Innovations in Computer Science 2010

Playing Games without Observing Payoffs

Michal Feldman[1,2] Adam Kalai[3] Moshe Tennenholtz[2,4]
[1]School of Business Administration and Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem, Israel.
[2]Microsoft Israel R&D Center.
[3]Microsoft Research, Cambridge, MA.
[4]Technion-Israel Institute of Technology, Haifa, Israel.
mfeldman@huji.ac.il adum@microsoft.com moshet@microsoft.com

Abstract: Optimization under uncertainty is central to a number of fields, and it is by now well-known that one can learn to play a repeated game without *a priori* knowledge of the game, as long as one observes the payoffs (even if one does not see the opponent's play). We consider the complementary scenario in which one does not observe the payoffs but does observe the opponent's play. Curiously, we show that for an interesting class of games one can still learn to play well. In particular, for any symmetric two-person game, one can nearly match the payoff of the opponent (who may have full knowledge of the game), without ever observing a single payoff. The approach employed is a familiar one: attempt to mimic the opponent's play. However, one has to be careful about how one mimics an opponent who may know that they are being mimicked.
This paper has two contributions: it (a) extends our understanding of optimization under uncertainty by modeling a new setting in which one can play games optimally; and (b) introduces a new algorithm for being a copycat, one which is strategically sound even against an opponent with a superior informational advantage.

Keywords: two-player symmetric games; learning; information asymmetry; unobservable payoffs

1 Introduction

Being a copycat is a fundamental strategy well-known to anyone who has ever found themselves in an unfamiliar environment. In online arenas where advertising decisions can be changed daily, companies often mimic the advertising campaign of a more experienced rival. Since measuring the effect of an advertising campaign is notoriously difficult, a less experienced company may benefit from observing the choices of an experienced opponent. The newcomer cannot afford to invest in research or wait until they learn consumer behavior; it needs to function effectively when competing with the existing, well-informed company. However, in a competitive environment, such behavior must be executed with great care. A good copycat can reap tremendous rewards with little experience and little research investments. A poor copycat may perform worse than making random decisions.

We model this phenomenon as a game against an opponent whose interests are not necessarily aligned with yours. The model is that of a two-person game that is repeated T times. For simplicity, we begin by assuming that the underlying game is zero-sum, i.e., the two players have interests that are completely opposed. (We later discuss general-sum games.) In the well-studied *bandit* scenario (see, e.g., Cesa-Bianchi and Lugosi [5]), a player observes her own payoff each period and nothing else – she does not observe the actions chosen by the opponent. Auer, *et al.* [1] gave an algorithm that guarantees a player her safety level (mixed-strategy min-max value) in such a context, even with worst-case assumptions on the game and the adversary. The guaranteed value is the maximum possible guarantee, even if the player knew the entire game, and in the case of a zero-sum game it is the classic von Neumann [9] min-max value. Moreover, the issues dealt with here are not merely theoretical – they reflect the tradeoff between exploration and exploitation that occurs in real life.

In contrast, little or no attention has been given to strategies that ignore payoffs and only observe the opponent's play. One reason is because, in the general case, it is impossible to perform well without any feedback besides the opponent's play. However, in the case of symmetric games – and a good many games are symmetric – one may be able to learn only from observing the opponent's play. We give a strategy for playing a repeated symmetric two-person game using only the history of observed play of the opponent. The strategy in no way depends on prior knowledge of the game (not even the number of strategies). No assumptions are made about the opponent. Hence

the strategy may, in extreme cases, even be used by a player that is completely oblivious to payoffs, one who never observes a single payoff, against an opponent that has complete knowledge of the game.

