Forthcoming in the *New Palgrave Dictionary of Economics, Second Edition,* edited by Steven N. Durlauf and Lawrence E. Blume. London: Macmillan.

adaptive heuristics

A 'heuristic' is a method or rule for solving problems; in game theory it refers to a method for learning how to play. Such a rule is 'adaptive' if it is directed towards higher payoffs and is reasonably simple to implement. This article discusses a variety of such rules and the forms of equilibrium that they implement. It turns out that even sophisticated solution concepts, like subgame perfect equilibrium, can be achieved by relatively simple and intuitive methods.

'Adaptive heuristics' are simple behavioural rules that are directed towards payoff improvement but may be less than fully rational. The number and variety of such rules are virtually unlimited; here we survey several prominent examples drawn from psychology, computer science, statistics and game theory. Of particular interest are the informational inputs required by different learning rules and the forms of equilibrium to which they lead. We shall begin by considering very primitive heuristics, such as reinforcement learning, and work our way up to more complex forms, such as hypothesis testing, which still, however, fall well short of perfectly rational learning.

One of the simplest examples of a learning heuristic is *cumulative payoff matching*, in which the subject plays actions next period with probabilities proportional to their cumulative payoffs to date. Specifically, consider a finite stage game G that is played infinitely often, where all payoffs are assumed to be strictly positive. Let $a_{ij}(t)$ denote the cumulative payoff to player i over all those periods $0 \leq t' \leq t$ when he played action j, including some *initial propensity* $a_{ij}(0) > 0$. The cumulative payoff matching rule stipulates that in period $t+1$, player i chooses action j with probability

$$p_{ij}(t+1) = a_{ij}(t) \Big/ \sum_k a_{ik}(t). \tag{1}$$

Notice that the distribution has full support given the assumption that the initial propensities are positive. This idea was first proposed by the psychologist Nathan Herrnstein (1970) to explain certain types of animal behaviour, and falls under the more general rubric of *reinforcement learning* (Bush and Mosteller, 1951; Suppes and Atkinson, 1960; Cross, 1983). The key feature of a reinforcement model is that the probability of choosing an action increases monotonically with the total payoff it has generated in the past (on the assumption that the payoffs are positive). In other words, taking an action and receiving a positive payoff *reinforces* the tendency to take that same action again. This means, in particular, that play can become concentrated on certain actions simply because they were played early and often, that is, play can be *habit-forming* (Roth and Er'ev, 1995; Er'ev and Roth, 1998).

Reinforcement models differ in various details that materially affect their theoretical behaviour as well as their empirical plausibility. Under cumulative payoff matching, for example, the payoffs are not discounted, which means that current payoffs have an impact on current behaviour that diminishes as $1/t$. Laboratory experiments suggest, however, that recent payoffs matter more than those long past (Er'ev and Roth, 1998); furthermore, the rate of discounting has implications for the asymptotic properties of such models (Arthur, 1991).

Another variation in this class of models relies on the concept of an *aspiration level*. This is a level of payoffs, sometimes endogenously determined by past play, that triggers a change in a player's behaviour when current payoffs fall below the level and inertial behaviour when payoffs are above the

level. The theoretical properties of these models have been studied for 2×2 games, but relatively little is known about their behaviour in general games (Börgers and Sarin, 2000; Cho and Matsui, 2005).

Next we turn to a class of adaptive heuristics based on the notion of minimizing *regret*, about which more is known in a theoretical sense. Fix a particular player and let $\alpha(t)$ denote the average per period payoff that she received over all periods $t' \leq t$. Let $\alpha_j(t)$ denote the average payoff she *would have* received by playing action j in every period through t, on the assumption that the opponents played as they actually did. The difference $r_j(t) = \alpha_j(t) - \alpha(t)$ is the subject's *unconditional regret* from not having played j in every period through t. (In the computer science literature this is known as *external regret*; see Greenwald and Gondek, 2002.)

