

Lecture 10/07

- **Monte-Carlo CFR**: standard sublinear method
  - Suppose only one leaf nonzero util



Then all other paths have util 0

- $\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$ "Coin with 3 faces" (uniform)
  values 1,2,3
  flip $1 \to \begin{pmatrix} 3v_1 \\ 0 \\ 0 \end{pmatrix}$, $2 \to \begin{pmatrix} 0 \\ 3v_2 \\ 0 \end{pmatrix}$, $3 \to \begin{pmatrix} 0 \\ 0 \\ 3v_3 \end{pmatrix}$

  "Importance sampling"
  can generalize to p_1, p_2, p_3 instead of uniform
  just $\frac{1}{p_i}$ instead of 3

Lecture 10/07 cont.

- An unbiased estimator for Ay^t can be computed by
  1. Pick unbiased estimator $\tilde{y}^t$ for y^t
  2. Compute $A\tilde{y}^t$

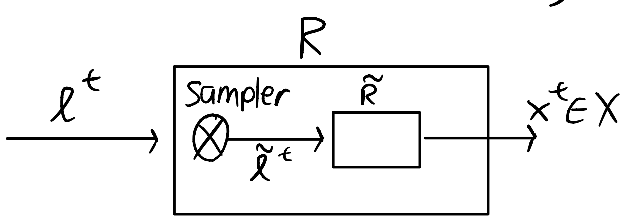
  Note: $\tilde{y}^t$ can be very sparse

- $u(x,y) = x^T A y$
  $= \sum_{z \text{ terminal}} u(z) \left(\prod \text{all actions for pl1 on path to } z\right)\left(\prod \text{all actions for pl2 on path to } z\right)\left(\prod \text{all nature actions on path to } z\right)$

  but recall x, y already in sequence form
  $= \sum_{z \text{ terminal}} u(z)\ x[\sigma_1(z)]\ y[\sigma_2(z)]\ P_{chance}(z)$

  we will show an unbiased estimator is
  1. Pick z with distribution $P_{chance}(z)\ y[\sigma_2(z)]\ \tilde{x}[\sigma_1(z)]$
  2. Consider vector $\frac{u(z)}{\tilde{x}[\sigma_1(z)]} e_{\sigma_1(z)} \leftarrow$ basis vector, | where $\sigma_1(z)$ is $\in \mathbb{R}^{|\Sigma|}$

  This is known as outcome sampling

- 

- The degradation in regret due to sampling
  $|R^T - \tilde{R}^T|$ is upper bounded by $\sqrt{2T \log \frac{1}{\delta}}\,(M + \tilde{M})$ with probability $\geq 1-\delta \quad \forall \delta \in (0,1)$

  maximum range of ℓ^t ; maximum range of $\tilde{\ell}^t$

  - Proof a little involved,
    Azuma Hoeffding on martingales
  - But overall, sampling does not hurt much
  - In practice: for games where CFR can handle it, just using CFR is faster
    but for huge games, MCCFR preferred for sublinear

- This concludes CFR, online learning. Now, offline learning.

- Note. First order offline optimization has no theoretical/practical benefits over online.

- Offline optimization
  - (A) First-order saddle point solvers
  - (B) First-order gradient descent based
  - (C) Methods bared on the linear programming formulation

- (A) $\max_{x \in X} \min_{y \in Y} x^T A y$
  - Excessive gap technique [Nesterov]
  - Mirror prox [Nemirovski]

  } $O(\frac{1}{T})$ convergence rate to saddle point

  POMD is more powerful nowadays

Lecture 10/07 cont.

- (B) $\max_{x\in X} \min_{y\in Y} x^T A y = \max_{x \in X} g(x)$

  where $g(x) = \min_{y\in Y} x^T A y$ is concave

  - Gradient ascent
  - ADAM (?)

