

There Can Be No Single Best Adaptive Poker AI

Néill Sweeney

Independent Researcher, Dublin, Ireland.
neill.sweeney4@mail.dcu.ie

Abstract

Adaptive strategies are popular in poker AI research. Three desirable properties of an adaptive poker bot are optimality, generality and speed of response to new opponents. These three properties though cannot be achieved simultaneously for most imperfect information games; some trade-off must be made between them. This general principle is connected to recent work on poker AI and particularly the total bankroll competitions at the Annual Computer Poker Competition (ACPC). This paper is meant to generate discussion between researchers about how to explain and quantify these trade-offs better and to possible future directions for research.

Introduction

First some terminology will be introduced. The aim is not to redefine standard terms but to emphasise certain aspects which will be important later.

The player: the player whose utility function we are trying to optimise.

The opponents: the other players in the repeated game.

Repeated game: A game consisting of repeatedly playing an identical game (called the stage game) and in which the reward for each player for the repeated game is a linear function of the rewards at each stage game and players can remember the action history between stages. On a side note: round-robin (everyone plays everyone-else) leagues where a single reward goes to the winner of the overall league are not strictly repeated games because the overall result is not a linear function of the results of the individual stage games. I suspect the Nash equilibrium for the league would include adaptive strategies but I know of no studies that describe the equilibria of such leagues even for very simple games. For the purpose of this paper, I ignore this complication and assume the aim of the player is to optimise its expected reward rather than having a “win or bust” attitude.

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Match length: the number of stage games to be played in the repeated game.

Static strategy: A strategy for a repeated game where the strategy played at each stage game is independent and identical.

Adaptive strategy: A strategy for a repeated game where the strategy for each stage game depends on the play of the previous stages.

Maximin solutions to 2-player zero-sum games are a central concept in game theory (which corresponds with Nash equilibrium solution in this case).

Outline of Argument

Firstly, the uninteresting case; if a strategy exists which weakly dominates all other strategies in the stage game, there is no incentive to for a rational player to play anything else. It can be demonstrated this is not the case for simplified versions of poker. For example, the [0,1] poker game used in the foundation of game theory (Von Neumann and Morgenstern 1944). In all other cases, it is not possible to play a *best* response to *any* new set of opponents *immediately*. This is meant as a statement of the obvious: the reason to phrase it this way is to emphasise that to sensibly frame the problem, a compromise must be accepted in one of three areas; optimality, generality or speed (corresponding to the terms in *italics*). These three headings will now be discussed assuming the uninteresting case does not apply.

Compromise on Optimality

The most common option is to seek a maximin strategy instead a best response. This maximises the minimum expected return a player can achieve no matter what his opponents do. This will be a best response only if the opponents actually play the other side of the maximin solution. The distance of a strategy from a true maximin can be measured by its exploitability; its performance

against the opponent (or team of opponents) which minimises the player's reward often called "the Nemesis". This presents no theoretical difficulties but does present the practical difficult of finding "the Nemesis". Recent work though has demonstrated that exploitabilities can be calculated for two-player limit hold'em (Johanson et al 2011).

Compromise on Generality

Instead of seeking a best response to any opponent over any time scale, it is possible to seek a best response to a set of opponent strategies drawn from a specific population with the number of stage games also drawn from a specific distribution. The reason this makes the framing of the problem sensible is that the expected value of a strategy can be calculated over the population. A strategy that is optimal for one population (of opponents and match lengths) will not necessarily be optimal for others because the expected value of the strategy depends on the population used in the calculation. Of course, techniques to find a best response to one population will tend to do well against similar populations.

Adaptive strategies could be useful in approximating a best response to a population for two reasons. Firstly, the early stages of the match can be used to try to narrow down which member of the opponent population is being faced which can then be used to alter the player's strategy in subsequent stage games. Secondly, if the opponents population includes adaptive strategies, it may be possible for the player to make decisions in the early stages which induce the opponents to play favourably towards him at the later stages.

