No-Regret Learning

Game Theory · Module 9 — Learning In Games

Node 9.2

Position in Knowledge Graph

System: Game Theory **Structural Domain:** Learning in Games **Node:** 9.2

No-regret learning is the framework that repairs the defects of fictitious play by replacing "best-respond to empirical opponent play" with a weaker but more robust guarantee: *the learner should not regret, in hindsight, having played a fixed action for all time.* This shift from optimality-against-a-model to regret-minimization-against-an-arbitrary-sequence is the conceptual core of modern online learning theory, and it has two consequences for game theory that reshape the field. First, when *all* players in a finite game use no-regret algorithms, the joint empirical frequency of play converges to the set of *coarse correlated equilibria* — a set that strictly contains Nash in general but equals Nash in two-player zero-sum games. Second, the convergence rate is $O(T^{-1/2})$, exponentially faster than fictitious play's sluggish polynomial, which makes no-regret methods computationally viable. The connection between no-regret learning and equilibrium is one of the deepest structural results in the theory: it shows that a *very weak individual rationality guarantee* (no external regret) is sufficient for *collective convergence* to a well-defined equilibrium concept. Node 9.2 develops the regret framework, proves the fundamental convergence theorem, and introduces the Hannan consistency criterion that links individual learning to equilibrium theory.

Formal Definition

Fix a finite action set $A = \{1, 2, \ldots, K\}$ for a single learner. At each round $t = 1, 2, \ldots, T$, the learner picks an action $a^t \in A$ (possibly randomly), and then observes a loss vector $\ell^t \in [0,1]^K$, where ℓ^t_k is the loss that would have been incurred by action k. The learner suffers loss $\ell^t_{a^t}$.

External regret. The external regret with respect to a fixed action k is

$$R^T_k \;=\; \sum_{t=1}^{T} \ell^t_{a^t} - \sum_{t=1}^{T} \ell^t_k.$$

The **maximum external regret** is

$$R^T \;=\; \max_{k \in A} R^T_k \;=\; \sum_{t=1}^{T} \ell^t_{a^t} - \min_{k \in A} \sum_{t=1}^{T} \ell^t_k.$$

This measures how much worse the learner did than the single best fixed action in hindsight.

No-regret (Hannan consistency). An algorithm is **no-regret** (or **Hannan consistent**) if, against *any* sequence of loss vectors (even adversarially chosen),

$$\frac{R^T}{T} \;\to\; 0 \qquad \text{as } T \to \infty.$$

A stronger quantitative guarantee: the algorithm achieves **sublinear regret** if $R^T = o(T)$, and achieves $O(\sqrt{T})$ regret if $R^T \le C\sqrt{T \ln K}$ for some constant C.

Internal regret. A finer notion: the internal regret of switching action j to action k whenever j was played is

$$R^T_{j\to k} \;=\; \sum_{t:a^t=j} \left(\ell^t_j - \ell^t_k\right).$$

An algorithm has **no internal regret** if $\max_{j,k} R^T_{j\to k}/T \to 0$. No internal regret is strictly stronger than no external regret.

Coarse correlated equilibrium (CCE). A distribution $\mu \in \Delta(S)$ over joint strategy profiles is a CCE if for every player i and every fixed action $s'_i \in S_i$:

$$\mathbb{E}_{s\sim\mu}\big[u_i(s)\big] \;\ge\; \mathbb{E}_{s\sim\mu}\big[u_i(s'_i, s_{-i})\big].$$

No player can improve their expected payoff by unilaterally switching to a *fixed* alternative, where the expectation is over the correlation device.

Correlated equilibrium (CE). Stronger: μ is a CE if no player can improve by a *conditional* switch — after observing their own recommended action, no player wants to deviate. Every CE is a CCE; the converse is false in general.

Intuition

Three ways to understand what no-regret buys you.

The insurance policy. External regret asks: "Could I have done better by committing to a single action for all time?" If the answer is "not by much," you have no external regret. This is a *minimal* rationality requirement — you are not comparing yourself to the best adaptive strategy, or the best sequence of actions, only to the best fixed action. The requirement is deliberately weak because the guarantee must hold against *arbitrary* loss sequences, including adversarial ones. A stronger benchmark (e.g., competing with the best *sequence* of actions) is unachievable against adversaries.

The equilibrium factory. The most surprising consequence of no-regret is not about individual learning but about what happens when *everyone* uses it simultaneously. If all players in a finite game independently run no-regret algorithms, the empirical distribution of joint play converges to the set of coarse correlated equilibria. In two-player zero-sum games, every CCE yields the same payoff as Nash (because the CCE polytope collapses to the Nash set), so no-regret learning *computes Nash equilibria* in zero-sum games as a side effect. This is the structural link between regret and equilibrium.

