CSCE 631 — Algorithmic Game Theory

Lecture 9: Regret Minimization III

Alan Kuhnle

Summer 2026

Texas A&M University

Today's Plan

Regret Minimization in Zero-Sum Games

Blackwell Approachability

Regret Matching+ (RM+)

Comparison of Regret Minimization Approaches

Connection to CFR (Preview)

Bandit Feedback vs. Full Information

Summary

Regret Minimization in Zero-Sum Games

The Special Case of Zero-Sum Games

Recall: in general games, no-regret dynamics converge to CCE / CE but **not** to NE.

Zero-sum games are different.

- Two players with payoff matrix A: row player gets $x^\top Ay$, column player gets $-x^\top Ay$.
- The minimax theorem guarantees $v^* = \max_x \min_y x^\top Ay = \min_y \max_x x^\top Ay$.
- NE = minimax strategies.

Key Insight

In zero-sum games, minimizing external regret suffices for convergence to Nash equilibrium via average strategies.

Setup: Repeated Zero-Sum Game

- Row player uses no-regret algorithm producing strategies $x^1, \ldots, x^T$.
- Column player uses no-regret algorithm producing strategies $y^1, \ldots, y^T$.
- At round t: row player's loss is $-u_1 = -x^{t\top} A y^t$;
  column player's loss is $-u_2 = x^{t\top} A y^t$.

Define average strategies:

$$\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x^t \in \Delta_m, \qquad \bar{y} = \frac{1}{T} \sum_{t=1}^{T} y^t \in \Delta_n$$

Goal: Show $(\bar{x}, \bar{y})$ is an approximate NE.

The Main Theorem

Theorem

If both players in a zero-sum game use no-regret algorithms with regret bounds R_T^1 and R_T^2, then the average strategy pair $(\bar{x}, \bar{y})$ is an ε-Nash equilibrium with

$$\varepsilon = \frac{R_T^1 + R_T^2}{T}$$

Using MWU for both players ($R_T \leq 2\sqrt{T \ln n}$):

$$\varepsilon \leq \frac{4\sqrt{T \ln n}}{T} = 4\sqrt{\frac{\ln n}{T}}$$

For an ε-NE: need $T = O(\ln n / \varepsilon^2)$ rounds.

Proof (1/2)

Row player's regret bound:

$$\frac{1}{T}\sum_{t=1}^{T} x^{t\top} A y^t \;\geq\; \max_{x\in\Delta_m} \frac{1}{T}\sum_{t=1}^{T} x^{\top} A y^t - \frac{R_T^1}{T}$$

The RHS equals:

$$\max_{x\in\Delta_m} x^{\top} A\bar{y} - \frac{R_T^1}{T}$$

Column player's regret bound (column player minimizes $x^{\top} A y$):

$$\frac{1}{T}\sum_{t=1}^{T} x^{t\top} A y^t \;\leq\; \min_{y\in\Delta_n} \frac{1}{T}\sum_{t=1}^{T} x^{t\top} A y + \frac{R_T^2}{T} \;=\; \min_{y\in\Delta_n} \bar{x}^{\top} A y + \frac{R_T^2}{T}$$

Proof (2/2)

Combining both bounds:

$$\max_{x \in \Delta_m} x^\top A\bar{y} - \frac{R_T^1}{T} \;\leq\; \frac{1}{T}\sum_{t=1}^{T} x^{t\top} A y^t \;\leq\; \min_{y \in \Delta_n} \bar{x}^\top A y + \frac{R_T^2}{T}$$

Therefore:

$$\max_x x^\top A\bar{y} \;\leq\; \min_y \bar{x}^\top A y \;+\; \frac{R_T^1 + R_T^2}{T}$$

But by the minimax theorem: $\max_x x^\top A\bar{y} \geq v^* \geq \min_y \bar{x}^\top A y$.

So both quantities are within $\varepsilon = (R_T^1 + R_T^2)/T$ of v^*.

