Data Biased Robust Counter Strategies

Michael Johanson
johanson@cs.ualberta.ca
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada

Michael Bowling
bowling@cs.ualberta.ca
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada

Abstract

The problem of exploiting information about the environment while still being robust to inaccurate or incomplete information arises in many domains. Competitive imperfect information games where the goal is to maximally exploit an unknown opponent's weaknesses are an example of this problem. Agents for these games must balance two objectives. First, they should aim to exploit data from past interactions with the opponent, seeking a best-response counter strategy. Second, they should aim to minimize losses since the limited data may be misleading or the opponent's strategy may have changed, suggesting an opponent-agnostic Nash equilibrium strategy. In this paper, we show how to partially satisfy both of these objectives at the same time, producing strategies with favourable tradeoffs between the ability to exploit an opponent and the capacity to be exploited. Like a recently published technique, our approach involves solving a modified game; however the result is more generally applicable and even performs well in situations with very limited data. We evaluate our technique in the game of two-player, Limit Texas Hold'em.

1 INTRODUCTION

Maximizing utility in the presence of other agents is a fundamental problem in game theory. In a zero-sum game, utility comes from the exploitation of opponent weaknesses, but it is important not to allow one's own strategy to be exploited in turn. Two approaches to such problems are well known: best response strategies and Nash equilibrium strategies. A best response strategy maximizes utility for an agent, assuming perfect knowledge of its static opponent. However, such strategies are *brittle*: against a worst case opponent, they have a high exploitability. In a two-player zero-sum game, a Nash equilibrium strategy maximizes its utility against a worst-case opponent. As a result, we say that such strategies are *robust*. If a perfect model of the opponent is available, then they can be exploited by a best response; if a model is not available, then playing a Nash equilibrium strategy is a sensible choice. However, if a model exists but it is somewhat unreliable (*e.g.*, if it is formed from a limited number of observations of the opponent's actions, or if the opponent is known to be changing strategies) then a better option may be to compromise: accepting a slightly lower worst-case utility in return for a higher utility if the model is approximately correct.

One simple approach for creating such a compromise strategy is to create both a best response strategy and a Nash equilibrium strategy, and then play a mixture of the two. Before each game, we will flip a biased coin. With probability p we will use the best response, and with probability $(1 - p)$ we will use the Nash equilibrium. By varying p, we can create a range of strategies that linearly trade off *exploitation* of the opponent and our own *exploitability* by a worst-case opponent. While this approach is a useful baseline, we would like to make more favourable tradeoffs between these goals.

McCracken and Bowling (2004) proposed ϵ-safe strategies as another approach. The set of ϵ-safe strategies contains all strategies that are exploitable by no more than ϵ. From this set, the strategies that maximize utility against the opponent are the set of ϵ-safe best responses. Thus, for a chosen ϵ, the set of ϵ-safe best responses achieve the best possible tradeoffs between exploitation and exploitability. However, their approach is computationally infeasible for large domains, and has only been applied to Ro-Sham-Bo (Rock-Paper-Scissors).

In previous work we proposed the restricted Nash response (Johanson et al., 2008) technique (RNR) as a practical approach for generating a range of strategies that provide good tradeoffs between exploitation and exploitabil-

Appearing in Proceedings of the 12^{th} International Confe-rence on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

ity. In this approach, a modified game is formed in which the opponent is forced to act according to an opponent model with some probability p, and is free to play the game as normal with probability $(1 - p)$. When p is 0 the result is a Nash equilibrium, and when p is 1 the result is a best response. When $0 < p < 1$ the technique produces a counter-strategy that provides different trade-offs between exploitation and exploitability. In fact, the counter-strategies generated are in the set of ϵ-safe best responses for the counter-strategy's value of ϵ, making them the best possible counter-strategies, assuming the model is correct. In a practical setting, however, the model is likely formed through a limited number of observations of the opponent's actions, and it may be incomplete (it cannot predict the opponent's strategy in some states) or inaccurate. As we will show in this paper, the restricted Nash response technique can perform poorly under such circumstances.

In this paper, we present a new technique for generating a range of counter-strategies that form a compromise between the exploitation of a model and its exploitability. These counter-strategies, called **data biased responses** (DBR), are more resilient to incomplete or inaccurate models than the restricted Nash response (RNR) counter-strategies. DBR is similar to RNR in that the technique involves computing a Nash equilibrium strategy in a modified game where the opponent is forced with some probability to play according to a model. Unlike RNR, the opponent's strategy is constrained on a per-information set basis, and depends on our confidence in the accuracy of the model. For comparison to the RNR technique, we demonstrate the effectiveness of the technique in the challenging domain of 2-player Limit Texas Hold'em Poker.

