School of Economics and Finance

On the Impossibility of Regret Minimization in Repeated Games

Karl Schlag and Andriy Zapechelnyuk

Working Paper No. 676 December 2010 ISSN 1473-0278

On the Impossibility of Regret Minimization in Repeated Games

Karl Schlag
University of Vienna*

Andriy Zapechelnyuk
Queen Mary, University of London†

December 26, 2010

Abstract

Regret minimizing strategies for repeated games have been receiving increasing attention in the literature. These are simple adaptive behavior rules that exhibit nice convergence properties. If all players follow regret minimizing strategies, their average joint play converges to the set of correlated equilibria or to the Hannan set (depending on the notion of regret in use), or even to Nash equilibrium on certain classes of games. In this note we raise the question of validity of the regret minimization objective. By example we show that regret minimization can lead to unrealistic behavior, since it fails to take into account the effect of one's actions on subsequent behavior of the opponents. An amended notion of regret that corrects this defect is not very useful either, since achieving a no-regret objective is not guaranteed in that case.

Keywords: Repeated games, regret minimization, no-regret strategy

JEL classification numbers: C73, D81

1. Introduction. In a repeated interaction, an individual follows a *regret-minimizing* strategy if, loosely speaking, she reinforces those actions that she regrets not having played enough in the past. A particularly simple strategy is *regret matching*, which is defined by the following rule:

*Department of Economics, University of Vienna, Hohenstaufengasse 9, 1010 Vienna, Austria. *E-mail:* karl.schlag $\alpha\tau$ univie.ac.at

†*Corresponding author.* School of Economics and Finance, Queen Mary, University of London, Mile End Road, London E1 4NS, UK. *E-mail:* a.zapechelnyuk $\alpha\tau$ qmul.ac.uk

> Switch next period to a different action with a probability that is *proportional* to the *regret* for that action, where regret is defined as the increase in payoff had such a change always been made in the past (Hart and Mas-Colell, 2000; Hart, 2005).

This strategy, in particular, includes rules of thumb which act according to "Never change a winning team," in other words, do not switch to a different action, as long as the current action keeps being a best reply to the observed (average) actions of the opponents.

Regret-minimizing strategies received a lot of attention in the recent literature.[1] The main value of these strategies is that they are simple adaptive behavior rules that are neither computationally demanding nor relying on common knowledge assumptions and yet exhibiting nice convergence properties. If all players follow regret-minimizing strategies, their average joint play converges to the set of correlated equilibria or to the Hannan set[2], depending on the notion of regret in use (Hart and Mas-Colell 2000; see also Lehrer 2003, Cesa-Bianchi and Lugosi 2006); or even to Nash equilibria on certain classes of games (Hart and Mas-Colell 2003; Marden, Arslan, and Shamma 2007).

In this note we raise the question of validity of the regret minimization objective in the context of games. On the one hand, according to the notions of regret used in the literature, an individual who contemplates whether she could have done better by having played a particular action more often in the past does not take into account the effect of her actions on the subsequent behavior of her opponent. This is perfectly fine in a decision making environment, but *not* in a game, where, by definition, players are responsive to the opponents' behavior. We show by example that failure to take the opponent's responsiveness into account may lead to unrealistic behavior.[3]

On the other hand, if we extend the notion of regret to take into account the above mentioned effect, then it becomes impossible to guarantee no regrets, even against a severely restricted set of the opponent's strategies. We show that if opponent is even

[1] A non-exhaustive list includes Littlestone and Warmuth (1994), Fudenberg and Levine (1995), Foster and Vohra (1998), Foster and Vohra (1999), Freund and Schapire (1999), Hart and Mas-Colell (2000), Hart and Mas-Colell (2001), Hart and Mas-Colell (2003), Lehrer (2003), Young (2004), Cesa-Bianchi and Lugosi (2003), Cesa-Bianchi and Lugosi (2006), Lehrer and Solan (2009).

[2] The Hannan set of a game is the set of all mixed action profiles that satisfy Hannan's (1957) no-regret condition. It is also known as the set of *coarse correlated equilibria* first appeared in Moulin and Vial (1978), but explicitly defined as a solution concept by Young (2004, Ch.3).

[3] This problem is recognized in the computer science literature, e.g., Farias and Megiddo (2004) and Cesa-Bianchi and Lugosi (2006, Ch.7.11). These works show that regret minimizing strategies fail to lead to the cooperative outcome in a repeated prisoner's dilemma. Our example is different and, as we believe, has a value on its own, as it illuminates failure to learn the Pareto dominant equilibrium of a *one-shot game*, whereas the above literature shows failure to learn playing strictly dominated actions.

"slightly" adaptive to the player's past behavior, the "no-regrets" objective cannot be achieved.