$\Rightarrow (\bar{x}, \bar{y})$ is an ε-NE. □

Interpretation

- The *average strategies* converge to NE, not the *last-iterate* strategies.
- Last-iterate strategies may oscillate (and typically do).
- Rate: $\varepsilon = O(\sqrt{\ln n / T})$ — logarithmic in game size!
- This is an **efficient algorithm** for computing approximate NE in zero-sum games.

Comparison to LP

- LP solves zero-sum NE exactly in poly time.
- Regret minimization gives ε-NE, but is simpler, parallelizable, and extends to structured games (e.g., extensive form).

Rate of Convergence: Summary

Algorithm	**Regret Bound**	**Rounds for ε-NE**
MWU (both players)	$O(\sqrt{T \ln n})$	$O(\ln n/\varepsilon^2)$
Regret Matching	$O(\sqrt{Tn})$	$O(n/\varepsilon^2)$
Regret Matching+	$O(\sqrt{Tn})$ (better constant)	$O(n/\varepsilon^2)$
Optimistic MWU	$O(\sqrt{T \ln n})$	$O(\ln n/\varepsilon^2)$

All achieve ε-NE in polynomial number of rounds.

The $\ln n$ vs. n distinction matters for large games.

Blackwell Approachability

The General Theorem Behind No-Regret

We have used no-regret guarantees repeatedly:

- no-external-regret $\Rightarrow$ CCE, no-swap-regret $\Rightarrow$ CE (L8);
- in zero-sum games, average strategies $\rightarrow$ Nash (just now).

Underneath all of these sits one geometric master theorem: **Blackwell approachability**.

The idea

Replace the *scalar* payoff of the minimax theorem with a *vector* payoff, and ask: can a player steer the *average* payoff vector into a target convex set, no matter what the opponent does? Regret minimization is exactly the special case where the target is the nonpositive orthant.

Vector-Payoff Games

Setup

- Player chooses $x \in \Delta_m$, opponent chooses $y \in \Delta_n$.
- Instead of a scalar, each action profile $a = (i, j)$ yields a **vector payoff** $\mathbf{g}(a) \in \mathbb{R}^d$; in expectation $\mathbf{g}(x, y) = \sum_{i,j} x_i\, y_j\, \mathbf{g}(i, j) \in \mathbb{R}^d$.
- Over rounds $t = 1, \ldots, T$, realized vectors $\mathbf{g}^t = \mathbf{g}(x^t, y^t)$ accumulate into the **running average** $\bar{\mathbf{g}}_T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{g}^t$.

Question: can the player force $\bar{\mathbf{g}}_T$ toward a target region, regardless of the opponent's sequence $y^1, \ldots, y^T$?

Approachable Sets

Let $\operatorname{dist}(\mathbf{z}, C) = \min_{\mathbf{c} \in C} \|\mathbf{z} - \mathbf{c}\|_2$ denote Euclidean distance to a closed set C.

Definition (Blackwell)

A nonempty closed convex set $C \subseteq \mathbb{R}^d$ is **approachable** for the player if there is a strategy such that, for *every* opponent sequence,

$$\limsup_{T \to \infty} \operatorname{dist}(\bar{\mathbf{g}}_T,\ C) = 0.$$

- Scalar minimax forces a *number* past a threshold; approachability forces a *vector* into a region.
- The guarantee is worst-case over the opponent — no stochastic assumption, exactly as with regret.

Blackwell's Theorem

Theorem (Blackwell, 1956)

A nonempty closed convex set $C \subseteq \mathbb{R}^d$ is approachable *if and only if* every halfspace containing C is forceable: for every direction $\lambda \in \mathbb{R}^d$ there is a mixed action $x_\lambda \in \Delta_m$ with

$$\langle \lambda,\, \mathbf{g}(x_\lambda, y) \rangle \;\le\; \sup_{\mathbf{c} \in C} \langle \lambda, \mathbf{c} \rangle \qquad \text{for all } y \in \Delta_n.$$

- A pointwise (per-direction) minimax condition: in each direction λ, the player keeps the expected vector payoff on the C-side of the supporting halfspace, whatever the opponent does.
- For convex C this condition is both necessary and sufficient.

Blackwell (1956), "An analog of the minimax theorem for vector payoffs," *Pacific J. Math.* 6(1):1–8.