2 BACKGROUND

A **perfect information extensive game** consists of a tree of game states and terminal nodes. At each game state, an action is taken by one player (or by "chance") causing a transition to a child state; this is repeated until a terminal state is reached. The terminal state defines the payoffs to the players. In **imperfect information extensive games** such as poker, the players cannot observe some piece of information (such as their opponent's cards) and so they cannot exactly determine which game state they are in. Each set of indistinguishable game states is called an **information set** and we denote such a set by $I \in \mathcal{I}$. A **strategy** for player i, σ_i, is a mapping from information sets to a probability distribution over actions, so $\sigma_i(I, a)$ is the probability player i takes action a in information set I. The space of all possible strategies for player i will be denoted Σ_i. In this paper, we will focus on two player games.

Given strategies for both players, we define $u_i(\sigma_1, \sigma_2)$ to be the expected utility for player i if player 1 uses the strategy $\sigma_1 \in \Sigma_1$ and player 2 uses the strategy $\sigma_2 \in \Sigma_2$. A best response to an opponent's strategy σ_2 is a strategy for player 1 that achieves the maximum expected utility of all strategies when used against the opponent's strategy. There can be many strategies that achieve the same expected utility; we refer to the set of best responses as $BR(\sigma_2) \subseteq \Sigma_1$. For example, the set of best responses for player 1 to use against σ_2 is defined as:

$$BR(\sigma_2) = \{\ \sigma_1 \in \Sigma_1 : \forall \sigma_1' \in \Sigma_1 u_1(\sigma_1, \sigma_2) \geq u_1(\sigma_1', \sigma_2)\}$$

A **strategy profile** σ consists of a strategy for each player in the game; *i.e.*, (σ_1, σ_2). In the special case where $\sigma_1 \in BR(\sigma_2)$ and $\sigma_2 \in BR(\sigma_1)$, we refer to σ as a Nash equilibrium. A **zero-sum extensive game** is an extensive game where $u_1 = -u_2$ (one player's gains are equal to the other player's losses). In such games, all Nash equilibrium strategies have the same utility for the players, and we refer to this as the **value of the game**. We define the term **exploitability** to refer to the difference between a strategy's utility when playing against its best-response and the value of the game for that player. We define **exploitation** to refer to the difference in utility between one strategy's utility against a specific opponent strategy and the value of the game for that player.

A strategy that can be exploited for no more than ϵ is called **ϵ-safe**, and is a member of the set of ϵ-safe strategies $\Sigma_1^{\epsilon\text{-safe}} \subseteq \Sigma_1$. A strategy profile where each strategy can be exploited by no more than ϵ is called an ϵ-Nash equilibrium. Given the set $\Sigma_1^{\epsilon\text{-safe}}$, there is a subset $BR^{\epsilon\text{-safe}}(\sigma_2) \subseteq \Sigma_1^{\epsilon\text{-safe}}$ that contains the strategies that maximize utility against σ_2:

$$BR^{\epsilon\text{-safe}}(\sigma_2) = \{\ \sigma_1 \in \Sigma^{\epsilon\text{-safe}} : \forall \sigma_1' \in \Sigma^{\epsilon\text{-safe}} u_1(\sigma_1, \sigma_2) \geq u_1(\sigma_1', \sigma_2)\}$$

3 TEXAS HOLD'EM POKER

Heads-Up Limit Texas Hold'em poker is a two-player wagering card game. In addition to being commonly played in casinos (both online and in real life), it is also the main event of the AAAI Computer Poker Competition (Zinkevich and Littman, 2006), an initiative to foster research into AI for imperfect information games. Texas Hold'em is a very large zero-sum extensive form game with imperfect information (the opponent's cards are hidden) and stochastic elements (cards are dealt at random). Each individual game is short, and players typically play a session of many games.

We will briefly summarize the rules of the game. A session starts with each player having some number of **chips**, which usually represent money. A single game of Heads-Up Limit Texas Hold'em consists of each player being forced to place a small number of chips (called a **blind**) into the **pot** before being dealt two private cards. The players will combine these private cards with five public cards

that are revealed as the game progresses. The game has four phases: the preflop (when two private cards are dealt), the flop (when three public cards are dealt), the turn (when one public card is dealt) and the river (when one final public card is dealt). If both players reach the end of the game (called a **showdown**), then both players reveal their private cards and the player with the best 5-card poker hand wins all of the chips in the pot. If only one player remains in the game, then that player wins the pot without revealing their cards. After the cards are dealt in each phase, the players engage in a **round of betting**, where they **bet** by placing additional chips in the pot that their opponent must match or exceed in order to remain in the game. To do this, the players alternate turns and take one of three actions. They may **fold** to exit the game and let the opponent win, **call** to match the opponent's chips in the pot, or **raise** to match, and then add a fixed number of additional chips (the "bet" amount). When both players have called, the round of betting is over, and no more than four bets are allowed in a single round.