Only in the smallest of toy problems will we have a fully specified population i.e. a list of opponent strategies with a probability for each. A sample might be available though. This means statistical analysis is possible; statistics being the inference of the properties of a population from a sample. The success or failure of any attempt to generate an adaptive strategy for any population may depend on how easy it is predict the behaviour of new opponents from the available sample (i.e. "do statisites").

Note that game theory is not without merit when seeking a population best-response. Any theory that can make predictions about the actions of opponents can form the basis of a statistical model. Whether a game theory model forms a successful basis for an optimal strategy for a specific population depends on how important the differences between the population and the assumptions of game theory prove to be.

Adaptive strategies are also attractive in multi-player games. They allow the player to use the early stages of a repeated game to effectively negotiate the formation of a

coalition with some of their opponents. This can occur even if explicit communication between players is not allowed in the game. This is aptly illustrated by the lemonade stand game (Zinkevitch, Bowling and Wunder . 2011).

Compromise on Speed of Adaptation

Online learning algorithms learn to solve stochastic decision problems by *repeatedly* trying possible strategies and adjusting their strategy in reaction to the results achieved so as to achieve optimal performance. The field of reinforcement learning is almost exclusively devoted to studying such online learning algorithms. Many techniques have solid asymptotic performance guarantees in limit of an infinite number of attempts at a static problem. (In practice, most techniques also perform well if the target problem is slowly changing.) Finding a best response to a single opponent (or even a population) is a stochastic decision problem. So seeking an online algorithm that will find a best response to *any* opponent *eventually* is perfectly sensible. But the theoretical results in reinforcement learning are usually asymptotic results. They say nothing about performance in the short term. No online algorithm can be expected to learn an optimal strategy before it has had a chance to explore a substantial part of the strategy space.

A particular type of opponents strategy which causes difficulty when considering optimal adaptive strategies should also be mentioned here. In theoretical work on learning in repeated games with finite automata, they are usually excluded. They are first mentioned by Moore (1956), who called them combination-lock automata. The possibility of these types of opponents nor the difficulties they cause generally optimal adaptive strategies is not limited to automata. Broadly, these are strategies where the opponent plays mostly like the opponent in a maximin solution but if the player makes a particular series of actions (the combination) in the early stages of the game, the opponent strategy changes from the most difficult possible to the most generous possible for the remainder of the game. The more complex (i.e. more decisions per stage game and more stages in a match) the game the easier it is for these combination-lock strategies to encode a combination that is unlikely to be guessed in decisions in the early stages of a game while still leaving enough time for the switch to the generous strategy to have a substantial effect on the rewards of the player. Hence combination-lock opponents are less of an issue in small "toy" games.

It is hard to envisage a "real-world" situation in which opponents would deliberately have combination-lock strategies but if the population of opponents is not limited in any way, they are always possible for reasonably

complex games. Hence, it is not possible to say a particular strategy is approximately optimal for all opponents, as the best response to a combination-lock opponent will be the one that uses the combination. A specific example of such a strategy will be given at the end of the next section.

If, though, the online player is allowed to play multiple matches against an opponent, it may find the best-response to even combination-lock opponents. Note that this is an asymmetric contest where the opponents memory is cleared at the start of each match but the online learning algorithm retains information from match to match. A good online algorithm always maintains an element of exploration (Singh et al. 2000) which means it will eventually find the combination for such opponents. But in a complex game, eventually can be a very long time as there is a large number of possible combinations to try.

Seeking a population best response by a statistical approach can bypass these problems in two ways. Firstly its standard of optimality is lower. It will only be optimal in the mean, not to every strategy in the population. Also the (preferably large) sample may provide a lot of the information that would otherwise need to be discovered from exploration. So adaptation could be much more rapid.

Connection to Poker AI

The accepted aim of the total bankroll competition is to encourage the development of adaptive Poker AI systems. The consensus of informal discussions at last years workshop is that it has not been a complete success. Either the results were dominated by minimax approximations (or equilibrium approximations for the 3-player contest) or the top positions were completely dependent on the presence of a single very weak entry.