We conclude that regret-minimizing behavior rules, either with the original notion of regret or with the one that takes into account the opponent's reaction, are not very appealing when describing behavior of real subjects in repeated interactions where one's past actions may affect the other's reaction, in particular, in repeated games.

2. Regrets. Consider a finite two-player game, with players named Alice and Bob.[4] Let A and B be sets of actions of Alice and Bob, respectively, and let $u : A \times B \to \mathbb{R}$ be Alice's payoff function. The game is played repeatedly in time periods $t = 1, 2, \ldots$, in which players choose actions (a_t, b_t). The history of realized actions is observable for both players.

Consider Alice before period $T+1$ and let $a^* = a_T$ be her most recent action. Denote by $\bar{U}_T$ the average payoff of Alice up to period T,

$$\bar{U}_T = \frac{1}{T} \sum_{t=1}^{T} u(a_t, b_t),$$

and denote by $U_T(a')$ the average payoff that Alice would have obtained had she played a' instead of the reference action a^* every time in the past when she actually played a^*,

$$U_T(a') = \frac{1}{T} \sum_{t=1}^{T} w_t(a'),$$

where

$$w_t(a') = \begin{cases} u(a', b_t), & \text{if } a_t = a^*, \\ u(a_t, b_t), & \text{if } a_t \neq a^*. \end{cases}$$

Alice's *regret* $r_T(a')$ for action a' after T periods is defined as the excess of $U_T(a')$ over $\bar{U}_T$,

$$r_T(a') = U_T(a') - \bar{U}_T.$$

According to the above definition, Alice evaluates her regret for some action a' relative to the reference action a^* (the most recently played one) by contemplating how much higher payoff, on average, she could have received had she played a' in every past period when she actually played a^*, assuming that *the play of the opponents would have remained*

[4]Bob can be considered as a set of players, so the arguments presented below trivially extend to n-player games.

unchanged. This definition is plausible in the context of decision making, when an individual's actions have no effect on the opponent, who can be perceived as an abstract environment. It is much less appealing if the individual is engaged in a game, where the opponent's future play can be responsive to the individual's present actions.

	Bob	
Alice	L	R
L	$1, 1$	$0, 0$
R	$0, 0$	$100, 100$

Figure 1

3. An example. For illustration, consider the following coordination game (Fig. 1). Suppose that the observed play up to period T is $((a_1, b_1), (a_2, b_2), \ldots, (a_T, b_T)) = ((L, L), (L, L), \ldots, (L, L))$. Given this history, from the perspective of Alice, playing L is a best reply to the average *realized* play of Bob.

Does Alice have regret for action R? Not according to the above definition, since in every period she would have miscoordinated with Bob, so $r_T(R) = -1$.

Could Alice have done better by having switched to R?

(I) No, if Bob's strategy is *independent* of Alice's actions.

(II) Possibly, if Bob's strategy is *adaptive*, so that Bob could have followed Alice after observing her trying to coordinate on a Pareto superior outcome.

Since Alice and Bob are engaged in a game, it would be unrealistic to assume that Bob ignores all information obtained during past play. In games, scenario (II) is much more plausible. Thus, the described notion of regret is not very appealing, and behavior rules based on this notion could lead to outcomes that are unlikely to occur in interactions of real subjects.

4. Regrets against history dependent behavior. Let us now introduce a different notion of regret that accounts for the opponent's reaction. Denote by $h_T = ((a_1, b_1), \ldots, (a_T, b_T))$ the history of play up to T, and let $\mathcal{H}$ be the set of all finite histories. Let $\alpha : \mathcal{H} \to \Delta(A)$ and $\beta : \mathcal{H} \to \Delta(B)$ be strategies of Alice and Bob, respectively, that prescribe mixed actions for every history $h_t \in \mathcal{H}$. Denote by $U_T(\alpha, \beta)$ the expected

average payoff of Alice up to period T when she plays α against Bob playing β,

$$U_T(\alpha,\beta) = E_{(\alpha,\beta)}\left[\frac{1}{T}\sum_{t=1}^{T} u(a_t,b_t)\right],$$

where the expectation is taken with respect to the probability measure over $\mathcal{H}$ induced by (α,β).

Fix Alice's strategy α. Denote by $\alpha_{(a^*|a')}$ the strategy obtained from α by replacement of a^* by a' in all periods where the realized action of α is a^*. Formally, for every history $h \in \mathcal{H}$ let

$$\begin{aligned}\alpha_{(a^*|a')}(h)[a^*] &= 0, \quad \text{and}\\ \alpha_{(a^*|a')}(h)[a'] &= \alpha(h)[a^*] + \alpha(h)[a'],\end{aligned}$$

where $\alpha(h)[k]$ denotes the probability that $\alpha(h)$ assigns to action $k \in A$.

