Gerard Vreeswijk, Intelligent Systems Group, Computer Science Department, Faculty of Sciences, Utrecht University, The Netherlands.
No-regret learning: motivation

- **Reinforcement Learning.** *Play those actions that were successful in the past.*

  - **Similarities:**
    1. Driven by past payoffs.
    2. Not interested in (the behaviour of) the opponent.
    3. Probabilistic.
    4. Smooth adaptation.
    5. Myopic.

- **No-regret learning:** might be considered as an extension of reinforcement learning
  
  No-regret $\overset{\text{Def}}{=} \text{play those actions that would have been successful in the past.}$

  - **Differences:**
    
    a) Keeping accounts of hypothetical actions rests on the assumption that a player is able to estimate payoffs of actions that were actually not played.  
    [Knowledge of the payoff matrix certainly helps, but is an even more severe assumption.]

    b) Bit more easy to obtain results regarding performance.
Qualitative features of reinforcement and regret

1. **Probabilistic choice.** A choice of action is never completely determined by history but has a random component.
   - The different magnitudes of the probabilities (arisen through experience) ensures exploitation of past experience.
   - The randomness ensures exploration.

2. **Smooth adaptation.** The mixed strategy adapts gradually, as with differential equations.
   - **No-regret learning.** (Conditionally) select a pure strategy that would have been most successful, given past play.
   - **Smoothed fictitious play.** Give a (perturbed) best response to the (recent) empirical frequency opponents’ actions.
   - **Hypothesis testing with smoothed best responses.** Give a best response to maintained beliefs about patterns of play.
Plan for today

Three parts.

1. **Basic concepts.**

2. **Proportional regret matching.** Hart and Mas-Colell (2000).


This presentation almost exclusively follows the second half of Ch. 2 of (Peyton Brown, 2004). This second half is dedicated to giving insight behind convergence results in (Foster & Vohra, 1999) and (Hart & Mas-Colell, 2000).

---


Peyton Young, H. (2004): *Strategic Learning and it Limits*, Oxford UP. Ch. 2: “Reinforcement and Regret”
Part I: Basic concepts
## No-regret: example

<table>
<thead>
<tr>
<th>Payoffs Player $A$</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actions Player $A$</td>
<td>L</td>
<td>R</td>
<td>L</td>
<td>L</td>
<td>R</td>
<td>R</td>
<td>L</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>Actions Player $B$</td>
<td>R</td>
<td>L</td>
<td>R</td>
<td>L</td>
<td>R</td>
<td>R</td>
<td>L</td>
<td>R</td>
<td>R</td>
<td>L</td>
<td>L</td>
</tr>
</tbody>
</table>

- Suppose $A$ is offered to replay the first 11 periods, under the proviso that he must play one pure strategy throughout.

<table>
<thead>
<tr>
<th>Payoff</th>
<th>Average Regret</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rounds 1-11:</td>
<td>3</td>
</tr>
<tr>
<td>Had $L$ been played:</td>
<td>6</td>
</tr>
<tr>
<td>Had $R$ been played:</td>
<td>5</td>
</tr>
</tbody>
</table>

- It is ignored that $B$ would likely have played different if he knew $A$ would play different.

- Thus, no-regret fails to take the interactive nature of play into account.
No-regret: some notation

- The average payoff up to and including round $t$ is
  $$\bar{u}^t = \text{Def} \frac{1}{t} \sum_{s=1}^{t} u(x_s^s, y_s^s)$$

- For each action $x$, the hypothetical average payoff for playing $x$ is
  $$\bar{h}^t_x = \text{Def} \frac{1}{t} \sum_{s=1}^{t} u(x, y_s^s)$$

- For each action $x$, the average regret from not having played $x$ is
  $$\bar{r}^t_x = \text{Def} \bar{h}^t_x - \bar{u}^t$$

- Multiple regret may be represented as a vector
  $$\bar{r}^t = \text{Def} (\bar{r}_1^t, \ldots, \bar{r}_k^t)$$

- A given realisation of play
  $$\omega = (x_1, y_1), \ldots, (x_t, y_t), \ldots$$
  is said to have no regret if, for all actions $x$,
  $$\lim_{T \to \infty} \sup_{t \to \infty} \bar{r}^t_x(\omega)$$
  $$= \lim_{T \to \infty} \sup \{ \bar{r}^t_x(\omega) \mid t \geq T \} \leq 0.$$
Part II: proportional regret matching
Strategies and regret matching

A strategy \( g : H \rightarrow \Delta(X) \) is said to have no regret if almost all its realisations of play have no regret.