The previous discussion about the three compromises an adaptive system designer has to choose from suggest possible reasons why this might be so. Consider someone trying to develop a system specifically tuned to the ACPC contest. Predicting entrants for next years competition is a particularly difficult task. At first glance, it might appear that there is a large data-set available in the form of the logs of previous competitions. But the the competition attracts a small number of entries (by statistical standards), so the effective sample size is small. Many of those strategies will be reasonably good approximations to a minimax strategy; due to the work of (Johanson et al. 2011) we have solid results to confirm this. It would be more important to predict how many highly exploitable strategies will be entered and what the weaknesses of those strategies will be. But in practice these are the most difficult entries to predict; they are usually a teams' first effort at constructing a poker bot. After a poor showing in their first entry, teams usually make radical changes to

their design or don't enter again. Note that it is much easier for a programmer to make a rapid and significant change to the strategy of a bot (for example, by fixing a bug) than for a human player to make a similar change in their own strategy. So statistical approaches are far likelier to succeed against a human population. This is not to say they can't succeed as evidenced by recent work (Bard et al. 2013).

To quantify the effect of the statistically small number of entries, it is instructive to generative some sampling variance estimates. A simple method for generating these is bootstrap resampling (Good 2001). In particular, consider the difference in mean score between first and second in a total-bankroll competition. The variance of this value due to variation in the entries can be estimated by resampling with replacement the other entries in that years competition. This assumes that the other entries are independently and identically drawn from some population which isn't strictly correct but they still provide a quantitative estimate of how important changes in the line-up are to the result. Results for the last three years heads-up competitions are in Table 1.

	Year	First	Second	No. of entries	Gap	S.E.
2p limit	2012	slumbot	littlerock	11	15	2.5
	2011	Calamari	Sartre	19	5	7.1
	2010	PULPO	Hyperborean.tbr	13	18	11.5
2p NL	2012	Little.rock	Hyperborean.tbr	8	521	321
	2011	Lucky7	SartreNL	7	265	1472
	2010	Tartanian4.tbr	PokerBotSLO	5	705	290

Table 1: Bootstrap estimates for the standard error for the gap (in millibets per hand) between first and second in the total-bankroll two player competitions in recent years

Someone trying to use an online learning approach for any of the usual competitions in ACPC is faced with a different problem. As previously stated, there are online algorithms whose asymptotic performance can be proven. The difficulty for any entrant is getting to point where this asymptotic performance kicks-in within the match length for the competition (3,000 deals for the heads-up competitions and 1,000 for the 3-player). Firstly note that neither player in a match is allowed to retain information from previous matches. This is a sensible rule as it means that each match is independent; without this rule, statistical analysis of the results would be very difficult. But it means any player is effectively dealing with their opponent for the current match for the first time; remember the opponent

may be using adaptive strategies as well. This is not the situation in which online algorithms are guaranteed to work.

Note also the combination-lock opponents described in the “Compromise on Speed of Adaptation” section could be entered into the competition. It might help to give a concrete example. Consider a strategy that almost always plays an approximation to a minimax strategy. The exception is if its opponent folds immediately for the first four hands and then raises at every opportunity for the next two (or some other unlikely combination). It will then raise at every opportunity before the river and then fold to any bet on the river. A general adaptive strategy can’t hope to get near the results of the best response of this strategy. If someone tried to slip such an obvious example of a combination-lock strategy into the competition under an assumed name and entered the best response to it under their own name, it would probably be spotted and barred for chip-dumping. This is an extreme example to illustrate the point. A less obvious and completely innocent example (maybe there is a rarely triggered bug in an entry) could still dominate the results of a total-bankroll competition.

Conclusion

My contention is that it is impossible for an adaptive strategy to be completely general (rather than specific to a population), optimal (rather than minimax robust) and rapidly adjust (rather than after long exploration). Hence no competition can be designed to identify a single best adaptive strategy for a complex game. This is not an original argument: it is effectively a restatement of the famous “no-free-lunch” theorems (Wolpert and Macready 1997) but I feel this general point is easily lost when the discussion shifts to the technical details of running a competition. Similar points have been made before in the literature; two examples are (Halck and Dahl 1999) and (Shoham, Powers and Grenager 2007). Note that this discussion applies not only to the total bankroll competitions in ACPC or to computer poker but to any scenario where a small number of distinct algorithms are competing in a complex domain.