The upgrade from fictitious play. Fictitious play best-responds to the empirical opponent distribution, which sounds strong but has no regret guarantee (Shapley's counterexample shows fictitious play can accumulate linear regret in non-zero-sum games). No-regret algorithms are weaker per-step — they do not necessarily best-respond — but they guarantee sublinear regret against *any* opponent, including adversarial ones. This robustness is what enables the equilibrium convergence theorem.

Structural Role

9.2 is the theoretical backbone of Module 9. It provides:

- **The regret framework** that defines the convergence criterion for all subsequent algorithms.
- **The CCE convergence theorem** that links individual regret bounds to collective equilibrium.

- **The conceptual foundation for 9.3 (CFR)**: CFR is no-regret learning applied independently to each information set in an extensive-form game.
- **The benchmark for 9.4 (MWU/Hedge)**: Hedge is a specific no-regret algorithm that achieves the optimal $O(\sqrt{T \ln K})$ regret bound.
- **The convergence criteria for 9.5**: the question "does a learning dynamic converge to Nash?" is answered by asking whether the associated regret goes to zero.

Without 9.2, the entire module is a collection of algorithms without a unifying performance criterion. With 9.2, every algorithm in Module 9 can be evaluated on the same axis: *how fast does its regret grow?*

Mathematical Development

The fundamental theorem: no-regret implies CCE convergence

Theorem. Let G be a finite n-player normal-form game. Suppose each player i uses a no-regret algorithm to select actions over rounds $t = 1, \ldots, T$. Let $\mu^T \in \Delta(S)$ denote the empirical distribution of joint play: $\mu^T(s) = \frac{1}{T}|\{t : s^t = s\}|$. Then every limit point of $\{\mu^T\}$ is a coarse correlated equilibrium.

Proof. Fix player i and a fixed action s_i'. The external regret of player i with respect to s_i' is

$$R_{i,s_i'}^T \;=\; \sum_{t=1}^{T} u_i(s_i', s_{-i}^t) - \sum_{t=1}^{T} u_i(s_i^t, s_{-i}^t).$$

(Using gains rather than losses; the sign flips.) No-regret guarantees $R_{i,s_i'}^T / T \to 0$ for every s_i'. Dividing by T:

$$\frac{1}{T}\sum_{t=1}^{T} u_i(s_i^t, s_{-i}^t) \;\geq\; \frac{1}{T}\sum_{t=1}^{T} u_i(s_i', s_{-i}^t) - \frac{R_{i,s_i'}^T}{T}.$$

As $T \to \infty$, the left side converges to $\mathbb{E}_\mu[u_i(s)]$ and the right to $\mathbb{E}_\mu[u_i(s_i', s_{-i})]$, with the regret term vanishing. So $\mathbb{E}_\mu[u_i(s)] \geq \mathbb{E}_\mu[u_i(s_i', s_{-i})]$, which is the CCE condition. □

Corollary: Nash in zero-sum games

In a two-player zero-sum game, every CCE yields the minimax value to both players (because the CCE polytope is a singleton in payoff space for zero-sum games — any correlation device that prevents unilateral improvement must give each player at least their minimax value, and zero-sum forces the payoffs to sum to zero). Therefore, when both players use no-regret, the empirical distribution converges to a Nash equilibrium *in payoff.*

The exploitability of the empirical mixed strategies $\bar{\sigma}_1^T, \bar{\sigma}_2^T$ is bounded by:

$$\text{exploitability} \;\leq\; \frac{R_1^T + R_2^T}{T}.$$

With $O(\sqrt{T})$ regret for each player, exploitability drops as $O(T^{-1/2})$ — much faster than fictitious play.

No internal regret implies CE convergence

Theorem (Foster and Vohra 1997, Hart and Mas-Colell 2000). If all players use no-*internal*-regret algorithms, the empirical distribution converges to the set of correlated equilibria (not just coarse correlated equilibria). This is a strictly stronger guarantee.

Algorithms achieving no internal regret exist (e.g., Blum–Mansour calibration-based algorithms), but they are more complex than external-regret algorithms and typically have higher computational cost. For game-solving

purposes, external regret (and hence CCE) is usually sufficient, especially in zero-sum games where CCE = Nash.

Regret matching (Hart and Mas-Colell 2000)

A specific no-regret algorithm that is the direct precursor to CFR:

1. Track cumulative regret for each action k: $R_k^t = \sum_{\tau=1}^{t}(u_i(k, s_{-i}^\tau) - u_i(s_i^\tau, s_{-i}^\tau))$.
2. Define the regret-positive part: $R_k^{t,+} = \max(R_k^t, 0)$.
3. At round $t+1$, play action k with probability proportional to $R_k^{t,+}$:

$$\sigma_i^{t+1}(k) \;=\; \begin{cases} \dfrac{R_k^{t,+}}{\sum_j R_j^{t,+}} & \text{if } \sum_j R_j^{t,+} > 0, \\ \dfrac{1}{K} & \text{otherwise.} \end{cases}$$

Theorem (Hart and Mas-Colell 2000). Regret matching achieves $R^T = O(\sqrt{T})$ external regret, and is therefore Hannan consistent.

