Complexity and Algorithms for Exploiting Quantal Opponents in Large Two-Player Games

David Milec[1], Jakub Černý[2], Viliam Lisý[1], Bo An[2]
[1] AI Center, FEE, Czech Technical University in Prague, Czech Republic
[2] Nanyang Technological University, Singapore
milecdav@fel.cvut.cz, cerny@disroot.org, viliam.lisy@fel.cvut.cz, boan@ntu.edu.sg

Abstract

Solution concepts of traditional game theory assume entirely rational players; therefore, their ability to exploit subrational opponents is limited. One type of subrationality that describes human behavior well is the quantal response. While there exist algorithms for computing solutions against quantal opponents, they either do not scale or may provide strategies that are even worse than the entirely-rational Nash strategies. This paper aims to analyze and propose scalable algorithms for computing effective and robust strategies against a quantal opponent in normal-form and extensive-form games. Our contributions are: (1) we define two different solution concepts related to exploiting quantal opponents and analyze their properties; (2) we prove that computing these solutions is computationally hard; (3) therefore, we evaluate several heuristic approximations based on scalable counterfactual regret minimization (CFR); and (4) we identify a CFR variant that exploits the bounded opponents better than the previously used variants while being less exploitable by the worst-case perfectly-rational opponent.

Introduction

Extensive-form games are a powerful model able to describe recreational games, such as poker, as well as real-world situations from physical or network security. Recent advances in solving these games, and particularly the Counterfactual Regret Minimization (CFR) framework (Zinkevich et al. 2008), allowed creating superhuman agents even in huge games, such as no-limit Texas hold'em with approximately 10^{160} different decision points (Moravčík et al. 2017; Brown and Sandholm 2018). The algorithms generally approximate a Nash equilibrium, which assumes that all players are perfectly rational, and is known to be inefficient in exploiting weaker opponents. An algorithm that would be able to take an opponent's imperfection into account is expected to win by a much larger margin (Johanson and Bowling 2009; Bard et al. 2013).

The most common model of bounded rationality in humans is the quantal response (QR) model (McKelvey and Palfrey 1995, 1998). Multiple experiments identified it as a *good predictor of human behavior* in games (Yang, Ordonez, and Tambe 2012; Haile, Hortaçsu, and Kosenok 2008). QR is also the *hearth of the algorithms successfully deployed in the real world* (Yang, Ordonez, and Tambe 2012; Fang et al. 2017). It suggests that players respond stochastically, picking better actions with higher probability. Therefore, we investigate **how to scalably compute a good strategy against a quantal response opponent in two-player normal-form and extensive-form games.**

If both players choose their actions based on the QR model, their behavior is described by quantal response equilibrium (QRE). Finding QRE is a computationally tractable problem (McKelvey and Palfrey 1995; Turocy 2005), which can be also solved using the CFR framework (Farina, Kroer, and Sandholm 2019). However, when creating AI agents competing with humans, we want to assume that **one of the players is perfectly rational, and only the opponent's rationality is bounded**. A tempting approach may be using the algorithms for computing QRE and increasing one player's rationality or using generic algorithms for exploiting opponents (Davis, Burch, and Bowling 2014) even though the QR model does not satisfy their assumptions, as in (Basak et al. 2018). However, this approach generally leads to a solution concept we call Quantal Nash Equilibrium (QNE), which we show is very inefficient in exploiting QR opponents and may even perform worse than an arbitrary Nash equilibrium.

Since the very nature of the quantal response model assumes that the sub-rational agent responds to a strategy played by its opponent, a more natural setting for studying the optimal strategies against QR opponents are Stackelberg games, in which one player commits to a strategy that is then learned and responded to by the opponent. Optimal commitments against quantal response opponents - Quantal Stackelberg Equilibrium (QSE) - have been studied in security games (Yang, Ordonez, and Tambe 2012), and the results were recently extended to normal-form games (Černý et al. 2020b). Even in these one-shot games, polynomial algorithms are available only for their very limited subclasses. In extensive-form games, we show that computing the QSE is NP-hard, even in zero-sum games. Therefore, it is very unlikely that the CFR framework could be adapted to closely approximate these strategies. Since we aim for high scalability, we focus on empirical evaluation of several heuristics, including using QNE as an approximation of QSE. We identify a method that is not only more exploitative than QNE, but also more robust when the opponent is rational.

Our contributions are: **1)** We analyze the relationship and properties of two solution concepts with quantal opponents that naturally arise from Nash equilibrium (QNE) and Stackelberg equilibrium (QSE). **2)** We prove that computing QNE is PPAD-hard even in NFGs, and computing QSE in EFGs is NP-hard. Therefore, **3)** we investigate the performance of CFR-based heuristics against QR opponents. The extensive empirical evaluation on four different classes of games with up to 10^8 histories identifies a variant of CFR-f (Davis, Burch, and Bowling 2014) that computes strategies better than both QNE and NE.

Background

Even though our main focus is on extensive-form games, we study the concepts in normal-form games, which can be seen as their conceptually simpler special case. After defining the models, we proceed to define quantal response and the metrics for evaluating a deployed strategy's quality.