  - $\nabla g = Ay^*$ where y^* solution to $\min_{y\in Y} x^T A y$

- (C) Linear programming

$$\max_{x\in Q_1} \min_{y \in Q_2} x^T A y$$

$$= \max_x \min_y x^T A y$$

$$F_2 y = f_2, \quad y \geq 0$$

$$F_1 x = f_1, \quad x \geq 0$$

note: sequence form polytope $Q \subseteq \mathbb{R}^{|\Sigma|}$ such that

$\forall j \; \sum_{a \in A_j} x[ja] = x[p_j], \quad x \geq 0$

i.e. $Q = \left\{ x : \begin{matrix} Fx = f \\ x \geq 0 \end{matrix} \right\}$

Not quite a LP but min is LP

$$= \max_x \max_v f_2^T v$$

$$F_2^T v \leq A^T x$$

$$F_1 x = f_1, \quad x \geq 0$$

$$= \max_{x,v} f_2^T v$$

$$F_2^T v \leq A^T x$$

$$F_1 x = f_1$$

$$x \geq 0$$

- Solvers
  - Simplex
  - Interior point/Barrier
  - Ellipsoid

  $\Big\}$ guarantee error $\leq \varepsilon$ in time $O(\log \frac{1}{\varepsilon})$

- Payoff matrix sparsification

$$A = UM^{-1}V^T + \hat{A}$$

size of sparsification $= \text{nnz } U + \text{nnz } M + \text{nnz } V + \text{nnz } \hat{A}$

$$A^T x = (\hat{A}^T + UM^{-1}V^T)x$$

$$= \hat{A}^T x + VM^{-T}U^T x$$

$$= \hat{A}^T x + VM^{-T} w$$

$$= \hat{A} x + Vz$$

$$A^T x = \hat{A} x + Vz$$

$$M^{-T} z = w$$

$$U^T x = w$$

- So sparsified

$$\max_{x,v} f_2^T v$$

$$F_2^T v \leq \hat{A}^T x + Vz$$

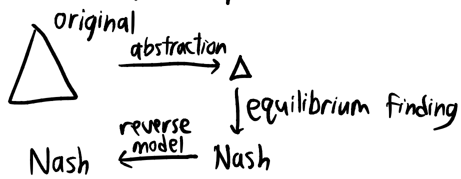
$$M^{-T} z = V^T x$$

$$F_1 x = f_1$$

$$x \geq 0$$

Lecture 10/12 SOTA Practical Game Abstraction

- Automated game abstraction [Gilpin & Sandholm EC-06, ACM07]
  - Used in all competitive Texas Holdem today



- Lossless game abstraction
  - Information filters
    - Can make game smaller by filtering the information a player receives
  - Signal tree
    - Each edge corresponds to the revelation of some signal by nature to at least one player
    - Abstraction algorithm operates on it
  - Isomorphic relation
    - Strategic symmetry between nodes
    - Recursively, two leaves in signal tree are isomorphic if for each action history in the game, the payoffs are the same.
    - Recursively, two internal nodes in signal tree are isomorphic if they are siblings and their children are isomorphic
      - Need custom perfect matching algorithm for isomorphism children matching

- Game Shrink
  - Bottom up pass: DP to mark isomorphic pairs of nodes in signal tree
  - Top down pass: starting from top of signal tree, perform transformation (merge isomorphic pairs) wherever possible
  - Thm. To do all transforms,
    - $\tilde{O}(n^2)$, $n = \#$ nodes in signal tree
    - Usually highly sublinear in game size
  - Solved AI challenge problem [Shi & Littman 01] Rhode Island Holdem
    - 3.1 billion nodes in game tree
    - No abstraction, LP has 91224226 rows and cols ⇒ unsolvable
    - After abstraction, LP has 1237238 rows and cols (50428638 nonzeros)
    - Abstraction runs in 1 second
    - CPLEX barrier took 8 days, 25 GB of RAM back then [2006/2007]
    - Exact Nash equilibrium

- Lossy game abstraction

- Texas holdem poker
  - 2-player limit 10^{18} nodes
  - 2-player no-limit 10^{165} nodes
  - Lossless abstraction still too big, need lossy abstraction
    - Usually 2 orders of magnitude, $10^{165} \to 10^{163}$ still eh

Lecture 10/12 cont.

- GameShrink can abstract more by not requiring a perfect matching ⇒ lossy
  - $|wins_{node1} - wins_{node2}| + |losses_{node1} - losses_{node2}| < k$
  - Greedy ⇒ lopsided abstractions

- Abstraction in each player's card tree separately [Gilpin & Sandholm AAMAS-07]
  - Clustering + Integer programming
    - For every betting round i, tell alg how many buckets k_i it is allowed to generate
    - First betting round ⇒ k_1-means clustering to bucket nodes
    - Later rounds ⇒ run IP to determine how many children each parent should be allowed to have so that total # of children doesn't exceed k_i
      - Value determined with k-means clustering for all k (up to 30) on each parent before IP