The objective is to formulate a strategy without regret. One such strategy (we can already say) is proposed by Hart and Mas-Colell (2000):

\[
q^{t+1}_x = \text{Def} \frac{[\bar{r}^t_x]^+}{\sum_{x' \in X} [\bar{r}^t_{x'}]^+}
\]

where \([z]^+ = \text{Def} z \geq 0 \ ? \ z : 0\). This rule is called proportional regret matching, or regret matching (RM for short).

**Theorem** (Hart & Mas-Colell, 2000). *In a finite game, regret matching by a given player almost surely yields no regret against every possible sequence of play by the opponent(s).*

(Proportional) regret differs from (proportional) reinforcement

\[
\begin{array}{cccccccccccc}
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
A & L & R & L & L & R & R & L & R & R & R & R & ? \\
B & R & L & R & L & R & L & R & L & R & L & L & ? \\
\end{array}
\]

Proportional regret:

<table>
<thead>
<tr>
<th></th>
<th>Payoff</th>
<th>Average regret</th>
<th>Mixed strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rounds 1-11:</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Had L been played</td>
<td>6</td>
<td>(\frac{(6-3)}{11})</td>
<td>(\frac{3}{5})</td>
</tr>
<tr>
<td>Had R been played</td>
<td>5</td>
<td>(\frac{(5-3)}{11})</td>
<td>(\frac{2}{5})</td>
</tr>
</tbody>
</table>

Proportional reinforcement:

<table>
<thead>
<tr>
<th></th>
<th>Accumulated payoff</th>
<th>Mixed strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action L:</td>
<td>1</td>
<td>(\frac{1}{3})</td>
</tr>
<tr>
<td>Action R:</td>
<td>2</td>
<td>(\frac{2}{3})</td>
</tr>
</tbody>
</table>
Regret matching cast into a form similar to reinforcement

- Through reinforcement with an aspiration level $\bar{u}^t$.
- Define the reinforcement increment for every $x$ in round $t$ as
  \[ \Delta r^t_x = u(x, y^t) - \bar{u}^t \]
- Define the propensities in round $t + 1$ as
  \[ \theta^{t+1}_x = \text{Def} \left[ \sum_{s=1}^{t} \Delta r^t_x \right]_+ \]
- This is like standard reinforcement, but now all actions in a given period are reinforced, whether or not they are actually played.
- Hypothetical reinforcement takes into account virtual payoffs. (Payoffs that never materialised.) The vector $\Delta r^t$ is a vector of “virtual” reinforcements—gains or losses relative to the current average that that would have materialised if a given action $x$ had been played at time $t$. 
Adequacy of regret matching: proof outline

- Let
  \[ r^t_x = \text{Def} \text{ accumulated regret for not playing } x, \text{ up to and including } t \]
  and
  \[ \bar{r}^t_x = \text{Def} \text{ average regret for not playing } x, \text{ up to and including } t. \]

Let
  \[ [\bar{r}^t_x]^+ = \text{Def} \bar{r}^t_x \geq 0 \quad \text{?} : \bar{r}^t_x : 0 \]

Dropping the subscript, \( x \), is a way to denote the complete vector \( [\bar{r}^t]^+ \) of average regrets.

- The objective is to show that
  \[ \lim_{t \to \infty} \sup [\bar{r}^t]^+ = 0 \]
  with probability one. (The zero on the RHS again is a vector.)
Incremental regret

Suppose there are only two actions, “1” and “2,” say.

1. If 1 is executed at $t + 1$ then
   - $r_{1}^{t+1}$ will not change.
   - $r_{2}^{t+1}$ changes with $u(2, y^{t+1}) - u(1, y^{t+1})$.
2. If 2 is executed at $t + 1$ then
   - $r_{1}^{t+1}$ changes with $u(1, y^{t+1}) - u(2, y^{t+1})$.
   - $r_{2}^{t+1}$ will not change.