The goal is to win as much money as possible from the opponent by the end of the session. This distinguishes poker from games such as Chess or Checkers where the goal is simply to win and the magnitude of the win is not measured. The performance of an agent is measured by the number of bet amounts (or just bets) they win per game across a session. Between strong computer agents, this number can be small, so we present the performance in millibets per game (mb/g), where a millibet is one thousandth of a bet. A player that always folds will lose 750 millibets per game to their opponent, and a strong player can hope to win 50 millibets per game from their opponent. Due to a standard deviation of approximately 6000 millibets per game, it can take more than one million games to distinguish with 95% confidence a difference of 10 millibets per game.

Since the goal of the game is to maximize the exploitation of one's opponent, the game emphasizes the role of exploitive strategies as opposed to equilibrium strategies. In the two most recent years of the AAAI Computer Poker Competition, the "Bankroll" event which rewards exploitive play has been won by agents that lost to some opponents, but won enough money from the weakest agents to have the highest total winnings. However, many of the top agents have been designed to take advantage of a suspected *a priori* weakness common to many opponents. A more promising approach is to observe an opponent playing for some fixed number of games, and use these observations to create a counter-strategy that exploits the opponent for more money than a baseline Nash equilibrium strategy or a strategy that exploits some expected weaknesses.

3.1 ABSTRACTION

The variant of poker described above has 9.17×10^{17} game states; computing best responses and Nash equilibria in a game of this size is intractable. Therefore, it is common practise to instead reduce the real game to a much smaller abstract game that maintains as many of the strategic properties as possible. The strategies of interest to us will be computed in this abstract game. To use the abstract game strategy to play the real game, we will map the current real game information set to an abstract game information set, and choose the action specified by the abstract game strategy.

The game is abstracted by merging information sets that result from similar chance outcomes. On the preflop, one such abstraction might reduce the number of chance outcomes from 52 choose 2 down to 5, and from (52 choose 2)(50 choose 3) to 25 on the flop. Each chance outcome is reduced to one of 5 outcomes, giving 625 possible combinations, resulting in a game that has 6.45×10^9 game states. In this abstract game, best response counter-strategies can be computed in time linear in the size of the game tree; on modern hardware, this takes roughly 10 minutes. Using recent advances for solving extensive form games (Zinkevich et al., 2008), a Nash equilibrium for this abstract game can be approximated to within 3 millibets per game in under 10 hours.

3.2 OPPONENT STRATEGIES

Much of the recent effort towards creating strong agents for Texas Hold'em has focused on finding Nash equilibrium strategies for abstract games (Zinkevich et al., 2008; Gilpin and Sandholm, 2006). We want to examine the ability to exploit opponent weaknesses, so we will examine results where the opponent is not playing an equilibrium strategy. Toward this end, we created an agent similar to "Orange", which was designed to be overly aggressive but still near equilibrium and competed in the First Man-Machine Poker Championship (Johanson, 2007, p. 82),. "Orange" is a strategy for an abstract non-zero-sum poker game where the winner gets 7% more than usual, while the loser pays the normal price. When this strategy is used to play the normal (still abstract) zero-sum game of poker, it is exploitable for 28 millibets per game. This value is the upper bound on the performance obtainable by any counter-strategy that plays in the same abstraction.

In this paper, we will also refer to an agent called "Probe" (Johanson et al., 2008). Probe is a trivial agent that never folds, and calls and raises whenever legal with equal probability. The Probe agent is useful for collecting observations about an opponent's strategy, since it forces them into all of the parts of the game tree that the opponent will consent to reach.

3.3 OPPONENT BELIEFS

A belief about the opponent's current strategy can simply be encoded as a strategy itself. Even a posterior belief derived from a complicated prior and many observations still can be summarized as a single function mapping an information set to a distribution over actions, the **expected posterior strategy**[1]. In this work, we will mainly take a frequentist approach to observations of the opponent's actions (although we discuss a Bayesian interpretation to our approach in Section 7). Each observation is one full information game of poker: both players' cards are revealed. The model of our opponent will consider all of the information sets in which we have observed the opponent acting. The probability of the opponent model taking an action a in such an information set I is then set to the ratio of the number of observations of the opponent playing a in I to the number of observations of I. There will likely be information sets in which we have never observed the opponent acting. For such information sets, we establish a **default policy** to always choose the call action (Johanson, 2007, p. 60)[2]

Since our opponent model is itself a strategy, it can be used to play against the counter-strategies that are designed to exploit it. We would expect the counter-strategies to perform very well in such cases, and this is demonstrated in our previous work on restricted Nash responses (Johanson et al., 2008). However, since the model is constructed only from (possibly a small number) observations of the opponent's strategy, it is more interesting to examine how the counter-strategies perform against the actual opponent's strategy.