- Potential aware abstraction
  - All prior algs had probability of winning as similarity metric
    - Assumes no more betting
  - Doesn't capture potential
  - Potential is multidimensional, not positive or negative
  - Bottom up pass for round 1
    - L_1 norm on transition probability vector to (oracle) next round's buckets
  - Last round, no more potential ⇒ probability of winning assuming rollout as similarity metric
  - See slides for details

- Important ideas for practical lossy abstraction 2007-2013
  - Integer programming
  - Potential-aware
  - Imperfect recall

- SOTA: Potential Aware Imperfect Recall Abstraction with Earth Mover distance in imperfect information games
  - Expected hand strength = ehs = equity is $P(\text{winning}) + \frac{1}{2}P(\text{tying})$
    - against uniform random draw of private cards for opponent
    - assuming uniform random rollout of remaining public cards
    - Used to cluster hands
    - But doesn't account for hand strength
  - Earth mover distance, distance metric for histograms
    - min cost turning one pile into another
    - cost = amount of dirt moved × distance moved
    - Linear time in 1D but challenging to compute in higher dimensions

- Potential-aware abstraction considers all future rounds, not just final round

Lecture 10/19 Action Abstraction

- See slides

Lecture 10/21 Libratus - SOTA 2-player no-limit Texas hold'em

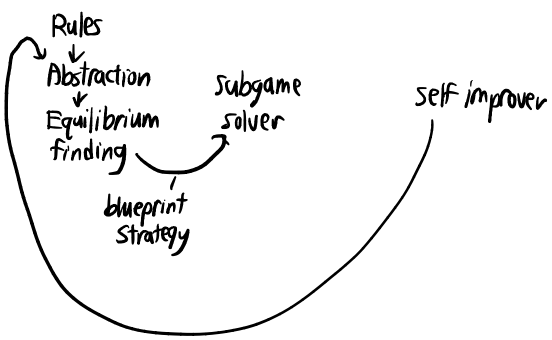
- AlphaGo extends to perfect information games only
- In perfect info games, subgames can be solved with info in subgame only
  - Not true in imperfect info games!

- Heads up (2 player) No Limit Texas Hold'em
  - 10^{161} situations
  - Main benchmark/challenge problem for imperfect info game
  - No AI beat humans prior to Libratus

- Libratus (rematch after prior AI lost)
  - 120k hands over 20 days, 4 players
  - Jan 2017
  - \$200k/pros based on performance – not NSF, private raise. \$20k base, bot nothing, top 3 by perf
  - Weren't confident that Libratus would win
  - Poker players are intense, ready to wake up/stop showering to play
  - Conservative experiment design
    - Slides for details
    - On avg human 21s per hand, AI 13s per hand

- AI vs ML
  - No data needed
  - Doesn't assume opponent will behave the same way
  - Not exploitable

- Libratus
  - Pgh Bridges supercomputer



  - Abstraction
    - Same algorithm as Tartanian8
    - But much finer abstraction
    - Abstracted player bet sizes, including radical bet sizes which were used

Lecture 10/21 cont.

- Equilibrium finding
  - Improved MCCFR
- System setup, see slides

- Subgame solver
  - NIPS17 best paper
  - 2015 unsafe subgame solving
    - No theoretical guarantees
    - Does well in practice for some domains
    - Assume other player plays according to blueprint strategy
  - 2014 Re-solve refinement
    - P1 picks between entering subgame or taking EV blueprint of subgame
  - 2016 Max margin refinement
    - $\text{Margin}_H = EV[\text{Alt}_H] - EV[\text{Enter}_H]$
    - Maximize minimum margin
  - 2017 Reach max margin refinement
    - Mistake by opponent is a gift
    - Split gifts among subgame by probability subgame reached
    - Can substitute lower bound estimates on the gift
  - Nested subgame solving
    - Solve subtree in realtime for off tree action taken

Lecture 10/26

- Self-improver
  - Intuition: use opponents actions as hints for where we are weak

- See slides for more on Libratus

- Depth-limited subgame solving and Pluribus, SOTA for multiplayer no limit Texas holdem
  - "Solve a middle game"

- Depth-limited search for imperfect information game

Lecture 11/02

- DeepMind SC2
  - Great talk
  - What is an action's representation?