Proof Idea: The Blackwell Operator

If $\bar{\mathbf{g}}_t \notin C$, let $\mathbf{c}_t = \Pi_C(\bar{\mathbf{g}}_t)$ be its projection and $\lambda_t = \bar{\mathbf{g}}_t - \mathbf{c}_t$ the separating direction.

- **Blackwell operator:** play $x^{t+1} = x_{\lambda_t}$, the halfspace action for λ_t (myopic best response to the distance functional). By the halfspace condition, $\langle \lambda_t, \mathbf{g}^{t+1} - \mathbf{c}_t \rangle \leq 0$.
- **One-step improvement** (Pythagoras): with D bounding $\|\mathbf{g}^{t+1} - \mathbf{c}_t\|$,

$$\operatorname{dist}^2(\bar{\mathbf{g}}_{t+1}, C) \;\leq\; \left(\tfrac{t}{t+1}\right)^2 \operatorname{dist}^2(\bar{\mathbf{g}}_t, C) + \frac{D^2}{(t+1)^2}.$$

- Telescoping gives $\operatorname{dist}(\bar{\mathbf{g}}_T, C) = O(D/\sqrt{T}) \to 0$.

The nonpositive cross term $\langle \lambda_t, \mathbf{g}^{t+1} - \mathbf{c}_t \rangle \leq 0$ is what makes the average *contract* toward C.

No-Regret = Approachability of the Orthant

Take $d = n$, one coordinate per action. After playing x^t and seeing loss ℓ^t, define the **regret vector** $\mathbf{g}^t \in \mathbb{R}^n$ by $g_a^t = \langle x^t, \ell^t \rangle - \ell_a^t$. Its running average is $(\bar{\mathbf{g}}_T)_a = R_a^T / T$, the average regret to action a.

- Target the **nonpositive orthant** $C = \mathbb{R}^n_{\leq 0}$ (closed, convex): its Euclidean distance is $\operatorname{dist}(\bar{\mathbf{g}}_T, C) = \left\| [\bar{\mathbf{g}}_T]^+ \right\|_2$.
- This $\to 0$ *iff* $\max_a [R_a^T / T]^+ = R_T / T \to 0$ — i.e. iff the learner has **no external regret**.

The payoff

No-external-regret learning *is* approachability of the nonpositive orthant; Hart & Mas-Colell (2000) derived regret matching as Blackwell's strategy for this target.

Approachability as the Engine

One theorem, many corollaries:

- **Regret matching** is the Blackwell operator for the orthant: the separating direction is $\lambda_t = [\bar{\mathbf{g}}_t]^+$, and playing $x^{t+1} \propto [R_a^t]^+$ recovers the L8 update. RM+ (next) refines the same operator.
- **Zero-sum Nash** (just proved) is the two-sided consequence: both players approaching their orthants drives the duality gap to 0.
- **CFR** (preview): run a *local* approachability instance at each information set — local regrets $\to 0$ forces global regret $\to 0$.

CCE/CE (L8), zero-sum Nash (today), and CFR (next week) are all the same geometric fact: driving an average vector into a convex target.

Regret Matching+ (RM+)

Motivation for RM+

- Standard regret matching: cumulative regrets can be very negative before becoming positive.
- Observation: negative regrets are "wasted" — they contribute nothing to the strategy (clipped to zero anyway).
- **Idea:** Truncate cumulative regrets at zero after each round.

Regret Matching+ (Tammelin, 2014)
Same as regret matching, but after each update:

$$R_i^t \leftarrow \max(R_i^t, 0)$$

Cumulative regrets are never allowed to go negative.

RM+: Algorithm

Regret Matching+

1. Initialize $R_i^0 = 0$ for all $i \in [n]$.
2. At round t:
   - Play $x_i^t = [R_i^{t-1}]^+ / \sum_j [R_j^{t-1}]^+$ (uniform if all zero).
   - Observe loss ℓ^t.
   - Update: $R_i^t = \max\big(R_i^{t-1} + \langle x^t, \ell^t \rangle - \ell_i^t,\ 0\big)$.