Two-player Normal-form Games

A two-player normal-form game (NFG) is a tuple $G = (N, A, u)$ where $N = \{\triangle, \triangledown\}$ is set of players. We use i and $-i$ for one player and her opponent. $A = \{A_\triangle, A_\triangledown\}$ denotes the set of ordered sets of *actions* for both players. The *utility function* $u_i : A_\triangle \times A_\triangledown \to \mathbb{R}$ assigns a value for each pair of actions. A game is called zero-sum if $u_\triangle = -u_\triangledown$.

Mixed strategy $\sigma_i \in \Sigma_i$ is a probability distribution over A_i. For any *strategy profile* $\sigma \in \Sigma = \{\Sigma_\triangle \times \Sigma_\triangledown\}$ we use $u_i(\sigma) = u_i(\sigma_i, \sigma_{-i})$ as the expected outcome for player i, given the players follow strategy profile σ. A *best response* (BR) of player i to the opponent's strategy σ_{-i} is a strategy $\sigma_i^{BR} \in BR_i(\sigma_{-i})$, where $u_i(\sigma_i^{BR}, \sigma_{-i}) \geq u_i(\sigma_i', \sigma_{-i})$ for all $\sigma_i' \in \Sigma_i$. An ϵ*-best response* is $\sigma_i^{\epsilon BR} \in \epsilon BR_i(\sigma_{-i}), \epsilon > 0$, where $u_i(\sigma_i^{\epsilon BR}, \sigma_{-i}) + \epsilon \geq u_i(\sigma_i', \sigma_{-i})$ for all $\sigma_i' \in \Sigma_i$. Given a normal-form game $G = (N, A, u)$, a tuple of mixed strategies $(\sigma_i^{NE}, \sigma_{-i}^{NE})$, $\sigma_i^{NE} \in \Sigma_i, \sigma_{-i}^{NE} \in \Sigma_{-i}$ is a *Nash Equilibrium* if σ_i^{NE} is an optimal strategy of player i against strategy σ_{-i}^{NE}. Formally: $\sigma_i^{NE} \in BR(\sigma_{-i}^{NE}) \quad \forall i \in \{\triangle, \triangledown\}$

In many situations, the roles of the players are asymmetric. One player (leader - $\triangle$) has the power to commit to a strategy, and the other player (follower - $\triangledown$) plays the best response. This model has many real-world applications (Tambe 2011); for example, the leader can correspond to a defense agency committing to a protocol to protect critical facilities. The common assumption in the literature is that the follower breaks ties in favor of the leader. Then, the concept is called a Strong Stackelberg Equilibrium (SSE).

A leader's strategy $\sigma^{SSE} \in \Sigma_\triangle$ is a *Strong Stackelberg Equilibrium* if $\sigma_\triangle$ is an optimal strategy of the leader given that the follower best-responds. Formally: $\sigma_\triangle^{SSE} = \arg\max_{\sigma'_\triangle \in \Sigma_\triangle} u_\triangle(\sigma'_\triangle, BR_\triangledown(\sigma'_\triangle))$. In zero-sum games, SSE is equivalent to NE (Conitzer and Sandholm 2006) and the expected utility is denoted *value of the game*.

Two-player Extensive-form Games

A two-player extensive-form game (EFG) consist of a set of players $N = \{\triangle, \triangledown, c\}$, where c denotes the chance. A is a finite set of all actions available in the game. $H \subset \{a_1 a_2 \cdots a_n \mid a_j \in A, n \in \mathbb{N}\}$ is the set of histories in the game. We assume that H forms a non-empty finite prefix tree. We use $g \sqsubset h$ to denote that h extends g. The *root* of H is the empty sequence $\varnothing$. The set of leaves of H is denoted Z and its elements z are called *terminal histories*. The histories not in Z are *non-terminal histories*. By $A(h) = \{a \in A \mid ha \in H\}$ we denote the set of actions available at h. $P : H \setminus Z \to N$ is the *player function* which returns who acts in a given history. Denoting $H_i = \{h \in H \setminus Z \mid P(h) = i\}$, we partition the histories as $H = H_\triangle \cup H_\triangledown \cup H_c \cup Z$. σ_c is the *chance strategy* defined on H_c. For each $h \in H_c, \sigma_c(h)$ is a probability distribution over $A(h)$. Utility functions assign each player utility for each leaf node, $u_i : Z \to \mathbb{R}$. The game is of *imperfect information* if some actions or chance events are not fully observed by all players. The information structure is described by *information sets* for each player i, which form a partition $\mathcal{I}_i$ of H_i. For any information set $I_i \in \mathcal{I}_i$, any two histories $h, h' \in I_i$ are indistinguishable to player i. Therefore $A(h) = A(h')$ whenever $h, h' \in I_i$. For $I_i \in \mathcal{I}_i$ we denote by $A(I_i)$ the set $A(h)$ and by $P(I_i)$ the player $P(h)$ for any $h \in I_i$.