While two-player symmetric games are a central object of study in game theory, our model of extreme informational asymmetry is new. Of course, such extremes are less common than other situations. But the point here is to test the limits of how much feedback a player needs to play a game. If the informed player has less knowledge, or the uninformed player has more feedback, the algorithm and analysis can still of course be applied. In cases where determining the payoff is difficult or costly, it is good to know that it is possible to simply copy, even in an adversarial setting. While this payoff-oblivious setting is less common than bandit settings, we do feel that there are sufficient applications to merit study. Furthermore, understanding this setting helps expand our general understanding of optimization and game play under uncertainty.

1.1 How To Be a *Bad* Copycat

At first, it may seem that being a copycat is easy. This is true in the case of optimization in an environment that does not react to your decisions. If other people are maximizing $f(x)$, then copying their choices may maximize $f(x)$ as well. For example, one may simply buy the car that an informed friend buys. In a competitive environment, however, it is not that easy.

The first try is to simply copy the opponent's play on the previous round. In the game Rock, Paper, Scissors (RPS), this will lose every round if the opponent cycles: $R, P, S, R, P, S, \ldots$, because it is always one step behind.

The second try is to copy the opponent's empirical frequency of play, akin to fictitious play [4]. Namely, copy the play of the opponent from a random previous round. This will also fail badly in RPS, for example if the opponent plays R for $T/2$ periods followed by P for $T/2$ periods.

1.2 The Copycat Strategy

Suppose player 1 never observes a single payoff and knows nothing about the game except the n available actions. Player 2, on the other hand, may have complete prior knowledge of the payoff matrix, $A \in \mathbb{R}^{n\times n}$, where $A(i,j)$ and $A(j,i) = -A(i,j)$ are the payoffs to players 1 and 2 when they play actions i and j, respectively. The *history* on period t is the sequence of actions played on periods $1, \ldots, t-1$. A *mixed strategy* $s : ([n] \times [n])^* \to \Delta([n])$ is a function from finite histories to probability distributions over $[n]$ (see, e.g., Myerson [7]).

We give a simple (easy to compute) repeated-game strategy for player 1 with the following guarantee. For any finite symmetric zero-sum n by n game $A \in \mathbb{R}^{n\times n}$, any number of rounds $T \geq 1$, and any opponent's strategy s_2, the expected average payoff of player 1 in the first T rounds is at least $-\frac{n}{\sqrt{2T}} \max_{i,j} |A(i,j)|$. We then consider several extensions. We first show that player 1 need not even know the set of strategies in advance. Second, in general-sum games, we show that the COPYCAT strategy guarantees nearly-equal expected payoffs for the two players. Third, we note that the COPYCAT strategy yields *learning equilibria* [3], a type of equilibrium for an entire family of games (a non-Bayesian notion of equilibrium for games with incomplete information).

2 The COPYCAT Strategy

The copycat strategy, defined below, is a relatively simple mixed strategy that can be computed, on period t, in time $\text{poly}(n, \log t)$ using linear programming. The copycat's goal is to equalize the number of occurrences of (i,j) and (j,i) for each $i, j \in [n]$. To this end, the copycat imagines playing a zero-sum *pretend game* P_t each period, whose payoff at (i,j) is equal to the difference between the number of times (j,i) and (i,j) have been played *so far*.

COPYCAT strategy

- On round 1, pick i^1 arbitrarily.
- On round $t = 2, \ldots$:
  - Let $V_t \in \mathbb{Z}^{n\times n}$ be the frequency matrix, where $V_t(i,j)$ is the number of times (i,j) has been played on periods $1, \ldots, t-1$.
  - Output any min-max mixed-strategy of the zero-sum game with payoffs $P_t = V_t^T - V_t$.

Since first-round play is arbitrary and there may be many possible min-max mixed strategies (found efficiently using linear programming), we refer to any such strategy as *a* COPYCAT *strategy.*

Theorem 2.1. *Fix any $n \geq 1$, any finite zero-sum symmetric game with payoffs $A \in \mathbb{R}^{n\times n}$, and any mixed strategy s_2 for player 2. The expected average payoff of player 1 (taken over the realized actions of player 1 playing the copycat strategy against s_2) over T periods, for any $T \geq 1$, is,*

$$\mathbf{E}\left[\ \left|\frac{1}{T}\sum_{t=1}^{T} A(i_t, j_t)\right|\ \right] \leq \frac{n}{\sqrt{2T}} \max_{i,j} |A(i,j)|.$$

Hence, players 1's expected payoff is not much worse (or better) than 0, for large T.