Key difference from RM: the $\max(\cdot, 0)$ truncation at each step, not just when forming the strategy.

RM+: Why It Helps

- In standard RM, cumulative regret for a bad action can drop to -1000 before the action becomes relevant again. Takes many rounds to climb back above zero.
- In RM+, regret is reset to 0 immediately $\Rightarrow$ the action can become active again quickly when conditions change.
- **Empirically:** RM+ converges much faster than RM, especially in large games.

Theoretical Guarantee

RM+ has the same asymptotic regret bound $O(\sqrt{Tn})$ as RM, but with better constants in practice.

RM+ and Counterfactual Regret Minimization

- RM+ was introduced in the context of **CFR**+ (Tammelin et al., 2015).
- CFR+ = Counterfactual Regret Minimization with RM+ as the local regret minimizer.
- Solved heads-up limit Texas hold'em poker (Bowling et al., 2015): the first nontrivial imperfect-information game to be "essentially solved."
- CFR+ converged orders of magnitude faster than original CFR.

Comparison of Regret Minimization Approaches

Algorithm Comparison

Algorithm	Regret	Memory	Practical Speed
MWU / Hedge	$O(\sqrt{T \ln n})$	$O(n)$	Fast
Regret Matching	$O(\sqrt{Tn})$	$O(n)$	Fast
Regret Matching+	$O(\sqrt{Tn})$	$O(n)$	Very fast
FTRL (entropy)	$O(\sqrt{T \ln n})$	$O(n)$	Fast
Optimistic MWU	$O(\sqrt{T \ln n})$	$O(n)$	Fast (last-iterate)

- MWU has better dependence on n ($\ln n$ vs. n).
- RM/RM+ are simpler to implement and often faster in practice.
- Choice depends on game size and structure.

Last-Iterate vs. Average-Iterate Convergence

Two notions of convergence:

1. **Average-iterate:** $\bar{x} = \frac{1}{T}\sum_t x^t \to \text{NE}$.
2. **Last-iterate:** $x^T \to \text{NE}$ as $T \to \infty$.

- Standard no-regret algorithms: average-iterate convergence in zero-sum games.
- Last-iterate behavior: MWU/RM can **cycle** in zero-sum games (Mertikopoulos et al., 2018).
- **Optimistic** variants (OMWU, optimistic FTRL): achieve last-iterate convergence.

Theorem (Daskalakis & Panageas, 2019)
Optimistic MWU converges in last iterate to NE in zero-sum games at rate $O(1/\sqrt{T})$.

Optimistic Multiplicative Weights Update

OMWU

1. At round t: use "prediction" $\hat{\ell}^t$ of the loss (typically $\hat{\ell}^t = \ell^{t-1}$).
2. Update:
$$\tilde{w}_i^{t+1} = w_i^t \cdot e^{-\eta \ell_i^t} \qquad w_i^{t+1} = \tilde{w}_i^{t+1} \cdot e^{-\eta \hat{\ell}_i^{t+1}}$$
3. Play $x^{t+1} \propto w^{t+1}$.

- "Optimism": the algorithm anticipates that the next loss will be similar to the current one.
- In zero-sum games (where the opponent also adapts slowly), this prediction is accurate $\Rightarrow$ faster convergence.
- Regret bound: $O(\sum_t \|\ell^t - \ell^{t-1}\|^2)$ — small when losses change slowly.

Connection to CFR (Preview)

From Normal Form to Extensive Form

- So far: **normal-form** (matrix) games.
- Real games (poker, chess) have **sequential** moves and **imperfect information**.
- These are modeled as **extensive-form games** (game trees).
- Problem: the normal-form representation of an extensive-form game is *exponentially large*.

Question: Can we apply regret minimization to extensive-form games without enumerating all strategies?

Counterfactual Regret Minimization (CFR): Preview

Key Idea (Zinkevich et al., 2007)

Decompose the global regret into **local** regrets at each information set, and minimize each one independently.