A *strategy* $\sigma_i \in \Sigma_i$ of player i is a function that assigns a distribution over $A(I_i)$ to each $I_i \in \mathcal{I}_i$. A *strategy profile* $\sigma = (\sigma_\triangle, \sigma_\triangledown)$ consists of strategies for both players. $\pi^\sigma(h)$ is the probability of reaching h if all players play according to σ. We can decompose $\pi^\sigma(h) = \prod_{i \in N} \pi_i^\sigma(h)$ into each player's contribution. Let π_{-i}^σ be the product of all players' contributions except that of player i (including chance). For $I_i \in \mathcal{I}_i$ define $\pi^\sigma(I_i) = \sum_{h \in I_i} \pi^\sigma(h)$, as the probability of reaching information set I_i given all players play according to σ. $\pi_i^\sigma(I_i)$ and $\pi_{-i}^\sigma(I_i)$ are defined similarly. Finally, let $\pi^\sigma(h, z) = \frac{\pi^\sigma(z)}{\pi^\sigma(h)}$ if $h \sqsubset z$, and zero otherwise. $\pi_i^\sigma(h, z)$ and $\pi_{-i}^\sigma(h, z)$ are defined similarly. Using this notation, *expected payoff* for player i is $u_i(\sigma) = \sum_{z \in Z} u_i(z) \pi^\sigma(z)$. BR, NE and SSE are defined as in NFGs.

Define $u_i(\sigma, h)$ as an expected utility given that the history h is reached and all players play according to σ. A *counterfactual value* $v_i(\sigma, I)$ is the expected utility given that the information set I is reached and all players play according to strategy σ except player i, which plays to reach I. Formally, $v_i(\sigma, I) = \sum_{h \in I, z \in Z} \pi_{-i}^\sigma(h) \pi^\sigma(h, z) u_i(z)$. And similarly counterfactual value for playing action a in information set I is $v_i(\sigma, I, a) = \sum_{h \in I, z \in Z, ha \sqsubset z} \pi_{-i}^\sigma(ha) \pi^\sigma(ha, z) u_i(z)$.

We define S_i as a set of sequences of actions only for player i. $inf_1(s_i), s_i \in S_i$ is the information set where last action of s_i was executed and $seq_i(I), I \in \mathcal{I}_i$ is sequence of actions of player i to information set I.

Quantal Response Model of Bounded Rationality

Fully rational players always select the utility-maximizing strategy, i.e., the best response. Relaxing this assumption leads to a "statistical version" of best response, which takes into account the inevitable error-proneness of humans and allows the players to make systematic errors (McFadden 1976; McKelvey and Palfrey 1995).

Definition 1. *Let $G = (N, A, u)$ be an NFG. Function $QR : \Sigma_\triangle \to \Sigma_\triangledown$ is a quantal response function of player $\triangledown$ if probability of playing action a monotonically increases as expected utility for a increases. Quantal function QR is* *called canonical if for some real-valued function q:*

$$QR(\sigma, a^k) = \frac{q(u_{\triangledown}(\sigma, a^k))}{\sum_{a^i \in A_{\triangledown}} q(u_{\triangledown}(\sigma, a^i))} \quad \forall \sigma \in \Sigma_{\triangle}, a^k \in A_{\triangledown}. \tag{1}$$

Whenever q is a strictly positive increasing function, the corresponding QR is a valid quantal response function. Such functions q are called *generators* of canonical quantal functions. The most commonly used generator in the literature is the exponential (logit) function (McKelvey and Palfrey 1995) defined as $q(x) = e^{\lambda x}$ where $\lambda > 0$. λ drives the model's rationality. The player behaves uniformly randomly for $\lambda \to 0$, and becomes more rational as $\lambda \to \infty$. We denote a logit quantal function as LQR.

In EFGs, we assume the bounded-rational player plays based on a quantal function in every information set separately, according to the counterfactual values.

Definition 2. *Let G be an EFG. Function $QR : \Sigma_{\triangle} \to \Sigma_{\triangledown}$ is a canonical couterfactual quantal response function of player $\triangledown$ with generator q if for a strategy $\sigma_{\triangle}$ it produces strategy $\sigma_{\triangledown}$ such that in every information set $I \in \mathcal{I}_{\triangledown}$, for each action $a^k \in A(I)$ it holds that*

$$QR(\sigma_{\triangle}, I, a^k) = \frac{q(v_{\triangledown}(\sigma, I, a^k))}{\sum_{a^i \in A(I)} q(v_{\triangledown}(\sigma, I, a^i))}, \tag{2}$$

where $QR(\sigma_{\triangle}, I, a^k)$ is the probability of playing action a^k in information set I and $\sigma = (\sigma_{\triangle}, \sigma_{\triangledown})$.

We denote the canonical counterfactual quantal response with the logit generator *counterfactual logit quantal response (CLQR)*. CLQR differs from the definition of logit agent quantal response (LAQR) (McKelvey and Palfrey 1998) in using counterfactual values instead of expected utilities. The advantage of CLQR over LAQR is that CLQR defines a valid quantal strategy even in information sets unreachable due to a strategy of the opponent, which is needed for applying regret-minimization algorithms explained later.

Because the logit quantal function is the most well-studied function in the literature with several deployed applications (Pita et al. 2008; Delle Fave et al. 2014; Fang et al. 2017), we focus most of our analysis and experimental results on (C)LQR. Without a loss of generality, we assume the quantal player is always player $\triangledown$.