Regret matching is the engine inside CFR (Node 9.3): at each information set, the player runs regret matching independently, and the global convergence follows from the per-info-set no-regret guarantee.

The Hannan set

Define the **Hannan set** as the set of payoff vectors achievable by strategies that are no-regret against any opponent. For two-player zero-sum games, the Hannan set equals the singleton $\{(V, -V)\}$ where V is the game value. For general games, the Hannan set contains the set of Nash equilibrium payoffs but may be larger. The CCE set is the equilibrium characterization of the Hannan set: the joint distributions reachable by no-regret play are exactly the CCEs.

Regret bounds and rates

Algorithm	External regret bound	Notes
Hedge / MWU (9.4)	$O(\sqrt{T \ln K})$	Optimal among oblivious algorithms
Regret matching	$O(\sqrt{T})$	Per-action tracking, precursor to CFR
Follow the perturbed leader	$O(\sqrt{T \ln K})$	Random perturbation to leader algorithm
EXP3 (bandit setting)	$O(\sqrt{TK \ln K})$	Only observes own payoff, not full vector

The $O(\sqrt{T \ln K})$ bound is optimal in the adversarial full-information setting (lower bound by Cesa-Bianchi et al.).

Worked Example

Two-player zero-sum: regret matching computes Nash

Consider Rock-Paper-Scissors with payoff matrix (for Row):

$$A = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}.$$

Both players run regret matching. Initialize all cumulative regrets to 0, so both play uniform $(1/3, 1/3, 1/3)$.

Round 1. Both play uniform. Expected payoff for Row from each action: 0 (by symmetry). Regret of each action: 0. Next-round strategy: still uniform.

By symmetry, regret matching stays at uniform forever in this game — the Nash equilibrium is reached in a single step. This is a degenerate example, but it illustrates the fixed-point property: if the current empirical play is Nash, regret matching stays there.

Now break the symmetry. Suppose Row plays R in round 1, Column plays S. Row's payoff: $u_1(R, S) = -1$. Counterfactual payoffs: $u_1(R, S) = -1$, $u_1(P, S) = -1$, $u_1(S, S) = 0$. Cumulative regret: $R_R^1 = 0$, $R_P^1 = 0$, $R_S^1 = 1$. (Wait — using the row-payoff against column's S: $u_1(R, S) = 1$, $u_1(P, S) = -1$, $u_1(S, S) = 0$.) Let me re-derive with the standard RPS matrix where Row = R beats Column = S:

Row played R, got payoff $u_1(R, S) = 1$. Regret of not playing P: $u_1(P, S) - 1 = -1 - 1 = -2$. Regret of not playing S: $u_1(S, S) - 1 = 0 - 1 = -1$. All regrets are negative, so $R_k^{1,+} = 0$ for all k. Row plays uniform next round. Column analogously updates. Over many rounds, the cumulative regrets oscillate around zero, and the empirical strategy converges to $(1/3, 1/3, 1/3)$.

Exploitability decay

After T rounds of regret matching by both players, the exploitability of the average strategies is bounded by $(R_1^T + R_2^T)/T \leq C/\sqrt{T}$. For RPS with 3 actions, after 10,000 rounds, exploitability is on the order of 0.01 — effectively Nash.

Poker Anchor

Concept mapping

Formal object	Poker instantiation
Action set A	Actions at a decision point: fold, call, raise (various sizes)
Loss vector ℓ^t	The EV of each action given the opponent's play this iteration
External regret	"How much better would I have done if I had always raised here instead of what I actually did?"
Regret matching	The per-info-set update rule inside every CFR solver
CCE convergence	The solver's average strategy profile converging to approximate Nash
Exploitability $O(T^{-1/2})$	The rate at which a CFR-based solver's output approaches GTO

Concrete situation: why solvers converge

When you run a solver (PioSolver, GTO+, MonkerSolver) on a postflop spot, the solver is running regret matching at every information set simultaneously. Each info set is an independent no-regret learner: it tracks cumulative regret for each available action (fold, call, bet 33%, bet 67%, etc.), and at each iteration it mixes in proportion to positive regrets. The fundamental theorem guarantees that the average strategy across all

info sets converges to a CCE, which in the two-player zero-sum poker game is a Nash equilibrium. The exploitability drops as $O(T^{-1/2})$, so after enough iterations the strategy is approximately GTO.