Proof. Without loss of generality, by scaling, we may assume that $\max_{i,j}|A(i,j)| = 1$. Let $\alpha = \sum_{t=1}^{T} A(i_t, j_t)$. Then, $|\alpha| \leq \frac{1}{2}\sum_{i,j}|P_{T+1}(i,j)|$ and, by Cauchy-Schwartz,

$$\begin{aligned}\left(\mathbf{E}[|\alpha|]\right)^2 &\leq \mathbf{E}[\alpha^2] \\ &\leq \mathbf{E}\left[\left(\sum_{i,j}\frac{1}{2}|P_{T+1}(i,j)|\right)^2\right] \\ &\leq \mathbf{E}\left[\frac{n^2}{4}\sum_{i,j}\left(P_{T+1}(i,j)\right)^2\right].\end{aligned}$$

We aim to show $\mathbf{E}[\alpha] \leq n\sqrt{T/2}$. Let $\beta_t = \sum_{i,j}\left(P_{t+1}(i,j)\right)^2$. It suffices to show that $\mathbf{E}[\beta_t] \leq 2t$ for all t. Simple algebra shows that a play of (i_t, j_t) on period t causes a change in β_t of,

$$\begin{aligned}\beta_{t+1} - \beta_t &= \begin{cases}0 & \text{if } i_t = j_t \\ 2 - 4P_t(i_t, j_t) & \text{otherwise}\end{cases} \\ &\leq 2 - 4P_t(i_t, j_t).\end{aligned}$$

The copycat plays the game P_t optimally on period t. Since P_t is symmetric, its value is 0, and hence the copycat is *guaranteed* an *expected* nonnegative pretend payoff on each period t, i.e., $\mathbf{E}[P_t(i_t, j_t)] \geq 0$. Hence $\mathbf{E}[\beta_{t+1} - \beta_t] \leq 2$. Since $\beta_0 = 0$, $\beta_t \leq 2t$ as required.

3 Extensions

In this section, we give a few extensions.

3.1 Unknown Action Sets

For the purposes of this section, it will be helpful to change notation and define a finite two-player symmetric game as $G = (S, u)$ where S is a finite set of actions and $u : S^2 \to \mathbb{R}^2$ such that $u_1(s_1, s_2) = u_2(s_2, s_1)$ for any $s_1, s_2 \in S$. For the purposes of this section, we assume that player 1 is only aware of a *single* action $s_0 \in S$. In particular, player 1 has no knowledge about S or even its size. Each period player 1 observes the play of the previous periods and then chooses an action. Let the known actions at round r be $K_t = \{s_0, s_1^1, s_2^1, s_1^2, s_2^2, \ldots, s_2^{t-1}\}$. A strategy for player 1 is a function which takes as input a finite history $(s_1^1, s_2^1), (s_1^2, s_2^2), \ldots, (s_1^{t-1}, s_2^{t-1}) \in S^2$ and outputs an action $s \in K_t$. That is player 1 is allowed to play either strategy s_0 or one of the strategies that she learned about from observing player 2. The *totally-ignorant* COPYCAT *strategy* is as follows.

Totally-ignorant COPYCAT strategy

- On round 1, let $K_1 = \{s_0\}$ and $j_1 = s_0$.
- On round $t = 2, \ldots$:
  - Define $V_t : K_t \times K_t \to \mathbb{Z}$ by letting $V_t(i,j)$ be the number of prior occurrences of (i,j).
  - Consider the symmetric game H_t with action set K_t and payoff function $u_t(i,j) = V_t(j,i) - V_t(i,j)$.
  - Output any min-max mixed-strategy of G_t (computed with a linear program).