The following simple heuristic was proposed by Hart and Mas-Colell (2000; 2001) and is known as *unconditional regret matching*: play each action with a probability that is proportional to the positive part of its unconditional regret, that is,

$$p_j(t+1) = [r_j(t)]_+ \Big/ \sum\nolimits_k [r_k(t)]_+ . \tag{2}$$

This learning rule has the following remarkable property: when used by any one player, his regrets become non-positive almost surely as t goes to infinity *irrespective of the behaviour of the other players*. When all players use the rule, their time average behaviour converges almost surely to a generalization of correlated equilibrium known as the *Hannan set* or the *coarse correlated equilibrium set* (Hannan, 1957; Moulin and Vial, 1978; Hart and Mas-Colell, 2000; Young, 2004). In general, a *coarse correlated equilibrium* (CCE) is a probability distribution over outcomes (joint actions) such that, given a choice between (*a*) committing *ex ante* to whatever joint action will be realized, and (*b*) committing *ex ante* to a fixed action, given that the others are committed to playing their part of whatever joint action will be realized, every player weakly prefers the former option. By contrast, a *correlated equilibrium* (CE) is a distribution such that, after a player's part of the realized joint action has been disclosed, he would just as soon play it as something else, given that the others are going to play their part of the realized joint action. It is straightforward to show that the coarse correlated equilibria form a convex set that contains the set of correlated equilibria (Young, 2004, ch. 3).

The heuristic specified in (2) belongs to a large family of rules whose time-average behaviour converges almost surely to the coarse correlated equilibrium set; equivalently, that assures no long-run regret for all players simultaneously. For example, this property holds if we let $p_j(t+1) = [r_j(t)]_+^\theta / \sum_k [r_k(t)]_+^\theta$ for some exponent $\theta > 0$; one may even take different exponents for different players. Notice that these heuristics put positive probability only on actions that would have done strictly better (on average) than the player's realized average payoff. These are sometimes called *better reply rules*. Fictitious play, by contrast, puts positive probability only on action(s) that would have done *best* against the opponents' frequency distribution of play.

Fictitious play does not necessarily converge to the coarse correlated equilibrium set (CCES); indeed, in some 2×2 coordination games fictitious play causes perpetual miscoordination, in which case both players have unconditional long-run regret (Fudenberg and Kreps, 1993; Young, 1993). By choosing θ to be very large, however, we see that there exist better reply rules that are arbitrarily close to fictitious play and that do converge almost surely to the CCES. Fudenberg and Levine (1995; 1998; 1999) and Hart and Mas-

Colell (2001) give general conditions under which stochastic forms of fictitious play converge in time average to the CCES.

Without complicating the adjustment process too much, one can construct rules whose time average behaviour converges almost surely to the *correlated equilibrium set* (CES). To define this class of heuristics we need to introduce the notion of conditional regret. Given a history of play through time t and a player i, consider the change in per period payoff if i had played action k in all those periods $t' \leq t$ when he actually played action j (and the opponents played what they did). If the difference is positive, player i has conditional regret – he wishes he had played k instead of j. Formally, i's *conditional regret* at playing j instead of k up through time t, $r^i_{jk}(t)$, is $1/t$ times the increase in payoff that would have resulted from playing k instead of j in all periods $t' \leq t$. Notice that the average is taken over all t periods to date; hence, if j was not played very often, $r^i_{jk}(t)$ will be small.

Consider the following *conditional regret matching* heuristic proposed by Hart and Mas-Colell (2000): if a given agent played action j in period t, then in period $t+1$ he plays according to the distribution

$$q_k(t+1) = \varepsilon r_{jk}(t)_+ \text{ for all } k \neq j, \text{ and } q_j(t+1) = 1 - \varepsilon \sum_{k \neq j} r_{jk}(t)_+. \quad (3)$$

In effect $1-\varepsilon$ is the degree of inertia, which must be large enough that $q_k(t+1)$ is non-negative for all realizations of the conditional regrets $r_{jk}(t)$. If all players use conditional regret matching and ε is sufficiently small, then almost surely the joint frequency of play converges to the set of correlated equilibria (Hart and Mas-Colell, 2000). Notice that *pointwise* convergence is not guaranteed; the result says only that the empirical distribution converges to a convex *set*. In particular, the players' time-average behaviour may wander from one correlated equilibrium to another. It should also be remarked that, if a single player uses conditional regret matching, there is no assurance that his conditional regrets will become non-positive over time unless we assume that the other players use the same rule. This stands in contrast to unconditional regret matching, which assures non-positive unconditional regret for any player who uses it irrespective of the behaviour of the other players. One can, however, design more sophisticated updating procedures that unilaterally assure no conditional regret; see for example Foster and Vohra (1999), Fudenberg and Levine (1998, ch. 4), Hart and Mas-Colell (2000), and Young (2004, ch. 4).