Metrics for Evaluating Quality of Strategy

In a two-player zero-sum game, the *exploitability* of a given strategy is defined as expected utility that a fully rational opponent can achieve above the value of the game. Formally, exploitability $\mathcal{E}(\sigma_i)$ of strategy $\sigma_i \in \Sigma_i$ is $\mathcal{E}(\sigma_i) = u_{-i}(\sigma_i, \sigma_{-i}) - u_{-i}(\sigma^{NE}), \quad \sigma_{-i} \in BR_{-i}(\sigma_i)$.

We also intend to measure how much we are able to exploit an opponent's bounded-rational behavior. For this purpose, we define *gain* of a strategy against quantal response as an expected utility we receive above the value of the game. Formally, gain $\mathcal{G}(\sigma_i)$ of strategy σ_i is defined as $\mathcal{G}(\sigma_i) = u_i(\sigma_i, QR(\sigma_i)) - u_i(\sigma^{NE})$.

General-sum games do not have the property that all NEs have the same expected utility. Therefore, we simply measure expected utility against LQR and BR opponents there.

One-Sided Quantal Solution Concepts

This section formally defines two one-sided bounded-rational equilibria, where one of the players is rational and the other is subrational – a saddle-point-type equilibrium called Quantal Nash Equilibrium (QNE) and a leader-follower-type equilibrium called Quantal Stackelberg Equilibrium (QSE). We show that contrary to their fully-rational counterparts, QNE differs from QSE even in zero-sum games. Moreover, we show that computing QSE in extensive-form games is an NP-hard problem. Full proofs of all our theoretical results are provided in the appendix[1]

Quantal Equilibria in Normal-form Games

We first consider a variant of NE, in which one of the players plays a quantal response instead of the best response.

Definition 3. *Given a normal-form game $G = (N, A, u)$ and a quantal response function QR, a strategy profile $(\sigma_{\triangle}^{QNE}, QR(\sigma_{\triangle}^{QNE})) \in \Sigma$ describes a* Quantal Nash Equilibrium *(QNE) if and only if $\sigma_{\triangle}^{QNE}$ is a best response of player $\triangle$ against quantal-responding player $\triangledown$. Formally:*

$$\sigma_{\triangle}^{QNE} \in BR(QR(\sigma_{\triangle}^{QNE})). \tag{3}$$

QNE can be seen as a concept between NE and Quantal Response Equilibrium (QRE) (McKelvey and Palfrey 1995). While in NE, both players are fully rational, and in QRE, both players are bounded-rational, in QNE, one player is rational, and the other is bounded-rational.

Theorem 1. *Computing a QNE strategy profile in two-player NFGs is a PPAD-hard problem.*

Proof (Sketch). We do a reduction from the problem of computing ϵ-NE (Daskalakis, Goldberg, and Papadimitriou 2009). We derive an upper bound on a maximum distance between best response and logit quantal response, which goes to zero with λ approaching infinity. For a given ϵ, we find λ, such that QNE is ϵ-NE. Detailed proofs of all theorems are provided in the appendix □

QNE usually outperforms NE against LQR in practice as we show in the experiments. However, it cannot be guaranteed as stated in the Proposition 2.

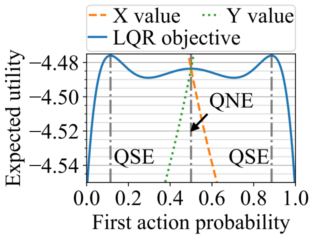
Proposition 2. *For any LQR function, there exists a zero-sum normal-form game $G = (N, A, u)$ with a unique NE - $(\sigma_{\triangle}^{NE}, \sigma_{\triangledown}^{NE})$ and QNE - $(\sigma_{\triangle}^{QNE}, QR(\sigma_{\triangle}^{QNE}))$ such that $u_{\triangle}(\sigma_{\triangle}^{NE}, QR(\sigma_{\triangle}^{NE})) > u_{\triangle}(\sigma_{\triangle}^{QNE}, QR(\sigma_{\triangle}^{QNE}))$.*

The second solution concept is a variant of SSE in situations, when the follower is bounded-rational.

Definition 4. *Given a normal-form game $G = (N, A, u)$ and a quantal response function QR, a mixed strategy $\sigma_{\triangle}^{QSE} \in \Sigma_{\triangle}$ describes a* Quantal Stackleberg Equilibrium *(QSE) if and only if*

$$\sigma_{\triangle}^{QSE} = \arg\max_{\sigma_{\triangle} \in \Sigma_{\triangle}} u_{\triangle}(\sigma_{\triangle}, QR(\sigma_{\triangle})). \tag{4}$$

[1]https://arxiv.org/abs/2009.14521



Game 1

	A	B	C	D
X	-4	-5	8	-4
Y	-5	-4	-4	8

Game 2

	A	B
X	-2	8
Y	-2.2	-2.5

Figure 1: (Left) An example of the expected utility against LQR in Game 1. The X-axis shows the strategy of the rational player, and the Y-axis shows the expected utility. The X- and Y-value curves show the utility for playing the corresponding action given the opponent's strategy is a response to the strategy on X-axis. A detailed description is in Example 1. (Right) An example of two normal-form games. Each row of the depicted matrices is labeled by a first player strategy, while the second player's strategy labels every column. The numbers in the matrices denote the utilities of the first player. We assume the row player is $\triangle$.