4 LIMITATIONS OF CURRENT METHODS

As discussed in the introduction, restricted Nash response counter-strategies form an envelope of possible counter-strategies to use against the opponent, assuming the opponent model is correct (Johanson et al., 2008). The restricted Nash response technique was designed to solve the brittleness of best response strategies. As was presented in Table 1 of that work, best response strategies perform well against their intended opponent, but they can perform very badly against other opponents, and are highly exploitable by a worst-case opponent. Restricted Nash response strategies are robust, and any new technique for producing counter-strategies should also be able to produce robust strategies. However, restricted Nash response strategies have three limitations. We will show that our new counter-strategy technique addresses these issues.

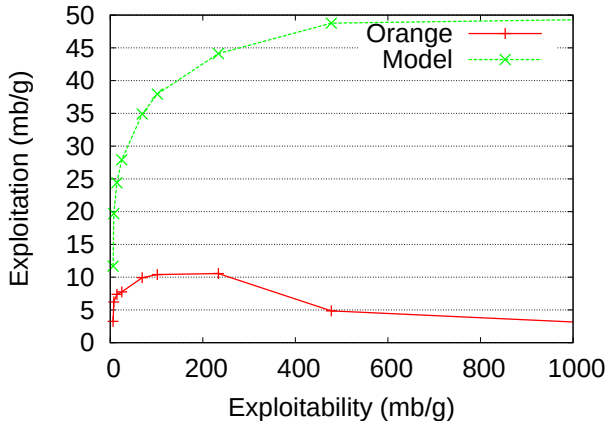
Before discussing the limitations, we first explain the exploitability-versus-exploitation graph that is used throughout the paper. For each counter-strategy, we can measure the exploitability (worst-case performance) and exploitation (performance against a specific opponent). So we can plot any counter-strategy as a point on a graph with these axes. Restricted Nash responses involve a family of counter-strategies attained by varying p. Hence, we plot a curve passing through a set of representative p-values to demonstrate the shape of the envelope of strategies. Since the exploitability is determined by the choice of p, we are (indirectly) controlling the exploitability of the resulting counter-strategy, and so it appears on the x-axis; the counter-strategy's exploitation of the specific opponent is the result, and is shown on the y-axis. In each of the following graphs, the values of p used were 0, 0.5, 0.7, 0.8, 0.9, 0.93, 0.97, 0.99, and 1. Each value of p corresponds to one datapoint on each curve. Unless otherwise stated, each set of counter-strategies was produced with 1 million observed games of Orange playing against Probe.

Restricted Nash response counter-strategies can overfit to the model. By varying p, the resulting restricted Nash response counter-strategies each present a different tradeoff of exploitation and exploitability when compared against their opponent model. As p increases, the counter-strategies exploit the opponent model to a higher degree, and are themselves more exploitable. However, as Figure 1a shows, this trend does not hold when we compare their performance against the actual opponent instead of the opponent model. As p increases, the counter-strategies begin to do markedly worse against the actual Orange strategy. The computed counter-strategy has overfit to the opponent model. As the number of observations approach the limit, the opponent model will perfectly match the actual opponent in the reachable part of the game tree, and this effect will lessen. In a practical setting, however, p must be chosen with care so that the resulting counter-strategies provide favourable trade-offs.

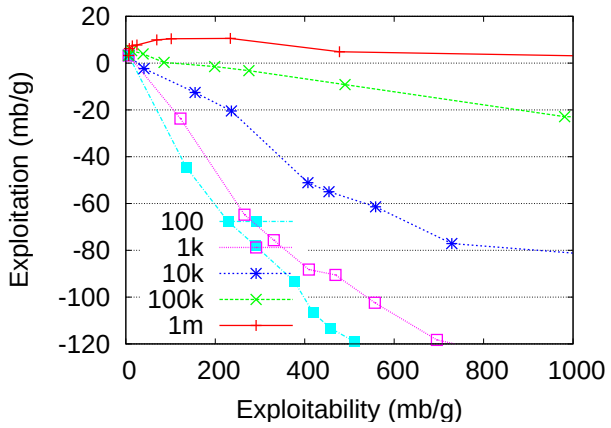
Restricted Nash response counter-strategies require a large quantity of observations. It is intuitive that, as any technique is given more observations of an opponent, the counter-strategies produced will grow in strength. This is true of the restricted Nash response technique. However, if there is not a sufficient quantity of observations, increasing p can make the resulting counter-strategies *worse* than the equilibrium strategy. This is another aspect of the restricted Nash response technique's capacity to overfit the model; if there is an insufficient number of observations, then the default policy plays a larger part of the model's strategy and the resulting counter-strategy is less applicable to the actual opponent. Figure 1b shows this effect. With less than 100 thousand observed games, increasing p causes the counter-