- At each information set h: run a separate regret minimizer over the available actions.
- The "loss" at h is the **counterfactual value**: expected utility conditional on reaching h.
- **Theorem:** if local regret at every h is bounded by R_T^h, then overall regret is bounded by $\sum_h R_T^h$.
- Iterates over the game tree (not the strategy space) $\Rightarrow$ polynomial per iteration.

Week 3: Full treatment of CFR and its variants.

CFR: Impact

- CFR (2007) and CFR+ (2014) are the backbone of modern **poker AI**.
- Libratus (2017) and Pluribus (2019): superhuman poker agents built on CFR variants.
- Key enabler: regret minimization theory provides convergence guarantees; game-tree decomposition provides scalability.

Milestone	**Year**
CFR introduced	2007
Heads-up limit hold'em solved (CFR+)	2015
Libratus beats pros (no-limit hold'em)	2017
Pluribus: 6-player no-limit hold'em	2019

Bandit Feedback vs. Full Information

Feedback Models

Full Information Feedback

After each round, the learner observes the **entire loss vector** $\ell^t \in [0,1]^n$.

Bandit Feedback

After each round, the learner observes **only** the loss of the chosen action $\ell^t_{i_t}$.

- Full information: MWU achieves $O(\sqrt{T \ln n})$ regret.
- Bandit: much harder — less information per round.
- Fundamental tradeoff: exploration vs. exploitation.

Bandit Regret Bounds

Theorem

- **Lower bound:** Any bandit algorithm has $R_T \geq \Omega(\sqrt{Tn})$.
- **Upper bound (EXP3, Auer et al. 2002):** $R_T \leq O(\sqrt{Tn \ln n})$.

Compare to full information:

Feedback	Lower Bound	Upper Bound
Full information	$\Omega(\sqrt{T \ln n})$	$O(\sqrt{T \ln n})$
Bandit	$\Omega(\sqrt{Tn})$	$O(\sqrt{Tn \ln n})$

The price of limited feedback: $\sqrt{n / \ln n}$ factor.

EXP3 Algorithm: Overview

EXP3 (Exponential-weight algorithm for Exploration and Exploitation)

1. Maintain weights w_i^t as in MWU.
2. Play action i with probability $p_i^t = (1-\gamma)\frac{w_i^t}{\sum_j w_j^t} + \frac{\gamma}{n}$ (mix with uniform for exploration).
3. Construct **importance-weighted estimate**:

$$\hat{\ell}_i^t = \begin{cases} \ell_i^t / p_i^t & \text{if } i = i_t \\ 0 & \text{otherwise} \end{cases}$$

4. Update weights: $w_i^{t+1} = w_i^t \cdot \exp(-\eta\, \hat{\ell}_i^t)$.

$\hat{\ell}^t$ is an **unbiased** estimate of ℓ^t: $\mathbb{E}[\hat{\ell}_i^t] = \ell_i^t$.

Bandit Feedback in Games

- In many real games, players observe only their own payoff (not the opponent's strategy or the full payoff matrix).
- This is the **bandit** feedback model.
- Convergence to equilibria still holds, but slower:
  - Bandit no-external-regret $\Rightarrow$ CCE convergence at rate $O(\sqrt{n/T})$.
  - Need $T = O(n/\varepsilon^2)$ rounds for ε-CCE (vs. $O(\ln n/\varepsilon^2)$ with full info).
- For zero-sum games: bandit regret minimization still yields ε-NE, but with worse dependence on n.

Summary

Summary

1. **Zero-sum key theorem:** If both players use no-regret algorithms, *average* strategies converge to NE at rate $O(\sqrt{\ln n / T})$.
2. **RM+:** Truncate cumulative regrets at zero; same asymptotic bound but much faster in practice; backbone of CFR+.
3. **Last-iterate convergence:** Standard MWU/RM may cycle; optimistic variants (OMWU) achieve last-iterate convergence.
4. **CFR preview:** Decompose extensive-form game regret into local information-set regrets; enables solving large games like poker.
5. **Bandit feedback:** Observing only own payoff costs a $\sqrt{n / \ln n}$ factor; EXP3 is the canonical algorithm.
6. **Next week:** Extensive-form games, CFR in detail, and applications to poker.