Lecture 11/04

- Certificates in extensive form games
- Deep RL [Alpha …]
  - Good practical perf
  - No exploitability bounds
- Bandit regret minimization [Farina 20]

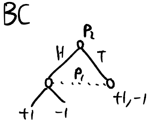
- Certificates
  - Compute Nash by incrementally expanding game tree

- Pseudogame
  - Game w/o known utils on all terminal nodes

- Small certificates
  - Small = $O(N^c)$, $c<1$, $N = \#$ nodes in entire game

- See slides…

- Matching pennies, C terminal nodes
  - BC

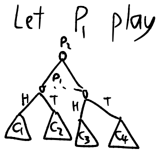


$C<4$: 0 rounds lost under optimistic best response

$C=4$: $\frac{1}{2}$ rounds lost

  - Inductive case

    Let P_1 play $H = x$



$C = C_1 + C_2 + C_3 + C_4$

$x + x\log_{16} C_1 + (1-x)\log_{16} C_2 \leftarrow u_{P_1 H}$

$(1-x) + (1-x)\log_{16} C_3 + x\log_{16} C_4 \leftarrow u_{P_1 T}$

want $\frac{C_1}{C_1+C_2} = x$, $\frac{C_4}{C_4+C_3} = x$, $p = \frac{C_1+C_2}{C_1+C_2+C_3+C_4}$

$\log_{16} C + \underbrace{\min(P_1 H, P_1 T)}_{\leq 0}$, so P_2 always wins that much

  - Oracle mode

Lecture 11/09

- Slides
- Exp4
  - Next page

Lecture 11/09 cont.

- Recall MWU, k experts
  - Initialize $\forall j \in [k]$, $P_j^0 = \frac{1}{k}$ and $R_j^0 = 0$
  - for t from 1 to ∞:

    select expert j according to P_j^{t-1}

    receive reward r_j^t for each expert j

    $\forall j,\ R_j^t \leftarrow R_j^{t-1} + r_j^t$

    $P_j \leftarrow \frac{e^{\eta R_j^t}}{\sum_i e^{\eta R_i^t}}$

missing fancy guarantees

Exp3

Receive reward r_j^t for expert j only

$R_j^t \leftarrow R_j^{t-1} + \frac{r_j^t}{P_j^{t-1}}$

Exp 4

Experts are strategies, e_j^t is the action recommended by expert j at time t

k experts, n actions

Initialize $\forall j \in [k]$, $P_j^0 = \frac{1}{k}$, $R_j^0 = 0$

For $t=1$ to ∞:

select an action i according to $\left(\sum_{j: e_j^t = i} P_j^{t-1}\right) := \mathcal{P}_i^{t-1}$

receive reward r_i^t for action i

update $R_j^t \leftarrow \begin{cases} R_j^{t-1} + \frac{r_i^t}{\mathcal{P}_i^{t-1}} & \text{if } j \text{ recommended } i \\ R_j^{t-1} & \text{otherwise} \end{cases}$

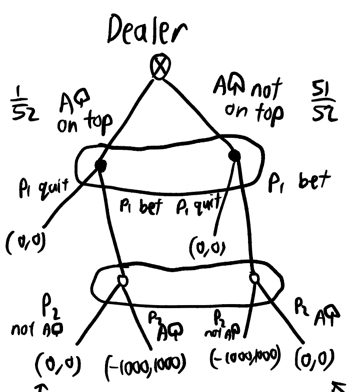
$P_j^t \leftarrow \frac{e^{\eta R_j^t}}{\sum_i e^{\eta R_i^t}}$

Lecture 11/11

- **Equilibrium refinements**
  - Traditionally from economics
  - Less/no opponent modeling

- Intuitively, **Nash equi optimizes for strong opponent**
  - Doesn't focus on parts of game tree where opponent wouldn't go

- **Guess the ace**



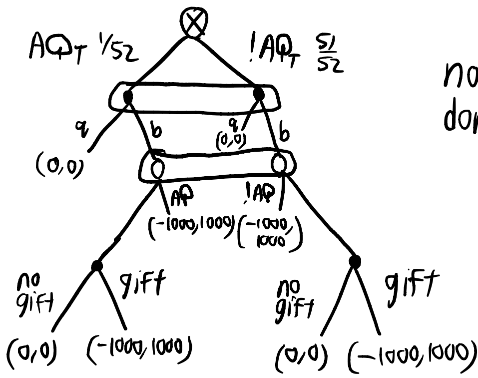
"sequentially rational"

"sequentially irrational"

- Nash eq not equally good when players make mistakes

Lecture 11/11 cont.