Thus, if $\alpha^{t+1} = \text{Def } u(2, y^{t+1}) - u(1, y^{t+1})$ then incremental regret will be either $(0, \alpha^{t+1})$ or $(-\alpha^{t+1}, 0)$.

Suppose in round $t + 1$ a mixed strategy $q^{t+1} = (q_{1}^{t+1}, q_{2}^{t+1})$ is played. Then the expected incremental regret is

$$E[\Delta r^{t+1}] = (q_{1}^{t+1} \cdot 0 + q_{2}^{t+1} \cdot \alpha^{t+1}, q_{1}^{t+1} \cdot -\alpha^{t+1} + q_{2}^{t+1} \cdot 0)$$
Decrease of expected regret

The objective is to find a (mixed) strategy $g : H \rightarrow \Delta(\{1, 2\})$ such that

$$E[\bar{\rho}^{t+1} \mid r^t, \ldots, r^1] < \bar{\rho}^t$$

(1)

because then again one of the martingale convergence theorems can be applied to conclude $\lim_{t \to \infty} [\bar{\rho}^t]_+ = 0$. Since $\Delta E[r^{t+1}]$ is known, we have

$$E[\bar{\rho}^{t+1} \mid r^t, \ldots, r^1] = E\left[\frac{r^t + \Delta r^{t+1}}{t+1} \mid r^t, \ldots, r^1\right]$$

$$= \frac{t}{t+1} E\left[\frac{r^t}{t} \mid r^t, \ldots, r^1\right] + \frac{1}{t+1} E[\Delta r^{t+1} \mid r^t, \ldots, r^1]$$

$$= \frac{t}{t+1} \bar{\rho}^t + \frac{1}{t+1} E[\Delta r^{t+1} \mid r^t, \ldots, r^1]$$

$$= \frac{t}{t+1} \bar{\rho}^t + \frac{1}{t+1} (\alpha^{t+1} q_{2}^{t+1}, -\alpha^{t+1} q_{1}^{t+1})$$

The objective is to find a strategy $q_{t+1} = g(\xi^t)$ such that Eq. (1) is satisfied.
A strategy $q$ such that $\lim_{t \to \infty} [\bar{r}^t]^+ = 0$: 1st attempt

Take $q^t_1 = q^t_2 = 1/2$ for all $t$.

Then

$$E[\bar{r}^t_1 + \bar{r}^t_2] = E\left[\frac{r^{t-1}_1 + \Delta r^t_1}{t} + \frac{r^{t-1}_2 + \Delta r^t_2}{t}\right]$$

$$= E\left[\frac{r^{t-1}_1 + \alpha^t/2}{t} + \frac{r^{t-1}_2 - \alpha^t/2}{t}\right]$$

$$= E[\bar{r}^{t-1}_1 + \bar{r}^{t-1}_2]$$

Hence,

$$E[\bar{r}^t_1 + \bar{r}^t_2] = 0,$$

so that $\lim_{t \to \infty} \bar{r}^t_1 + \bar{r}^t_2 = 0$ with probability one.

However, the two regrets themselves may be unbounded and neutralise each other. (Which is not what we want—we want all regrets to be non-positive.)
A strategy $q$ such that $\lim_{t \to \infty} [\bar{r}^t]_+ = 0$ : 2nd attempt

- Each round $t$, choose an action that would have minimised regret in the previous round.

- Matching Pennies:
  - Switch if wrong action in previous round; else stay.
  - Won’t work: suppose you meet an opponent who happens to switch every round as well . . .
  - Won’t work in general: your corrections may by coincidence be “in phase” with the path of play of your opponent. Peyton Young:
    “Recall the no-regret must hold even when Nature is malevolent.”
    (p. 26)
A strategy $q$ such that $\lim_{t \to \infty} [\bar{r}_t]^+ = 0$: $E[\Delta r^{t+1}] \perp \bar{r}_t^+$

- Recall: our objective is $[\bar{r}_t]^+ \to 0$.
- To this end, choose $q^{t+1}$ such that
  