In QSE, player $\triangle$ is fully rational and commits to a strategy that maximizes her payoff given that player $\triangledown$ observes the strategy and then responds according to her quantal function. This is a standard assumption, and even in problems where the strategy is not known in advance, it can be learned by playing or observing. QSE always exists because all utilities are finite, the game has a finite number of actions, player $\triangle$ utilities are continuous on her strategy simplex, and the maximum is hence always reached.

Observation 3. *Let G be a normal-form game and a q be a generator of a canonical quantal function. Then QSE of G can be formulated as a non-convex mathematical program:*

$$\max_{\sigma_\triangle \in \Sigma_\triangle} \frac{\sum_{a_\triangledown \in A_\triangledown} u_\triangle(\sigma_\triangle, a_\triangledown) q(u_\triangledown(\sigma_\triangle, \pi_\triangledown))}{\sum_{a_\triangledown \in A_\triangledown} q(u_\triangledown(\sigma_\triangle, a_\triangledown))}. \tag{5}$$

Example 1. *In Figure 1 we depict a utility function of $\triangle$ in Game 1 against LQR with $\lambda = 0.92$. As we show, both actions have the same expected utility in the QNE. Therefore, it is a best response for Player $\triangle$, and she has no incentive to deviate. However, the QNE does not reach the maximal expected utility, which is achieved in two global extremes, both being the QSE. Note that even such a small game like Game 1 can have multiple local extremes: in this case 3.*

Example 1 shows that finding QSE is a non-concave problem even in zero-sum NFGs, and it can have multiple global solutions. Moreover, facing a bounded-rational opponent may change the relationship between NE and SSE. They are no longer interchangeable in zero-sum games, and QSE may use strictly dominated actions, e.g in Game 2 from Figure 1.

Quantal Equilibria in Extensive-form Games

In EFGs, QNE and QSE are defined in the same manner as in NFGs. However, instead of the normal-form quantal response, the second player acts according to the counterfactual quantal response. QSE in EFGs can be computed by a mathematical program provided in the appendix. The natural formulation of the program is non-linear with non-convex constraints indicating the problem is hard. We show that the problem is indeed NP-hard, even in zero-sum games.

Theorem 4. *Let G be a two-player imperfect-information EFG with perfect recall and QR be a quantal response function. Computing a QSE in G is NP-hard if one of the following holds: (1) G is zero-sum and QR is generated by a logit generator $q(x) = exp(\lambda x)$, $\lambda > 0$; or (2) G is general-sum.*

Proof (Sketch). We reduce from the partition problem. The key subtree of the constructed EFG is zero-sum. For each item of the multiset, the leader chooses an action that places the item in the first or second subset. The follower has two actions; each gives the leader a reward equal to the sum of items in the corresponding subset. If the sums are not equal, the follower chooses the lower one because of the zero-sum assumption. Otherwise, the follower chooses both actions with the same probability, maximizing the leader's payoff.

A complication is that the leader could split each item in half by playing uniformly. This is prevented by combining the leader's actions for placing an item with an action in a separate game with two symmetric QSEs. We use a collaborative coordination game in the non-zero-sum case and a game similar to game in Figure 1 in the zero-sum case. □

The proof of the non-zero-sum part of Theorem 4 relies only on the assumption that the follower plays action with higher reward with higher probability. This also holds for a rational player; hence, the theorem provides an independent, simpler, and more general proof of NP-hardness of computing Stackelberg equilibria in EFGs, which unlike (Letchford and Conitzer 2010) does not require the tie-breaking rule.

Computing One-Sided Quantal Equilibria

This section describes various algorithms and heuristics for computing one-sided quantal equilibria introduced in the previous section. In the first part, we focus on QNE, and based on an empirical evaluation; we claim that regret-minimization algorithms converge to QNE in both NFGs and EFGs. The second part then discusses gradient-based algorithms for computing QSE and analyses cases when regret minimization methods will or will not converge to QSE.

Algorithms for Computing QNE

CFR (Zinkevich et al. 2008) is a state-of-the-art algorithm for approximating NE in EFGs. CFR is a form of regret matching (Hart and Mas-Colell 2000) and uses iterated self play to minimize regret at each information set independently. CFR-f (Davis, Burch, and Bowling 2014) is a modification capable of computing strategy against some opponent models. In each iteration, it performs a CFR update for one player and computes the response for the other player. We use CFR-f with a quantal response and call it CFR-QR. We initialize the rational player's strategy as uniform and compute the quantal response against it. Then, in each iteration, we update the regrets for the rational player, calculate the corresponding strategy, and compute a quantal response to this new strategy again. In normal-form games, we use the same approach with simple regret matching (RM-QR).

Conjecture 5. *(a) In NFGs, RM-QR converges to QNE; while (b) in EFGs, CFR-QR converges to QNE.*

Conjecture 5 is based on empirical evaluation on more than 2×10^4 games. In each game, the resulting strategy of player $\triangle$ was ϵ-BR to the quantal response of the opponent with epsilon lower than 10^{-6} after less than 10^5 iterations.