- Guess the ace with gifts



not just removing dominated actions!

- Trembling-hand refinement
  - Introduce $\varepsilon > 0$
  - For any $\varepsilon > 0$, imagine computing a Nash eq. in the game s.t. the fact that every action is selected with lower bound probability $f(\varepsilon)$
  - Return a limit point of those Nash eq. as $\varepsilon \to 0^+$

  } conceptual framework

- Extensive form perfect equilibrium [Setten 75] Nobel prize
  - $f(\varepsilon) = \varepsilon$
  - $x[ja] > \varepsilon\, x[p_j]$ if $p_j \neq \emptyset$
  - $x[ja] \geq \varepsilon$ if $p_j = \emptyset$

  $\} \to M_1(\varepsilon)\, x \geq m_1(\varepsilon)$

  - $\max\limits_x \min\limits_y x^T U y$

    $F_2 y = f_2$ (seq form constraints)

    $M_2(\varepsilon)\, y \geq m_2(\varepsilon)$ (trembling constraints)

    $F_1 x = f_1$

    $M_1(\varepsilon)\, x \geq m_1(\varepsilon)$

- Quasi-perfect equilibrium [van Damme 84]
  - The lower bound probability constraints (trembling constraints) are in sequence form
    - The probability of every sequence $\sigma \geq \varepsilon^{|\sigma|}$
    - $x[ja] \geq \varepsilon^{depth(ja)}$ $\Rightarrow$ $x[ja] \geq \ell_1(\varepsilon)$

  - $\max\limits_x \min\limits_y x^T U y$

    $F_2 y = f_2$ (seq form constraints)

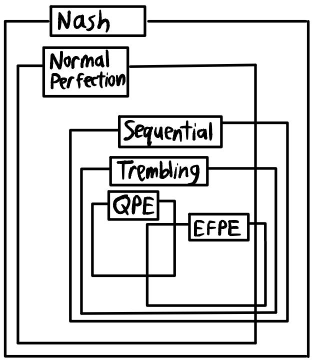
    $y \geq \ell_2(\varepsilon)$ (trembling constraints)

    $F_1 x = f_1$

    $x \geq \ell_1(\varepsilon)$

Lecture 11/11 cont.

- Relationships



Computational complexity

Solution concept	General sum	Zero sum
Nash eq	PPAD-complete [Daskalakis 2009] [Chan & Rong]	FP [Romanovski 62] [von Stengel 96]
QPE	PPAD-complete [Miltersen & Sørensen 2006]	FP ['']
EFPE	PPAD-complete [Farina & Gatti 2017]	FP ['']

- **Trembling LPs** $P(\varepsilon)$

$$\begin{aligned} \max_x \quad & c(\varepsilon)^T x \\ \text{st} \quad & A(\varepsilon)x = b(\varepsilon) \\ & x \geq 0 \end{aligned}$$

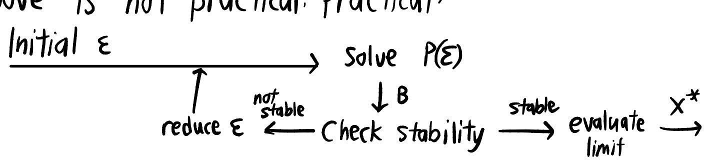
A, b, c "only" depend polynomially in ε

- Goal: compute a limit point of optimal solutions to $P(\varepsilon)$ as $\varepsilon \to 0^+$

- **Stable basis**
  - The LP basis B is stable if there exists $\bar{\varepsilon} > 0$ such that B is optimal for $P(\varepsilon)$ $\forall\, 0 < \varepsilon \leq \bar{\varepsilon}$

- **Negligible positive perturbation** NPP
  - $\varepsilon^* > 0$ st $\forall\, 0 < \bar{\varepsilon} < \varepsilon^*$ any optimal basis for the numerical LP $P(\bar{\varepsilon})$ is stable

- Thm. A NPP ε^* exists and it can be computed in polytime in the input size
- Above is not practical. Practical:



Lecture 11/16

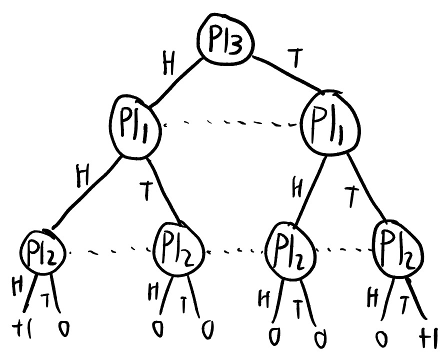
- 3 different notions of "equilibria"
  - Free communication → Nash equilibrium for the "meta player"
  - No communication ever → team maxmin equilibrium TME, favored in RL
  - No communication during game but players can discuss common tactics before playing
    - ↳ TMECor TME with correlation device
    - ↳ convex unlike TME

- Today: only discussing 0 sum

- Communication

requires common signal and tactics → TMECor; able to freely communicate → Nash

	TME	TMECor	Nash
Convex?	✗	✓	✓
Bilinear saddlepoint? $\max_\alpha \min_\beta \alpha^T M \beta$	✗	✓	✓
Is set of strategies a low dimensional (poly $\lvert A \rvert$) polytope?	n/a	✓	✓
Min max thm?	✗	✓	✓
Complexity?	hard, ??	hard	poly
Team utility	low	higher	highest



Pl1, Pl2 team, Pl3
$\uparrow x \quad \uparrow y \quad \uparrow z$

TME team: $\max_{x,y} \min_z \{x_H y_H z_H + x_T y_T z_T\}$
st $x_H + x_T = 1$
$y_H + y_T = 1$
$z_H + z_T = 1$
$x, y, z \geq 0$
→ solves to $\frac{1}{4}$

- $= \max_{x,y} \min \{x_H y_H, x_T y_T\}$

wlog $x_H y_H \leq x_T y_T$
$\Leftrightarrow x_H y_H \leq (1-x_H)(1-x_T)$
$\Leftrightarrow x_H + y_H \leq 1$

$\max_{x,y} \; x_H y_H$
st $x_H + x_T = 1$, $y_H + y_T = 1$ } irrelevant
$x, y \geq 0$
$x_H + y_H \leq 1$

→ $\max_x \; x_H(1-x_H)$
since last constraint tight
$0 \leq x_H \leq 1 \Rightarrow x_H = \frac{1}{4}$

But TME opponent
$\min_z \max_{x,y} (x_H y_H z_H, x_T y_T z_T)$
$\geq \min_z (z_H, z_T) = \frac{1}{2}$

so value maxmin $\neq$ value minmax

Lecture 11/16 cont.

- Recall Π_i = deterministic strategies
- A TMECor is a distribution $H \in \Delta(\Pi_1 \times \Pi_2)$

$$\max_{H \in \Delta(\Pi_1 \times \Pi_2)} \min_{z \in Q_s} \sum_{\substack{(\pi_1, \pi_2) \\ \cap \quad \cap \\ \Pi_1 \quad \Pi_2}} H(\pi_1, \pi_2) \Big(\sum_{\substack{w \in W \\ \uparrow \\ \text{terminal states}}} u_w \cdot \pi_1[\sigma_1(w)] \cdot \pi_2[\sigma_2(w)] \cdot z[\sigma_z(w)] \cdot c[\sigma_c(w)] \Big)$$

$$= \max_H \min_z \sum_{w \in W} \underbrace{\Big(\sum_{(\pi_1, \pi_2)} H(\pi_1, \pi_2) \cdot \pi_1[\sigma_1(w)] \cdot \pi_2[\sigma_2(w)] \Big)}_{:= \gamma[z] \text{ change of var}} u_w \cdot z[\sigma_z(w)] \cdot c[\sigma_c(w)]$$

linear in H!
linear in z!
But the polytope, exponentially big simplex

Let $f: \Delta(\Pi_1 \times \Pi_2) \to \mathbb{R}^{|W|}$

$$H \to \Big(\sum_{(\pi_1, \pi_2)} H(\pi_1, \pi_2) \cdot \pi_1[\sigma_1(w)] \cdot \pi_2[\sigma_2(w)] \Big)_{w \in W}$$

Then above

$$= \max_{\substack{\gamma \in \Omega \\ \gamma \subseteq \mathbb{R}^{|W|}}} \min_z \sum_{w \in W} \gamma[w] \cdot u_w \cdot z[\sigma_z(w)] \cdot c[\sigma_c(w)]$$

where $\Omega = \text{Image}(f)$

Some day I'll have the time to improve these notes...