[1] If $f : \Sigma_2 \rightarrow \Re$ is the posterior density function over strategies, then the expected posterior strategy chooses action a at information set I with probability $\bar{\sigma}_1(I, a) = \int_{\sigma_1 \in \Sigma_1} \sigma_1(I, a) f(\sigma_1)$

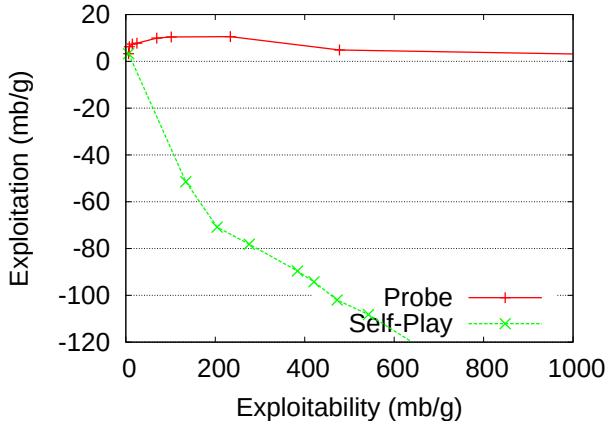
[2] Alternative default policies were tried in this previous work, but all performed far inferior.



(a) Overfitting



(b) Observations



(c) Training

Figure 1: Exploitation versus exploitability curves that illustrate three problems in the restricted Nash response technique. In 1a, we note the difference in performance when counter-strategies play against the opponent model and against the actual opponent. In 1b, we see how a scarcity of observations results in poor counter-strategies. In 1c, we see that the technique performs poorly when self-play data is used. Note that the red, solid curve is the same in each graph.

strategies to be both more exploitable and less exploitive.

Restricted Nash response counter-strategies are sensitive to the choice of training opponent. Ideally, a technique for creating counter-strategies based on observations should be able to accept any reasonably diverse set of observations as input. However, the restricted Nash response technique requires a very particular set of observations in order to perform well. Figure 1c shows the performance of two sets of restricted Nash response counter-strategies. The set labelled Probe uses an opponent model that observed one million games of Orange playing against Probe; the set labelled Self-Play uses an opponent model that observed one million games of Orange playing against itself. One might think that a model constructed from self-play observations would be ideal, because it would be accurate in the parts of the game tree that the opponent is likely to reach. Instead, we find that self-play data is of no use when constructing a restricted Nash response counter-strategy. If an agent will not play to reach some part of the game tree, then the opponent model has no observations of the opponent in that part of the tree, and is forced to turn to the default policy which may be very dissimilar from the actual opponent's strategy. The Probe agent forces the the opponent to play into all of the parts of the tree reachable because of the opponent's strategy, however, and thus the default policy is used less often.

5 DATA BIASED RESPONSE

The guiding idea behind the restricted Nash response technique is that the opponent model may not be perfect. The parameter p can be thought of as a measure of confidence in the model's accuracy. Since the opponent model is based on observations of the opponent's actions, there can be two types of flaws in the opponent model. First, there may be information sets in which we never observed the opponent, and so the opponent model must provide a default policy to be taken at this information set. Second, in information sets for which there were a small number of observations, the observed frequency of actions may not match the true opponent's action probabilities.

We claim that the restricted Nash response technique's selection of one parameter, p, is not an accurate representation of the problem, because the accuracy of the opponent model is not uniform across all of the reachable information sets. Consider the two cases described above. First, in unobserved information sets, the opponent model uses the default policy and is unlikely to accurately reflect the opponent's strategy. If we could select a value of p for just this information set, then p would be 0. Second, the number of observations of a particular information set will vary wildly across the game tree. In information sets close to the root, we are likely to have many observations, and so we expect the model to be accurate. In information sets

that are far from the root, we will tend to have fewer observations, and so we expect the model to be less accurate. If we were selecting a value of p for one information set, it should depend on how accurate we expect the model to be; one measure of this is the number of times we have observed the opponent acting in this information set.

This is the essential difference between the restricted Nash response technique and the data biased response technique. Instead of choosing one probability p that reflects the accuracy of the entire opponent model, we will assign one probability to each information set I and call this mapping P_{conf}. We will then create a modified game in the following way. Whenever the restricted player reaches I, they will be forced to play according to the model with probability $P_{\text{conf}}(I)$, and can choose their actions freely with probability $(1 - P_{\text{conf}}(I))$. The other player has no restrictions on their actions. When we solve this modified game, the unrestricted player's strategy becomes a robust counter-strategy to the model.

One setting for P_{conf} is noteworthy. If $P_{\text{conf}}(I)$ is set to 0 for some information sets, then the opponent model is not used at all and the player is free to use any strategy. However, since we are solving the game, this means that we assume a worst-case opponent and essentially compute a Nash equilibrium in these subgames.