QNE provides strategies exploiting a quantal opponent well, but performance is at the cost of substantial exploitability. We propose two heuristics that address both gain and exploitability simultaneously. The first one is to play a convex combination of QNE and NE strategy. We call this heuristic algorithm **COMB**. We aim to find a parameter α of the combination that maximizes the utility against LQR. However, choosing the correct α is, in general, a non-convex, non-linear problem. We search for the best α by sampling possible αs and choosing the one with the best utility. The time required to compute one combination's value is similar to the time required to perform one iteration of the RM-QR algorithm. Sampling the αs and checking the utility hence does not affect the scalability of COMB. The gain is also guaranteed to be greater or equal to the gain of the NE strategy, and as we show in the results, some combinations achieve higher gains than both the QNE and the NE strategies.

The second heuristic uses a restricted response approach (Johanson, Zinkevich, and Bowling 2008), and we call it **restricted quantal response (RQR)**. The key idea is that during the regret minimization, we set probability p, such that in each iteration, the opponent updates her strategy using (i) LQR with probability p and (ii) BR otherwise. We aim to choose the parameter p such that it maximizes the expected payoff. Using sampling as in COMB is not possible, since each sample requires to rerun the whole RM. To avoid the computations, we start with $p = 0.5$ and update the value during the iterations. In each iteration, we approximate the gradient of gain with respect to p based on a change in the value after both the LQR and the BR iteration. We move the value of p in the gradient's approximated direction with a step size that decreases after each iteration. However, the strategies do change tremendously with p, and the algorithm would require many iterations to produce a meaningful average strategy. Therefore, after a few thousands of iterations, we fix the parameter p and perform a clean second run, with p fixed from the first run. RQR achieves higher gains than both the QNE and the NE and performs exceptionally well in terms of exploitability with gains comparable to COMB.

We adapted both algorithms from NFGs also to EFGs. The COMB heuristic requires to compute a convex combination of strategies, which is not straightforward in EFGs. Let p be a combination coefficient and σ_i^1, $\sigma_i^2 \in \Sigma_i$ be two different strategies for the player i. The convex combination of the strategies is a strategy $\sigma_i^3 \in \Sigma$ computed for each information set $I_i \in \mathcal{I}_i$ and action $a \in A(I_i)$ as follows:

$$\sigma_i^3(I_i)(a) = \frac{\pi_i^{\sigma^1}(I_i)\sigma_i^1(I_i)(a)p + \pi_i^{\sigma^2}(I_i)\sigma_i^2(I_i)(a)(1-p)}{\pi_i^{\sigma^1}(I_i)p + \pi_i^{\sigma^2}(I_i)(1-p)} \quad (6)$$

We search for a value of p that maximizes the gain, and we call this approach the counterfactual COMB. Contrary to COMB, the RQR can be directly applied to EFGs. The idea is the same, but instead of regret matching, we use CFR. We call this heuristic algorithm the counterfactual RQR.

Algorithms for Computing QSE

In general, the mathematical programs describing the QSE in NFGs and EFGs are non-concave and non-linear. We use the gradient ascent (GA) methods (Boyd and Vandenberghe 2004) to find the local optima. For concave program, the GA will reach a global optimum. However, both formulations of QSE contain a fractional part, corresponding to a definition of the follower's canonical quantal function. Because concavity is not preserved under division, accessing conditions of the concavity of these programs is difficult. We construct provably globally optimal algorithms for QSE in our concurrent papers (Černý et al. 2020a,b). The GA performs well on small games, but it does not scale at all even for moderately sized games, as we show in the experiments.

Because QSE and QNE are usually non-equivalent concepts even in zero-sum games (see Figure 1), the regret-minimization algorithms will not converge to QSE. However, in case a quantal function satisfies the so-called *pretty-good-response* condition, the algorithm converges to a strategy of the leader exploiting the follower the most (Davis, Burch, and Bowling 2014). We show that a class of simple (i.e., attaining only a finite number of values) quantal functions satisfy a pretty-good-responses condition.

Proposition 6. *Let $G = (N, A, u)$ be a zero-sum NFG, QR a quantal response function of the follower, which depends only on the ordering of expected utilities of individual actions. Then the RM-QR algorithm converges to QSE.*

An example of a quantal response depending only on the ordering of expected utilities is, e.g., a function assigning probability 0.5 to the action with the highest expected utility, 0.3 to the action with the second-highest utility and $0.2/(|A_2|-2)$ to all remaining actions. The class of quantal functions satisfying the conditions of pretty-good-responses still takes into account the strategy of the opponent (i.e., the responses are not static), but it is limited. In general, quantal functions are not pretty-good-responses.

Proposition 7. *Let QR be a canonical quantal function with a strictly monotonically increasing generator q. Then QR is not a pretty-good-response.*

Experimental Evaluation

The experimental evaluation aims to compare solutions of our proposed algorithm RQR with QNE strategies computed by RM-QR for NFGs and CFR-QR for EFGs. As baselines, we use (i) Nash equilibrium (NASH) strategies, (ii) a best convex combination of NASH and QNE denoted COMB, and (iii) an approximation of QSE computed by gradient ascent (GA), initialized by NASH. We use regret matching+ in the regret-based algorithms. We focus mainly on zero-sum games, because they allow for a more straightforward interpretation of the trade-offs between gain and exploitability. Still, we also provide results on general-sum NFGs. Finally, we show that the performance of RQR is stable over

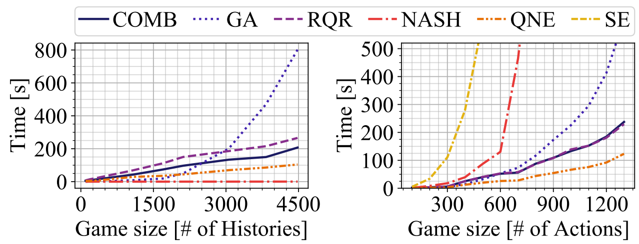


Figure 2: Running time comparison of COMB, GA, RQR, NASH, SE and QNE on (left) square general-sum NFGs and (right) zero-sum EFGs.