Consider Alice before period $T+1$ and let $a^* = a_T$ be her most recent action. For a given strategy β of Bob, $U_T(\alpha_{(a^*|a')},\beta)$ is the expected average payoff that Alice would have obtained had she played a' every time in the past when her strategy α stipulated to play a^*, and when at every stage $t \le T$ Bob would have responded according to β to the new history.

Let $\mathcal{B}$ be a set Bob's feasible strategies. Then Alice's regret for a' is given by

$$\rho_T(a') = \max_{\beta\in\mathcal{B}} U_T(\alpha_{(a^*|a')},\beta) - \bar{U}_T.$$

Thus, if $\rho_T(a') \le 0$, then Alice can conclude that she could not have done better by switching a^* to a' in the past, no matter what is the actual strategy of Bob.

A strategy of Alice is called a *no-regret strategy against* $\mathcal{B}$ if it guarantees that Alice's regrets become non-positive in the limit for every Bob's strategy in $\mathcal{B}$,

$$\limsup_{T\to\infty}\left\{\max_{a'\in A}\rho_T(a')\right\} \le 0 \quad \text{with probability one.}$$

It is known that there exist no-regret strategies against an irresponsive opponent, i.e., when $\mathcal{B}$ contains only deterministic sequences (e.g., Hannan, 1957; Hart and Mas-Colell, 2000, 2001; Cesa-Bianchi and Lugosi, 2003). Yet, as we show below, a minimum of adaptiveness of Bob's strategies to Alice's past actions leads to an impossibility result.

Bob's strategy is called *q-fictitious play* if in every period $t = 2,3,\ldots$, with probability

$1-q$ Bob repeats his last-period action, and with probability q he best-replies to Alice's average past play. The initial play of Bob is arbitrary.

For some $\varepsilon > 0$ denote by $\mathcal{B}_\varepsilon$ the set of q-fictitious play strategies with $q \in [0, \varepsilon]$. In particular, $\mathcal{B}_\varepsilon$ contains non-adaptive strategies where Bob plays a constant action (0-fictitious play).

Proposition. *For every $\varepsilon > 0$, there does not exist a no-regret strategy against $\mathcal{B}_\varepsilon$.*

Proof. The proof is by example. Consider the coordination game described earlier (Fig. 1). Fix $\varepsilon > 0$ and suppose that Bob plays q-fictitious play, $\beta^q \in \mathcal{B}_\varepsilon$, $q \in [0, \varepsilon]$, and let his initial action be L.

Observe that the only possible source of regret for Alice is her inability to distinguish the case of $q = 0$ from $q > 0$. Indeed, if Alice knew that $q = 0$, then her best reply would be to always play L, since Bob is non-adaptive and repeats L forever, so $\bar{U}_T = U_T(L) \to 1$. On the other hand, if she knew that $q > 0$, then her best reply would be to always play R, since eventually, with probability 1, Bob would switch to R after observing Alice's past average play being R, and the further play would be locked in (R, R) forever, so $\bar{U}_T = U_T(R) \to 100$.

Denote by z_t the frequency of R in Alice's past actions, $z_t = \frac{1}{t}|\{k \leq t : a_k = R\}|$. Then in every period $t \geq 2$, with probability q Bob plays R if $z_t > 1/100$ and plays L otherwise (the tie can be resolved arbitrarily); with probability $1-q$ Bob repeats his last-period action. Consider the subsequence of periods, $\{t_s\}_{s=1}^S$, where the event $\{z_{t_s} > 1/100\}$ occurs. For every s, the probability that Bob has never played R up to t_s is equal to $(1-q)^s$. First, suppose that S is finite. Then for $q > 0$ Alice's regret for action R is at least

$$\lim_{T\to\infty} U_T(R) - \bar{U}_T \geq 100 - [(1-(1-q)^S) \cdot 100 + (1-q)^S \cdot 1] = 99(1-q)^S,$$

which, for any given S, is bounded away from zero for every small enough q.

Next, let $S = \infty$. Then $\lim_{s\to\infty}(1-q)^s \to 0$ if and only if $q > 0$. So, for every $q > 0$, $\lim_{T\to\infty} U_T(R) - \bar{U}_T \to 0$, and Alice has no regrets. However, for $q = 0$, this strategy of Alice requires $z_{t_s} > 1/100$ for every $s = 1, 2, \ldots$, and hence

$$\bar{U}_{t_s} = z_{t_s} \cdot 0 + (1 - z_{t_s}) \cdot 1 < 99/100,$$

while $U_{t_s}(L) = 1$. Thus, on the subsequence $\{t_s\}$ of periods, the regret for L is at least $1/100$.