5.1 SOLVING THE GAME

Given an opponent model σ_{fix} and P_{conf}, the restricted player chooses a strategy σ_2' that makes up part of their restricted strategy σ_2. The resulting probability of σ_2 taking action a at information set I is given as:

$$\sigma_2(I,a) = P_{\text{conf}}(I) \times \sigma_{\text{fix}}(I,a) + (1 - P_{\text{conf}}(I)) \times \sigma_2'(I,a) \quad (1)$$

Define $\Sigma_2^{P_{\text{conf}},\sigma_{\text{fix}}}$ to be the set of strategies for the restricted player, given the possible settings of σ_2'. Among this set of strategies, we can define the subset of best responses to an opponent strategy σ_1, $BR^{P_{\text{conf}},\sigma_{\text{fix}}}(\sigma_1) \subseteq \Sigma_2^{P_{\text{conf}},\sigma_{\text{fix}}}$. Solving a game with the opponent restricted accordingly, finds a strategy profile (σ_1^*, σ_2^*) that is a restricted equilibrium, where $\sigma_1^* \in BR(\sigma_2^*)$ and $\sigma_2^* \in BR^{P_{\text{conf}},\sigma_{\text{fix}}}(\sigma_1)$. In this pair, the strategy σ_1^* is a **P_{conf}-restricted Nash response to the opponent model σ_{fix}**, which we call a data biased response counter-strategy.

5.2 CHOOSING P_{conf}

We will now present four ways in which P_{conf} can be chosen, all of which have two aspects in common. First, each approach sets $P_{\text{conf}}(I)$ for an information set I as a function of the number of observations we have of the opponent acting in information set I, n_I. As the number of observations of our opponent acting in I increase, we will become more confident in the model's accuracy. If $n_I = 0$, then we set $P_{\text{conf}}(I)$ to zero, indicating that we have no confidence in the model's prediction. Note that this choice in setting P_{conf} removes the need for a default policy. As mentioned in Section 5, this means the restricted player will become a worst-case opponent in any information sets for which we have no observations. Second, each approach accepts an additional parameter $P_{\max} \in [0,1]$, which acts in a similar fashion to p in the restricted Nash response technique. It is used to set a maximum confidence for P_{conf}. Varying $P_{\max}$ in the range $[0,1]$ allows us to set a tradeoff between exploitation and exploitability, while n_I indicates places where our opponent model should not be trusted.

Removing the default strategy. First, we consider a simple choice of P_{conf}, which we call the 1-Step function. In information sets where we have never observed the opponent, P_{conf} returns 0; otherwise, it returns $P_{\max}$. This choice of P_{conf} allows us to isolate the modelling error caused by the default policy from the error caused by the opponent model's action probabilities not matching the action probabilities of the actual opponent.

Requiring more observations. Second, we consider another simple choice of P_{conf}, which we call the 10-Step function. In information sets where we have observed the opponent fewer than 10 times, P_{conf} returns 0; otherwise, it returns $P_{\max}$. Thus, it is simply a step function that requires ten observations before expressing any confidence in the model's accuracy.

Linear confidence functions. Third, we consider a middle ground between our two step functions. The 0-10 Linear function returns $P_{\max}$ if $n_I > 10$ and $(n_I \times P_{\max})/10$ otherwise. Thus, as we obtain more observations, the function expresses more confidence in the accuracy of the opponent model.

Curve confidence functions. Fourth, we consider a setting of P_{conf} with a Bayesian interpretation. The s-Curve function returns $P_{\max} \times (n_I/(s + n_I))$ for any constant s; in this experiment, we used $s = 1$. Thus, as we obtain more observations, the function approaches $P_{\max}$. The foundation for this choice of P_{conf} is explained further in Section 7.

6 RESULTS

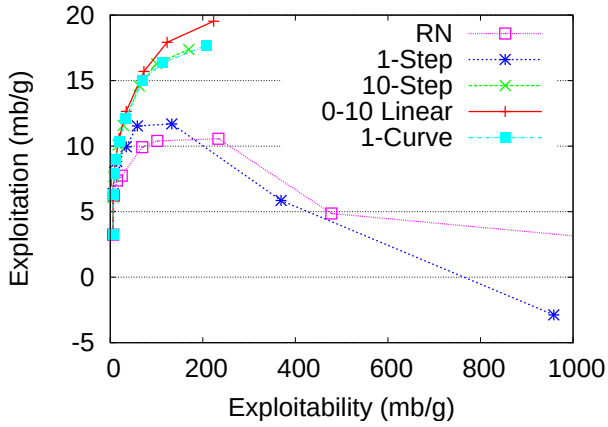
In Section 3, we presented three problems with restricted Nash response strategies. In this section, we will revisit these three problems and show that data biased response counter-strategies overcome these weaknesses. In each experiment, the sets of restricted Nash response and data biased response counter-strategies were created with p and $P_{\max}$ (respectively) parameters of 0, 0.5, 0.7, 0.8, 0.9, 0.93, 0.97, 0.99, and 1. Unless otherwise stated, each set of

counter-strategies was produced with 1 million observed games of Orange playing against Probe.