  $E[\Delta r^{t+1}] \perp [\bar{r}_t]^+$

Thus:

$$E[\Delta r^{t+1}] \cdot \bar{r}_t^+ = 0$$

$$\iff (\alpha^{t+1} q_2^{t+1}, -\alpha^{t+1} q_1^{t+1}) \cdot \bar{r}_t^+ = 0$$

$$\iff \alpha^{t+1} (q_2^{t+1} [\bar{r}_1]^+ - q_1^{t+1} [\bar{r}_2]^+) = 0$$

$$\iff q_1^{t+1} : q_2^{t+1} = [\bar{r}_1]^+ : [\bar{r}_2]^+$$

The last equation precisely amounts to proportional regret matching.

(Notice that $\alpha^{t+1}$ has left the stage.)

- Boundary cases are obvious and can be treated as follows:
  - If $\bar{r}_1^t > 0$ and $\bar{r}_2^t \leq 0$, then let $q^{t+1} = \text{Def} (1, 0)$.
  - If $\bar{r}_1^t \leq 0$ and $\bar{r}_2^t > 0$, then let $q^{t+1} = \text{Def} (0, 1)$.
  - If all regret is non-positive, then play a random action.
Stochastic dynamics of regret matching

• Expected incremental regret, \( E[\Delta r^{t+1}] \) is made orthogonal to the current regret, independently of the unknown \( \alpha^{t+1} \).
  (Because \( A \) does not know what \( B \) plays at \( t + 1 \), it is crucial that \( q^{t+1} \) does not depend on \( \alpha^{t+1} \).)

• \( E[\bar{r}^{t+1}] \) is a convex combination of \( \bar{r}^t_+ \) and \( E[\Delta r^{t+1}] \).

• Since \( E[\Delta r^{t+1}] \perp \bar{r}^t_+ \), \( E[\bar{r}^{t+1}] \) lies closer to the non-positive orthant than does \( \bar{r}^t_+ \), provided \( t \) is large.

• \( \bar{r}^t_+ \to 0 \) follows from Blackwell’s approachability theorem (PY, 2004, Ch. 4).
Stochastic dynamics of regret matching
Part III: 
$\epsilon$-Greedy Off-policy Regret Matching
**ε-Greedy regret matching** (Foster & Vohra, 1998)

**ε-greedy regret matching.** Let $\epsilon > 0$ small.

1. **Explore.** Play randomly $\epsilon\%$ of the time. **Only then compute regret.**
2. **Exploit.** Else, play off-policy no regret.

Define **off-policy no regret** for $x$ in round $t$ as

$$\tilde{r}_x^t = \text{Def } \tilde{r}_x^t (E) - \bar{u}^t,$$

where

$$\tilde{r}_x^t (E) = \left[ \frac{1}{|E_x|} \sum_{t \in E_x} u(x^t, y^t) \right]$$

and

$$E_x = \{ t \mid A \text{ experimented in round } t \text{ and played } x \}.$$  

- Proposed as a forecasting heuristic by Foster and Vohra (1993).
- Can be conceived as a way of estimating regrets without knowing, or having to care for, the actions of the opponent.
**ε-Greedy regret matching (outline of proof)**

**Theorem** (Foster et al., 1998). For all $\delta > 0$ there exists an $\epsilon > 0$ such that $\epsilon$-greedy regret matching has at most $\delta$ regret against all realisations of play.

By letting $\epsilon \to 0$ through time at a sufficiently slow rate, one can guarantee there is no regret in the long run.