A natural question now arises: do there exist simple heuristics that allow the players to learn *Nash* equilibrium instead of correlated or still coarser forms of equilibrium? The answer depends on how demanding we are about the long-run convergence properties of the learning dynamic. Notice that the preceding results on regret matching were concerned solely with time-average behaviour; no claim was made that period-by-period behaviour converges to any notion of equilibrium. Yet surely it is period-by-period behaviour that is most relevant if we want to assert that the players have 'learned' to play equilibrium. It turns out that it is very difficult to design adaptive learning rules under which period-by-period behaviour converges almost surely to Nash equilibrium in any finite game, unless one builds in some form of coordination among the players (Hart and Mas-Colell, 2003; 2006). The situation becomes even more problematic if one insists on fully rational, Bayesian learning. In this case it can be shown that there exist games of incomplete information in which no form of Bayesian rational learning causes period-by-period behaviours to come close to Nash equilibrium behaviour even in a probabilistic sense (Jordan, 1991, 1993; Foster and Young, 2001; Young, 2004; see also BELIEF LEARNING).

If one does not insist on full rationality, however, one can design stochastic adaptive heuristics that cause period-by-period behaviours to come close to Nash equilibrium – indeed close to subgame perfect equilibrium – most of the time (without necessarily *converging* to an equilibrium). Here is one approach due to Foster and Young (2003); for related work see Foster and Young (2006) and Germano and Lugosi (2007). Let G be a finite n-person game that is played infinitely often. At each point in time, each player thinks that the others are playing i.i.d. strategies. Specifically, at time t player i thinks that j is playing the i.i.d strategy $p_j(t)$ on j's action space, and that the opponents are playing independently; that is, their joint strategies are given by the product distribution $\mathrm{p}_{-i}(t) = \prod_{j \neq i} p_j(t)$. Suppose that i's best response is to play a smoothed best response to $p_{-i}(t)$. Specifically, assume that i plays each action j with a probability proportional to $e^{\beta u_i(j, p_{-i})}$, where $u_i(j, p_{-i})$ is i's expected utility from playing j in every period when the opponents play p_{-i}, and $\beta > 0$ is a *response parameter*. This is known as a *quantal* or *log linear* response function. For brevity, denote i's response in period t by $q_i^{\beta}(t)$; this depends, of course, on $p_{-i}(t)$. Player i views $p_{-i}(t)$ as a hypothesis that he wishes to test against data. After first adopting this hypothesis he waits for a number of periods (say s) while he observes the opponents' behaviour, all the while playing $q_i^{\beta}(t)$. After s periods have elapsed, he compares the empirical frequency distribution of the opponents' play during these periods with his hypothesis. Notice that both the empirical frequency distribution and the hypothesized distribution lie in the same compact subset of Euclidean space. If the two differ by more than some tolerance level τ (in the Euclidean metric), he rejects his current hypothesis and chooses a new one.

In choosing a new hypothesis, he may wish to take account of information revealed during the course of play, but we shall also assume he engages in some *experimentation*. Specifically, let us suppose that he chooses a new hypothesis according to a probability density that is uniformly bounded away from zero on the space of hypotheses. One can show the following: given any $\varepsilon > 0$, if the response parameter β is sufficiently large, the test tolerance τ is sufficiently small (given β), and the amount of data collected s is sufficiently large (given β and τ), then the players' *period-by-period* behaviours constitute an ε-equilibrium of the stage game G at least $1 - \varepsilon$ of the time (Foster and Young, 2003). In other words, classical statistical hypothesis testing is a heuristic for learning Nash equilibria of the stage game. Moreover, if the players adopt hypotheses that condition on history, they can learn complex equilibria of the repeated game, including forms of subgame perfect equilibrium.