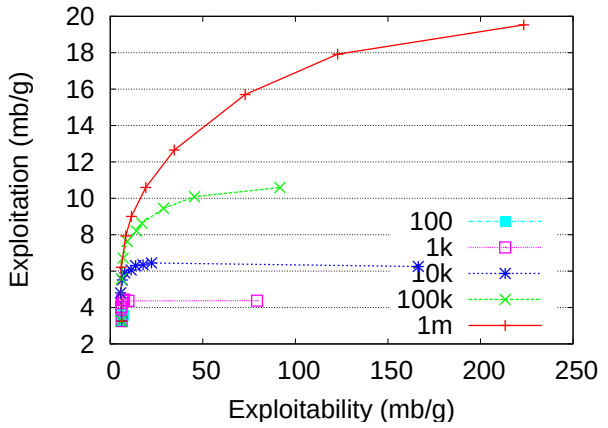
Overfitting to the model. We begin with the problem of overfitting to the model. Figure 2a shows the results of sets of restricted Nash response and 1-Step, 10-Step and 0-1 Linear data biased response counter-strategies playing against Orange and the opponent model of Orange. Two of the results are noteworthy. First, we observe that the set of 1-Step data biased response counter-strategies overfit the model. Since the 1-Step data biased response counter-strategies did not use the default policy, this shows us that the error caused by the opponent model's action probabilities not agreeing with the actual opponent's action probabilities is a nontrivial problem and that the default policy is not the only weakness. Second, we notice that the 0-10 Linear, 10-Step and 1-Curve data biased response counter-strategies do not overfit the opponent model, even at the last datapoint where $P_{\max}$ is set to 1.

Quantity of observations. Next, we examine the problem of the quantity of observations necessary to produce useful counter-strategies. In Figure 1b, we showed that with insufficient quantities of observations, restricted Nash counter-strategies not only did not exploit the opponent but in fact performed worse than a Nash equilibrium strategy (which makes no attempt to exploit the opponent). In Figure 2b, we show that the 0-10 Linear data biased response counter-strategies perform well, regardless of the quantity of observations provided. While the improvement in exploitation from having 100 or 1000 observations is very small, for $P_{\max} < 1$ the counter-strategies became only marginally more exploitable. This is a marked difference from the restricted Nash response results in Figure 1b.

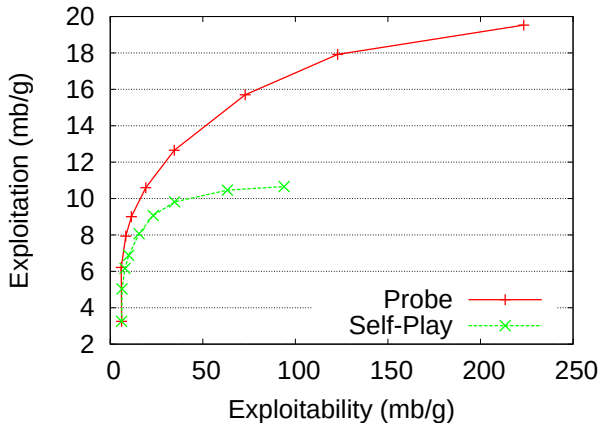
Source of observations. Finally, we consider the problem of the source of the observations used to create the model. In Figure 1c, we showed that the restricted Nash response technique required observations of the opponent playing against an opponent such as Probe in order to create useful counter-strategies. In Figure 2c, we show that while the data biased response counter-strategies produced are more effective when the opponent model observes games against Probe, the technique does still produce useful counter-strategies when provided with self-play data.



(a) Overfitting



(b) Observations



(c) Training

Figure 2: Exploitation versus exploitability curves for data biased response counter-strategies. 2a shows that restricted Nash and 1-Step counter-strategies overfit the model, while 10-Step, 0-10 Linear, and 1-Curve counter-strategies do not. 2b shows that the 0-10 Linear counter-strategies are effective with any quantity of training data. 2c shows that the 0-10 Linear counter-strategies can accept any type of training data. Note that the red, solid curve is the same in each graph.