This is the *entire* theoretical justification for why poker solvers produce Nash strategies. There is nothing else underneath — it is regret matching at every info set, plus the CCE-to-Nash collapse in zero-sum games.

Concrete situation: the meaning of "iterations" in a solver

When PioSolver reports "500 iterations, exploitability 0.3% pot," it is saying: regret matching has run for 500 rounds, and the sum of both players' maximum external regrets divided by iterations is 0.3% of pot. The $O(T^{-1/2})$ rate means that cutting exploitability in half requires quadrupling the iterations. This is why getting from 1% to 0.1% exploitability takes 100x more compute than getting from 10% to 1%.

What this understanding buys you

Understanding regret is understanding why solvers work. If you know that the solver is running no-regret learning at every info set, you understand why convergence is guaranteed (the CCE theorem), why it is slow (the $\sqrt{T}$ rate), and why the output is Nash in zero-sum games (the CCE = Nash collapse). You also understand what the solver is *not* doing: it is not searching for a Nash equilibrium directly, it is not doing backward induction, and it is not reasoning about opponent rationality. It is just minimizing regret locally, and the Nash equilibrium emerges from the global interaction of local no-regret learners.

Regret matching is the link between 9.2 and 9.3. CFR (Node 9.3) is regret matching lifted to extensive-form games via the counterfactual regret decomposition. If you understand regret matching in the normal-form setting of 9.2, the only new ingredient in CFR is the trick that decomposes global regret into per-info-set counterfactual regrets.

Cross-Links

- **9.1 Fictitious play** — predecessor with no regret guarantee; motivates the move to no-regret.
- **9.3 CFR** — no-regret learning applied to extensive-form games via counterfactual regret.
- **9.4 Multiplicative weights / Hedge** — the canonical no-regret algorithm with optimal bounds.
- **9.5 Convergence to equilibrium** — CCE convergence is the main positive result.
- **2.1 Nash equilibrium** — the target in zero-sum games.
- **2.5 Correlated equilibrium** — the target under no-internal-regret.
- **Online learning / optimization** — the broader field in which regret minimization originates.

Structural Compression

Invariant: If every player in a finite game independently minimizes external regret (achieving sublinear $R^T = o(T)$), the empirical distribution of joint play converges to the set of coarse correlated equilibria; in two-player zero-sum games this set collapses to Nash, giving a $O(T^{-1/2})$ equilibrium computation method.

Mechanism: Each player tracks cumulative regret per action and mixes in proportion to positive regrets (regret matching) or exponential weights (Hedge). The CCE convergence theorem translates individual sublinear-regret guarantees into collective equilibrium convergence via a telescoping-sum argument on the regret definitions.

Cross-System Echo: The regret framework is the game-theoretic instantiation of **online convex optimization**: the learner picks a point in a convex set, nature reveals a loss function, and the learner's goal is sublinear regret against the best fixed point. In machine learning, this is the foundation of online gradient descent, AdaGrad, and the entire adaptive-learning-rate literature. In statistics, it connects to sequential prediction and the minimax theory of estimation. The structural invariant — *local sublinear regret implies global consistency* — is the same across all these domains.

Exercises

1. Prove that regret matching achieves $O(\sqrt{T})$ external regret. (Hint: use a potential function $\Phi^t = \sum_k (R_k^{t,+})^2$ and bound the per-round change.)
2. Show that every Nash equilibrium payoff vector is a CCE payoff vector, but construct a 2-player game where the set of CCE payoff vectors strictly contains the set of Nash payoff vectors.
3. Verify the CCE-to-Nash collapse in zero-sum games: prove that if μ is a CCE of a two-player zero-sum game, then the marginals μ_1, μ_2 form a Nash equilibrium.
4. An algorithm has regret bound $R^T \leq 3\sqrt{T \ln K}$. After 10,000 rounds in a zero-sum game with $K = 5$ actions per player, bound the exploitability of the average strategy profile.
5. Construct an adversarial loss sequence against which fictitious play accumulates linear regret (i.e., $R^T = \Omega(T)$), showing that fictitious play is *not* Hannan consistent in general.
6. Explain why no-*internal*-regret algorithms converge to correlated equilibria rather than coarse correlated equilibria. What is the structural difference between the two regret concepts that produces the tighter equilibrium set?
7. **Argue, structurally,** why the $O(\sqrt{T})$ regret bound is tight — no algorithm can guarantee $o(\sqrt{T})$ regret against adversarial loss sequences — and what this implies about the fundamental speed limit of equilibrium computation via self-play.

Game Theory · Module 9 — Learning In Games · Node 9.2
Concepts: convergence, nash-equilibrium, cost-function
Prerequisites: 9.1, 2.1, 2.3
Enables: 9.3, 9.4, 9.5

hbar.university — own.your.cognition