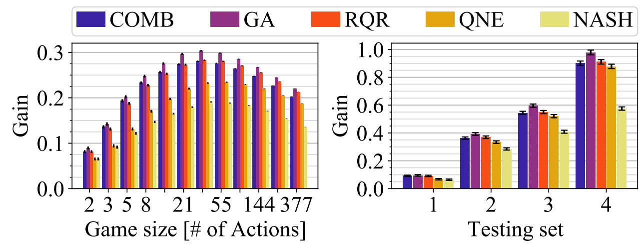


Figure 3: Gain comparison of GA, Nash(SE), QNE, RQR and COMB in (left) random square zero-sum NFGs, and (right) random zero-sum EFGs.

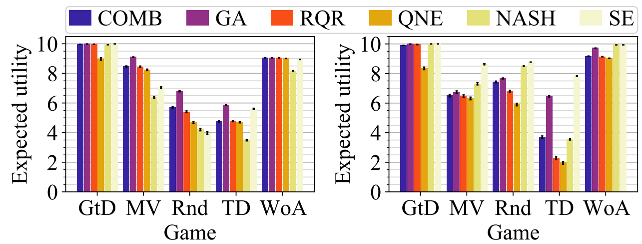


Figure 4: Gain and robustness comparison of GA, Nash, Strong Stackelberg (SE), QNE, COMB and RQR in general sum NFGs – the expected utility (left) against LQR and (right) against BR that maximizes leader's utility.

different rationality values and analyze the EFG algorithms more closely on well-known Leduc Hold'em game. The experimental setup and all the domains are described in the appendix. The implementation is publicly available.[2]

Scalability

The first experiment shows the difference in runtimes of GA and regret-minimization approaches. In NFGs, we use random square zero-sum games as an evaluation domain, and the runtimes are averaged over 1000 games per game size. In EFGs, the random generation procedure does not guarantee games with the same number of histories, so we cluster games with a similar size together, and report runtimes averaged over the clusters. The results on the right of Figure 2 show that regret minimization approaches scale significantly better – the tendency is very similar in both NFGs and EFGs, and we show the results for NFGs in the appendix.

We report scalability in general-sum games on the left in Figure 2. We generated 100 games of Grab the Dollar, Majority Voting, Travelers Dilemma, and War of Attrition with an increasing number of actions for both players and also 100 random general-sum NFGs of the same size. Detailed game description is in the appendix. In the rest of the experiments, we use sets of 1000 games with 100 actions for each class. We use a MILP formulation to compute the NE (Sandholm, Gilpin, and Conitzer 2005) and solve for SE using multiple linear programs (Conitzer and Sandholm 2006). The performance of GA against CFR-based algorithm is similar to the zero-sum case, and the only difference is in NE and SE, which are even less scalable than GA.

Gain Comparison

Now we turn to a comparison of gains of all algorithms in NFGs and EFGs. We report averages with standard errors for zero-sum games in Figure 3 and general-sum games in Figure 4 (left). We use the NE strategy as a baseline, but as different NE strategies can achieve different gains against the subrational opponent, we try to select the best NE strategy. To achieve this, we first compute a feasible NE. Then we run gradient ascent constrained to the set of NE, optimizing the expected value. We aim to show that RQR performs even better than an optimized NE. Moreover, also COMB strategies outperform the best NE, despite COMB using the (possibly suboptimal) NE strategy computed by CFR.

[2]https://gitlab.fel.cvut.cz/milecdav/aaai_qne_code.git

The results show that GA for QSE is the best approach in terms of gain in zero-sum and general-sum games if we ignore scalability issues. The scalable heuristic approaches also achieve significantly higher gain than both the NE baseline and competing QNE in both zero-sum and general-sum games. On top of that, we show that in general-sum games, in all games except one, the heuristic approaches perform as well as or better than SE. This indicates that they are useful in practice even in general-sum settings.

Robustness Comparison

In this work, we are concerned primarily with increasing gain. However, the higher gain might come at the expense of robustness–the quality of strategies might degrade if the model of the opponent is incorrect. Therefore, we study also (i) the exploitability of the solutions in zero-sum games and (ii) expected utility against the best response that breaks ties in our favor in general-sum games. Both correspond to performance against a perfectly rational selfish opponent.

First, we report the mean exploitability in zero-sum games in Figure 5. Exploitability of NE is zero, so it is not included. We show that QNE is highly exploitable in both NFGs and EFGs. COMB and GA perform similarly, and RQR has significantly lower exploitability compared to other modeling approaches. Second, we depict the results in general-sum games on the right in Figure 4. By definition, SE is the optimal strategy and provides an upper bound on achievable value. Unlike in zero-sum games, GA outperforms CFR-based approaches even against the rational opponent. Our heuristic approaches are not as good as entirely rational solution concepts, but they always perform better than QNE.