7 DISCUSSION

We motivated data biased responses by noting that the confidence in our model is not uniform over all information sets, and suggesting p should be some increasing function of the number of observations at a particular information set. We can give an alternative motivation for this approach by considering the framework of Bayesian decision mak-

ing. In the Bayesian framework we choose a prior density function ($f : \Sigma_2 \to \Re$) over the unknown opponent's strategy. Given observations of the opponent's decisions $\mathcal{Z}$ we can talk about the posterior probability $\Pr(\sigma_2|\mathcal{Z}, f)$. If only one more hand is to be played, decision theory instructs us to maximize our expected utility given our beliefs.

$$\operatorname*{argmax}_{\sigma_1} \int_{\sigma_2 \in \Sigma_2} u_1(\sigma_1, \sigma_2) \Pr(\sigma_2|\mathcal{Z}, f) \qquad (2)$$

Since utility is linear in the sequence form representation of strategy, we can move the integral inside the utility function allowing us to solve the optimization as the best-response to the expected posterior strategy (see Footnote 1).

However, instead of choosing a single prior density, suppose we choose a set of priors (F), and we want to play a strategy that would have large utility for anything in this set. A traditional Bayesian approach might require us to specify our uncertainty over priors from this set, and then maximize expected utility given such a hierarchical prior. Suppose, though, that we have no basis for specifying such a distribution over distributions. An alternative then is to maximize utility in the worst case.

$$\operatorname*{argmax}_{\sigma_1} \min_{f \in F} \int_{\sigma_2 \in \Sigma_2} u_1(\sigma_1, \sigma_2) \Pr(\sigma_2|\mathcal{Z}, f) \qquad (3)$$

In other words, employ a strategy that is robust to the choice of prior. Notice that if F contains a singleton prior, this optimization is equivalent to the original decision theoretic approach, *i.e.*, a best response strategy. If F contains all possible prior distributions, then the optimization is identical to the game theoretic approach, *i.e.*, a Nash equilibrium strategy. Other choices of the set F admit optimizations that trade-off exploiting data with avoiding exploitation.

Theorem 1 *Consider F to be the set of priors composed of independent Dirichlet distributions for each information set, where the strength (sum of the Dirichlet parameters) is at most s. The strategy computed by data biased response when $P_{\text{conf}}(I) = n_I/(s + n_I)$ is the solution to the optimization in 3.*

PROOF. *(Sketch)* If we let Σ_2^s be the set of resulting expected posterior strategies for all choices of priors $f \in F$. It suffices to show that $\Sigma_2^s = \Sigma^{P_{\text{conf}}, \sigma_{\text{fix}}}$. For any prior $f \in F$, let $\alpha^f_{I,a}$ be the Dirichlet weight for the outcome a at information set I. Let $\sigma_{\text{fix}}(I, a) = \alpha^f_{I,a} / \sum_{a'} \alpha^f_{I,a'}$, in other words the strategy where the opponent plays the expected prior strategy when given the opportunity. The resulting expected posterior strategy is the the same as σ_2 from Equation 1 and so is in the set $\Sigma^{P_{\text{conf}}, \sigma_{\text{fix}}}$. Similarly, given σ_{fix} associated with a strategy σ_2 in $\Sigma^{P_{\text{conf}}, \sigma_{\text{fix}}}$, let $\alpha_{I,a} = s\sigma_{\text{fix}}(I, a)$. The resulting expected posterior strategy is the same as σ_2. The available strategies to player 2 are equivalent, and so the resulting min-max optimizations are equivalent. ∎

In summary, we can choose P_{conf} in data biased response so that it is equivalent to finding strategies that are robust to a set of independent Dirichlet priors.

8 CONCLUSION

The problem of exploiting information about a suspected tendency in an environment while minimizing worst-case performance occurs in several domains, and becomes more difficult when the information may be limited or inaccurate. We reviewed restricted Nash response counter-strategies, a recent work on the opponent modelling interpretation of this problem in the Poker domain, and highlighted three shortcomings in that approach. We proposed a new technique, data biased responses, for generating robust counter-strategies that provide good compromises between exploiting a tendency and limiting the worst case exploitability of the resulting counter-strategy. We demonstrated that the new technique avoids the three shortcomings of existing approaches, while providing better performance in the most favourable conditions for the existing approaches.

9 ACKNOWLEDGEMENTS

We would like to thank the members of the University of Alberta Computer Poker Research Group. This research was supported by NSERC and iCore.

References

A. Gilpin and T. Sandholm. Finding equilibria in large sequential games of imperfect information. In *ACM Conference on Electronic Commerce*, 2006.

M. Johanson. Robust strategies and counter-strategies: Building a champion level computer poker player, 2007. MSc thesis.

M. Johanson, M. Zinkevich, and M. Bowling. Computing robust counter-strategies. In *Neural Information Processing Systems 21*, 2008.

P. McCracken and M. Bowling. Safe strategies for agent modelling in games. In *AAAI Fall Symposium on Artificial Multi-agent Learning*, October 2004.

M. Zinkevich and M. Littman. The AAAI computer poker competition. *Journal of the International Computer Games Association*, 29, 2006. News item.

M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. In *Neural Information Processing Systems 21*, 2008.