The theoretical literature on strategic learning has advanced rapidly in recent years. A much richer class of learning models has been identified since the mid-1990s, and more is known about their long-run convergence properties. There is also a greater understanding of the various kinds of equilibrium that different forms of learning deliver. An important open question is how these theoretical proposals relate to the empirical behaviour of laboratory subjects. While there is no reason to think that any of these rules can fully explain subjects' behaviour, they can nevertheless play a useful role by identifying phenomena that experimentalists should look for. In particular, the preceding discussion suggests that weaker forms of equilibrium may turn out to be more robust predictors of long-run behaviour than is Nash equilibrium.

H. Peyton Young

See also

behavioural game theory;
belief learning;
learning.

Bibliography

Arthur, W.B. 1991. Designing agents that act like human agents: a behavioral approach to bounded rationality. *American Economic Association, Papers and Proceedings* 81, 353–9.

Börgers, T. and Sarin, R. 2000. Naïve reinforcement learning with endogenous aspirations. *International Economic Review* 31, 921–50.

Bush, R.R. and Mosteller, F. 1951. A mathematical model for simple learning. *Psychological Review* 58, 313–23.

Cho, I.-K. and Matsui, A. 2005. Learning aspiration in repeated games. *Journal of Economic Theory* 124, 171–201.

Cross, J. 1983. *A Theory of Adaptive Economic Behavior*. Cambridge: Cambridge University Press.

Er'ev, I. and Roth, A.E. 1998. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* 88, 848–81.

Foster, D.P. and Vohra, R. 1999. Regret in the on-line decision problem. *Games and Economic Behavior* 29, 7–35.

Foster, D.P. and Young, H.P. 2001. On the impossibility of predicting the behavior of rational agents. *Proceedings of the National Academy of Sciences of the USA* 98(222), 12848–53.

Foster, D.P. and Young, H.P. 2003. Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior* 45, 73–96.

Foster, D.P. and Young, H.P. 2006. Regret testing: learning Nash equilibrium without knowing you have an opponent. *Theoretical Economics* 1, 341–67.

Fudenberg, D. and Kreps, D. 1993. Learning mixed equilibria. *Games and Economic Behavior* 5, 320–67.

Fudenberg, D. and Levine, D. 1995. Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control* 19, 1065–90.

Fudenberg, D. and Levine, D. 1998. *The Theory of Learning in Games*. Cambridge MA: MIT Press.

Fudenberg, D. and Levine, D. 1999. Conditional universal consistency. *Games and Economic Behavior* 29, 104–30.

Germano, F. and Lugosi, G. 2007. Global Nash convergence of Foster and Young's regret testing. *Games and Economic Behavior* (forthcoming).

Greenwald, A. and Gondek, D. 2002. On no-regret learning and game-theoretic equilibria. *Journal of Machine Learning* 1, 1–20.

Hannan, J. 1957. Approximation to Bayes risk in repeated plays. In *Contributions to the Theory of Games*, vol. 3, eds. M. Dresher, A.W. Tucker and P. Wolfe. Princeton, NJ: Princeton University Press.

Hart, S. and Mas-Colell, A. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68, 1127–50.

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

Hart, S. and Mas-Colell, A. 2003. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review* 93, 1830–6.

Hart, S. and Mas-Colell, A. 2006. Stochastic uncoupled dynamics and Nash equilibrium. *Games and Economic Behavior* 57, 286–303.

Herrnstein, R.J. 1970. On the law of effect. *Journal of the Experimental Analysis of Behavior* 13, 243–66.

Jordan, J.S. 1991. Bayesian learning in normal form games. *Games and Economic Behavior* 3, 60–81.

Jordan, J.S. 1993. Three problems in learning mixed-strategy equilibria. *Games and Economic Behavior* 5, 368–86.

Moulin, H. and Vial, J.P. 1978. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory* 7, 201–21.

Roth, A.E. and Er'ev, I. 1995. Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior* 8, 164–212.

Suppes, P. and Atkinson, R. 1960. *Markov Learning Models for Multiperson Interaction*. Stanford CA: Stanford University Press.

Young, H.P. 1993. The evolution of conventions. *Econometrica* 61, 57–84.

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