Corollary 3.1. *Fix any $n \geq 1$, any finite zero-sum symmetric game $G = (S, u)$, and any mixed strategy s_2 for player 2. The expected average payoff of player 1 (taken over the realized actions of player 1 playing the totally-ignorant copycat strategy against s_2) over T periods, for any $T \geq 1$, is,*

$$\mathbf{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T} u(i_t, j_t)\right|\right] \leq \frac{|S|}{\sqrt{2T}}\max_{i,j}|u(i,j)|.$$

Proof. WLOG we can assume that $S = [n]$ for $n = |S|$, and $s_0 = 1$. We claim that the totally-ignorant COPYCAT strategy above is in fact a COPYCAT strategy as defined earlier. To see this, note that the game defined in the above loop, H_t, is a *subgame* of the game defined in the earlier version of the COPYCAT strategy, G_t, in the sense that the set of known strategies are a subset of $[n]$ and that the payoffs agree on this subset of strategies. An optimal strategy in H_t must guarantee player 1 non-negative payoff, since H_t is also symmetric and zero-sum and hence has value 0, as well. This strategy, if used in G_t, would also guarantee player 1 a non-negative payoff since the payoff of each strategy in K_t versus $[n] \setminus K_t$ is 0 in G_t. Thus the strategies defined above are optimal for G_t and hence the new COPYCAT strategy is a special case of the old COPYCAT strategy.

3.2 General-sum Games

Let G be a finite two-person general-sum symmetric game, played in the same setting where player 2 knows the game and player 1 does not. The only difference is that we have lifted the restriction that $u(i,j) + u(j,i) = 0$ for all $i, j \in S$. Then a COPYCAT strategy (totally ignorant or not) guarantees player 1 nearly the same payoff as player 2.

Corollary 3.2. *Fix any $n \geq 1$, any finite* general-sum *symmetric game $G = (S, u)$, and any mixed strategy s_2 for player 2. The expected average payoff of player 1 (taken over the realized actions of player 1 playing the totally-ignorant copycat strategy against*

s_2) over T periods, for any $T \geq 1$, is,

$$\begin{aligned} & \mathbf{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T} u(i_t, j_t) - u(j_t, i_t)\right|\right] \\ \leq & \frac{|S|}{\sqrt{2T}} \max_{i,j} |u(i,j) - u(j,i)|. \end{aligned}$$

Proof. Let G' be the zero-sum game with payoffs $u'(i,j) = u(i,j) - u(j,i)$. Now, player 1 uses the COPYCAT strategy in exactly the same manner in G' as G. (In fact, she cannot distinguish between the two.) Hence, the COPYCAT strategy guarantees an average payoff in G' of,

$$\mathbf{E}\left[\left|\frac{1}{T}\sum_{r=1}^{T} u'(i,j)\right|\right] \leq \frac{|S|}{\sqrt{2T}} \max_{i,j} |u'(i,j)|.$$

4 Learning Equilibrium

One nice property of the COPYCAT strategy is that it guarantees to the uninformed player (almost) the value of an underlying zero-sum game, despite the fact the game is initially unknown, and all that it can observe are the actions selected. In the context of symmetric games the COPYCAT strategy yields the uninformed player (almost) the same expected payoff as the one obtained by the informed player. Notice however that in principle in a symmetric game the players can optimize social welfare by alternating between playing the joint action (p, q) and the joint action (q, p), where p and q are selected to maximize $A(i,j) + A(j,i)$ over all actions i, j in the symmetric one-shot game.