Computer poker is uniquely positioned to provide quantitative estimates for the size of the trade-offs for a complex imperfect information game people already have strong intuitions about. This already has happened with work on the restricted Nash technique (Johanson, Zinkevich and Bowling 2007) which includes estimates of the trade-off between robustness and optimality.

It is tempting to try force all poker AI research onto a single axis and hence rank the bots produced by that research in a competition. But it is neither possible nor helpful. I see three broad directions of research currently

active in poker AI, each with its own merits. The first broad direction is equilibrium approximation; the aim is to minimise the exploitability. The second is “black-box” best response calculation; the aim is to calculate the best-response to all opponents within a specified class. For a game as complex as poker this takes a long-time (i.e. longer than the match lengths in the competition) but not computationally infeasible lengths (i.e. less than a month). The third broad direction is population best response; the aim is performance against the specified population. A strategy that is optimal for an interesting population is itself of interest even if it does poorly outside that population. This has been somewhat neglected because it is difficult for researchers to agree on what constitutes an interesting population. While everyone is free to select their own population of interesting opponents, some of the value of the research is lost if comparisons cannot be made between the work of different research teams.

The most interesting population to humans is other humans. AI performance versus humans has already been explored in the Man-Machine contests (Johanson 2007). Poker is perhaps uniquely positioned for studies into human decision-making and its interaction with AI’s because of its popularity. This means there is a better chance of finding significant numbers of motivated subjects to play against poker AI’s than for most games studied. There are practical difficulties and it will require salesmanship to recruit motivated subjects, but it is still an avenue that is worth exploring.

Finally, this paper does not imply that the APCP or adaptive entries to the competition are without research value. The cross-table of results provide invaluable data-points from which conjectures about how complex AI’s interact in imperfect information games can be developed. My point is that the poker AI community should not fret about designing a competition to identify the single “best” bot as the task is not well defined.

References

- Bard, N., Johanson, M., Burch, N., & Bowling, M. 2013. Online Implicit Agent Modelling. Proceedings of the Twelfth International Conference on Autonomous Agents and Multiagent Systems (to appear).
- Good, P. I. 2001. *Resampling methods: A practical guide to data analysis*. Springer.
- Johanson, M. 2007 M.Sc. Thesis, Robust Strategies and Counter-Strategies: Building a Champion Level Computer Poker Player, University of Alberta .
- Johanson, M., Zinkevich, M., Bowling, M. Computing robust counter-strategies. *Advances in neural information processing systems* 20 2007: 721-728.

Johanson, M., Waugh, K., Bowling, M., & Zinkevich, M. 2011. Accelerating best response calculation in large extensive games. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume One* (pp. 258-265). AAAI Press.

Halck, O.M. and Dahl, F.A. 1999 On classification of games and evaluation of players—with some sweeping generalizations about the literature. In: Proceedings of the ICML-99 Workshop on Machine Learning in Game Playing.

Moore, E. F. 1956. Gedanken-experiments on sequential machines. *Automata studies*, 34, 129-153.

Singh, S., Jaakkola, T., Littman, M.L. and Szepesvári, C. 2000 Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3), pp.287-308.

Sklansky, D. (2005). *Theory of Poker*. Two Plus Two Publishing.

Shoham, Y., Powers, R., & Grenager, T. 2007. If multi-agent learning is the answer, what is the question?. *Artificial Intelligence*, 171(7), 365-377.

Von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.

Wolpert, D. H., & Macready, W. G. 1997. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1), 67-82.

Zinkevich, M. A., Bowling, M., & Wunder, M. 2011. The lemonade stand game competition: solving unsolvable games. *ACM SIGecom Exchanges*, 10(1), 35-38.