**Proof.** Suppose there are $k$ different actions. Let $e^t \in R^k$ such that

$$ e^t_x = \begin{cases} 1 : & (A \text{ explores at } t \text{ and chooses } x) \text{ ?} \\ 0 & \end{cases} $$

For each action $i$

$$ \Pr(x^t = i \mid A \text{ explores at round } t) = 1/k. $$

Hence,

$$ E[e^t] = (\epsilon/k, \ldots, \epsilon/k). $$
\( \epsilon \)-Greedy regret matching (outline of proof)

Define

\[
 z_t^x = \text{Def} \left( \frac{k}{\epsilon} \cdot e_t^x \cdot u(x, y^t) \right) - u(x, y^t).
\]

In words, \( z_t^x \) is the difference between (properly magnified) empirical payoff for \( x \) and (correct but) virtual payoff for \( x \). Hence,

\[
 E[z_t^x] = E \left[ \left( \frac{k}{\epsilon} \cdot e_t^x \cdot u(x, y^t) \right) - u(x, y^t) \right]
 = E \left[ \frac{k}{\epsilon} \cdot e_t^x \cdot u(x, y^t) \right] - E[u(x, y^t)]
 = \frac{k}{\epsilon} \cdot E[e_t^x] \cdot E[u(x, y^t)] - E[u(x, y^t)] \quad (e^t \text{ and } u^t \text{ are independent})
 = \frac{k}{\epsilon} \cdot E[u(x, y^t)] - E[u(x, y^t)]
 = 0.
\]
\( \epsilon \)-Greedy regret matching (outline of proof)

**Strong law of large numbers for dependent random variables.** Let \( \{w^t\}^t \) be a bounded sequence of possibly dependent random variables in \( \mathbb{R}^k \). Let 
\[
z^t = E[w^t | w^{t-1}, w^{t-2}, \ldots, w^1] - w^t,
\]
and \( \bar{z}^t \) the average of the \( z^t \)'s. Then
\[
\lim_{t \to \infty} \bar{z}^t = 0 \text{ with probability one.}
\]

We have
\[
z^t_x = \text{Def} \left( \frac{k}{\epsilon} \cdot e^t_x \cdot u(x, y^t) \right) - u(x, y^t).
\]
and
\[
E[z^t_x] = 0.
\]

If
\[
\bar{z}^t = \text{Def} \text{ average of } z^s, s \leq t
\]
then it follows from the above strong law of large numbers that \( \lim_{t \to \infty} \bar{z}^t = 0 \) with probability one. (PY refers to Loève, 1978, Book II, Th. 32.E.1.)
Estimated vs. true regret

Now write $\bar{z}_x^t$ as follows (!):

$$\bar{z}_x^t = \frac{1}{t} \frac{k}{\epsilon} \sum_{s=1}^{t} e_s \cdot u(x, y^s) - \bar{u}^t - \frac{1}{t} \sum_{s=1}^{t} u(x, y^s) - \bar{u}^t$$

\[ \text{regret estimated in experimentation} \]
\[ \text{proper regret} \]

a) The first term approaches regret as estimated in experimental rounds, since $\epsilon t/k \to \infty$ when $t \to \infty$.

b) The second term is true regret.

c) Since we know $\lim_{t \to \infty} \bar{z}^t = 0$, empirical regret converges to true regret almost certainly.

d) $(1 - \epsilon)\%$ of the time $A$ plays estimated regret $\Rightarrow$ true regret.

e) $\epsilon\%$ of the time $A$ explores.

f) In the long run, regret is within $2\epsilon$ of minimal regret.

g) If $\epsilon$ is set to $\delta/2$, the error of regret remains within $2 \cdot \delta/2$. □
### Literature

- No-regret learning can be traced to Blackwell’s approachability theorem and Hannan’s notion of universal consistency.

- The diagram on regret matching is taken from Peyton Young, and Foster and Vohra (who formulate the problem from a decision theoretic point of view).

- The regret-matching algorithm and the analysis of its convergence to correlated equilibria (a generalisation of Nash equilibria) is given by Hart and Mas-Colell.

---


What next?

- **Fictitious Play.** Monitor actions of opponent(s) and play a best response to most frequent actions. As opposed to no-regret, fictitious play is interested in the opponent’s behaviour to predict future play.

- **Smoothed fictitious play.** With fictitious play, the probability to play sub-optimal responses is zero. Smoothed fictitious play plays sub-optimal responses proportional to their expected payoff, given opponents’ play.

- **Conditional no-regret.** Conditions on particular actions. There is regret if there is a pair of actions \((x, x')\) such that, with hindsight, playing \(x'\) was better than playing \(x\).

- **Satisficing Play.** While payoffs equal or supersede the average of past payoffs, keep playing the same action.