Define the *social optimum welfare* to be $\text{opt}_G = \max_{i,j}(A(i,j) + A(j,i))$ in the given symmetric game G, and let (p_G, q_G) be a socially optimal action profile, i.e., a profile achieving $A(p_G, q_G) + A(q_G, p_G) = \text{opt}_G$. Consider the following strategy profile $s = (s_1, s_2)$ for the informed and uniformed players in the infinitely repeated game. The informed player's strategy:

1. In odd iterations play p_G and in even iterations play q_G, unless the uninformed player *failed.*
2. The uninformed player *fails* if there exists an iteration $l \geq 2$, in which she does not copy the informed player's action from the previous iteration $l-1$, i.e., in which $i_l \neq j_{l-1}$.
3. If the uninformed player *fails* then punish her indefinitely by choosing a mixed action that minimizes her maximal expected payoff.

The uninformed player's strategy:

1. In the first iteration choose an arbitrary action. In any iteration $l \geq 2$, unless the informed player *failed*, perform the action chosen by the informed player in iteration $l-1$.
2. The informed player *fails* if she did not perform the same action in all odd iterations and the same action in all even iterations.
3. If the informed player *fails*, the uninformed player adopts the COPYCAT strategy previously developed.

Assume that the number of available actions and the utilities are bounded by some constants, it is now easy to verify the following:

Corollary 4.1. *Given $\epsilon, c > 0$, there exists T, such that for every symmetric game G with payoffs in $[-c, c]$ we have that:*

1. *If the players adopt p then for every $t > T$ the average payoff of each player in the first t iterations is at least $\frac{\text{opt}_G}{2} - \epsilon$.*
2. *For every $t > T$, every possible deviation strategy for either player gives them an expected payoff of at most $\frac{\text{opt}_G}{2} + \epsilon$.*

The corollary follows straightforwardly from Corollary 3.2. Roughly speaking, if both players adopt the suggested strategies then they trivially obtain half of the optimal social welfare each. If the informed player deviates then the uniformed player, by switching to the COPYCAT strategy, will guarantee (almost) equal expected payoffs to both players, which implies no gain (to the deviating informed player) is obtained with respect to half of the optimal social welfare. The fact that the uninformed player can be easily punished when deviating is immediate. In game-theoretic terms what we have just proven is that in the above setting, there exists a *learning equilibrium* for two-person symmetric games: a strategy profile such that unilateral deviation from it is not beneficial for *any* symmetric game in that setting. Moreover, in that equilibrium the optimal social welfare is obtained.

Previous results on the existence of learning equilibrium relied on having higher sensing capabilities, and in particular observing the agents' payoffs. Hence, finding general cases where we do not have that sensing capability is of major importance. Our work exposes such general, highly natural class of games, and show how learning equilibrium can be efficiently constructed for that setting.

References

[1] Auer, P., N. Cesa-Bianchi, Y. Freund, R. Schapire, *The Nonstochastic Multiarmed Bandit Problem/*, SIAM J. Comput 32(1) (2002), 48–77.

[2] Blackwell, D., *An Analog of the MinMax Theorem For Vector Payoffs./*, Pacific J. of Math. 6 (1956), 1–8.
[3] Brafman, R. and M. Tennenholtz, *Efficient Learning Equilibrium./*, Artificial Intelligence. 159(1-2) (2004), 27–47.
[4] Brown, G.W. *Iterative Solutions of Games by Fictitious Play./*, Activity Analysis of Production and Allocation. T.C. Koopmans (Ed.), New York: Wiley.
[5] Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games./*, Cambridge University Press (2006).
[6] Freund, Y. and Schapire, R. *Adaptive game playing using multiplicative weights./*, Games and Economic Behavior. 29 (1999), 79–103.
[7] Myerson, R. *Game theory: analysis of conflict./*, Harvard University Press (1997).
[8] Nash, J. *Equilibrium points in n-person games./*, Proceedings of the National Academy of Sciences. 36(1) (1950), 48–49.
[9] von Neumann, J. *Zur Theorie der Gesellschaftsspiele./*, Mathematische Annalen, 100, (1928), 295–300.