It follows that no matter what Alice plays, there exists a strategy in $\mathcal{B}^*$ of Bob such that lim sup of Alice's regret for one of the actions is bounded away from zero. $\square$

The reason for this negative result is that the probability that Bob's type ($q = 0$ or $q > 0$) is revealed does not converge to one uniformly across $\mathcal{B}_\varepsilon$, as $T \to \infty$. That is, after every T, if Bob has never played R so far, there is no upper bound on Alice's posterior belief that $q = 0$. In other words, Alice cannot distinguish the cases $q = 0$ and $q > 0$, no matter how long she observes Bob's behavior.

	Bob		
Alice	L	M	R
L	$2, 2$	$1, 1$	$0, 0$
R	$0, 0$	$1, 1$	$2, 2$

Figure 2

Another example is less subtle and shows that *no regret* cannot be achieved if an opponent uses *trigger strategies*. Consider the game on Fig. 2 and suppose that the set of strategies of Bob includes the following:

(*non-adaptive*) Bob constantly plays M.

(*adaptive-*L) Bob starts with M. Then, if Alice played L in the initial period, then Bob will play L from period 2 forever, otherwise he will play M forever.

(*adaptive-*R) the same as *adaptive-*L except L is replaced by R.

In this game, Alice's long-run average payoff is determined entirely by Bob's type and Alice's initial action, since Bob's actions are constant from period 2 on. Now observe that no matter what Alice plays in period 1, there is a type of Bob, either *adaptive-*L or *adaptive-*R, that would make her regret for action L or R, respectively, in all subsequent periods. Indeed, if Alice chooses, for instance, L in the first period and Bob's type is *adaptive-*R, then the following play of Bob will be constantly M, and Alice's average payoff will be 1. However, Alice could have obtained the average payoff of 2, had she started her play with R.

5. Conclusion. To sum up, the notion of regret used in the literature is not satisfactory in the context of repeated games as it fails to take into account possible reaction of opponents to changes in one's actions. We define an extended notion of regret, with

respect to opponents' strategies (rather than realized actions) and show that in this case no-regret strategies need not exist when the opponents are adaptive. Two examples provide the intuition for this result: the regrets persist because the opponent's strategy cannot be statistically identified (as in the former example) or because the opponent uses *trigger* strategies, where an early decision of the player (which is payoff-relevant for the the entire infinitely repeated interaction) has to be made when the player has not been yet informed about the opponent's strategy.

We conclude that the notion of regret, whether the original or the extended one, should be used with caution in the context of repeated games where players may respond to one another's behavior.

References

Cesa-Bianchi, N. and G. Lugosi (2003). Potential-based algorithms in on-line prediction and game theory. *Machine Learning 51*, 239–261.

Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Farias, D. P. and N. Megiddo (2004). How to combine expert (or novice) advice when actions impact the environment. In *Advances in Neural Information Processing Systems*, Volume 16. MIT Press.

Foster, D. and R. Vohra (1998). Asymptotic calibration. *Biometrika 85*, 379–390.

Foster, D. and R. Vohra (1999). Regret in the online decision problem. *Games and Economic Behavior 29*, 7–35.

Freund, Y. and R. Schapire (1999). Adaptive game playing using multiplicative weights. *Games and Economic Behavior 29*, 79–103.

Fudenberg, D. and D. Levine (1995). Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control 19*, 1065–1089.

Hannan, J. (1957). Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe (Eds.), *Contributions to the Theory of Games, Vol. III*, Annals of Mathematics Studies 39, pp. 97–139. Princeton University Press.

Hart, S. (2005). Adaptive heuristics. *Econometrica 73*, 1401–1430.

Hart, S. and A. Mas-Colell (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica 68*, 1127–1150.

Hart, S. and A. Mas-Colell (2001). A general class of adaptive procedures. *Journal of Economic Theory 98*, 26–54.

Hart, S. and A. Mas-Colell (2003). Continuous-time regret-based dynamics. *Games and Economic Behavior 45*, 375–394.

Lehrer, E. (2003). A wide range no-regret theorem. *Games and Economic Behavior 42*, 101–115.

Lehrer, E. and E. Solan (2009). Approachability with bounded memory. *Games and Economic Behavior 66*, 995–1004.

Littlestone, N. and M. Warmuth (1994). The weighted majority algorithm. *Information and Computation 108*, 212–261.

Marden, J. R., G. Arslan, and J. S. Shamma (2007). Regret based dynamics: convergence in weakly acyclic games. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS07)*, Honolulu, Hawaii, USA, pp. 194–201.

Moulin, H. and J. P. Vial (1978). Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory 7*, 201–221.

Young, H. P. (2004). *Strategic Learning and Its Limits*. Oxford University Press.

This working paper has been produced by
the School of Economics and Finance at
Queen Mary, University of London

School of Economics and Finance
Queen Mary, University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 5096
Fax: +44 (0)20 8983 3580
Web: www.econ.qmul.ac.uk/papers/wp.htm