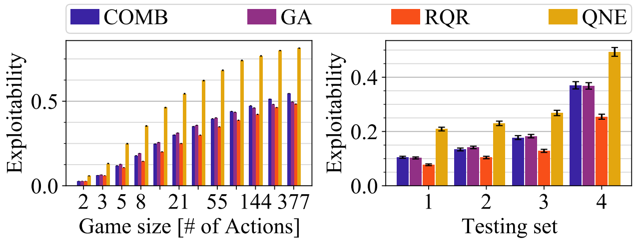


Figure 5: Robustness comparison of GA, QNE, COMB and RQR in (left) random square zero-sum NFGs, and (right) random zero-sum EFGs ($\mathcal{E}(Nash) = 0$).

Different Rationality

In the fourth experiment, we access the algorithms' performance against opponents with varying rationality parameter λ in the logit function. For $\lambda \in \{0, \dots 100\}$ we report the expected utility on the left in Figure 6. For smaller values of λ (i.e., lower rationality), RQR performs similarly to GA and QNE, but it achieves lower exploitability. As rationality increases, the gain of RQR is found between GA and QNE, while having the lowest exploitability. For all values of λ, both QNE and RQR report higher gain than NASH. We do not include COMB in the figure for the sake of better readability as it achieves similar results to RQR.

Standard EFG Benchmarks

Poker. Poker is a standard evaluation domain, and continual resolving was demonstrated to perform extremely well on it (Moravčík et al. 2017). We tested our approaches on two poker variants: one-card poker and Leduc Hold'em. We used $\lambda = 2$ because for $\lambda = 1$, QNE is equal to QSE. We report the values achieved in Leduc Hold'em on the right in Figure 6. The horizontal lines correspond to NE and GA strategies, as they do not depend on p. The heuristic strategies are reported for different p values. The leftmost point corresponds to the CFR-BR strategy and rightmost to the QNE strategy. The experiment shows that RQR performs very well for poker games as it gets close to the GA while running significantly faster. Furthermore, the strategy computed by RQR is much less exploitable consistently throughout various λ values. This suggests that the restricted response can be successfully applied not only against strategies independent of the opponent as in (Johanson, Zinkevich, and Bowling 2008), but also against adapting opponents. We observe similar performance also in the one-card poker and report the results in the appendix.

Goofspiel 7. We demonstrate our approach on Goofspiel 7, a game with almost 100 million histories to show a practical scalability. While CFR-QR, RQR, and CFR were able to compute a strategy, the games of this size are beyond the computing abilities of GA and memory requirements of COMB. CFR-QR has exploitability 4.045 and gains 2.357, RQR has exploitability 3.849 and gains 2.412, and CFR gains 1.191 with exploitability 0.115. RQR hence performs the best in terms of gain and outperforms CFR-QR in exploitability. All algorithms used 1000 iterations.

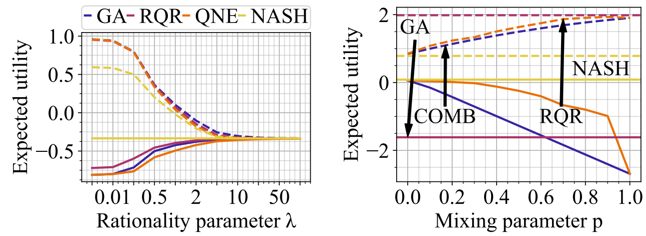


Figure 6: (Left) Mean expected utility against CLQR (dashed) and BR (solid) over 100 games from set 2 of EFGs with changing λ. (Right) Expected utility of different algorithms against CLQR (dashed) and BR (solid) in Leduc Hold'em. QNE is the value of COMB or RQR with $p = 1$.

Summary of the Results

In the experiments, we have shown three main points. (1) GA approach does not scale even to moderate games, making regret minimization approaches much better suited to larger games. (2) In both NFGs and EFGs, the RQR approach outperforms NASH and QNE baseline in terms of gain and outperforms QNE in terms of exploitability, making it currently the best approach against LQR opponents in large games. (3) Our algorithms perform better than the baselines, even with different rationality values, and can be successfully used even in general games. Visual comparison of the algorithms in zero-sum games is provided in the Table 1. Scalability denotes how well the algorithm scales to larger games. The marks range from three minuses as the worst to three pluses as the best; NE is a 0 baseline.

Conclusion

Bounded rationality models are crucial for applications that involve human decision-makers. Most previous results on bounded rationality consider games among humans, where all players' rationality is bounded. However, artificial intelligence applications in real-world problems pose a novel challenge of computing optimal strategies for an entirely rational system interacting with bounded-rational humans. We call this optimal strategy Quantal Stackelberg Equilibrium (QSE) and show that natural adaptations of existing algorithms do not lead to QSE, but rather to a different solution we call Quantal Nash Equilibrium (QNE). As we observe, there is a trade-off between computability and solution quality. QSE provides better strategies, but it is computationally hard and does not scale to large domains. QNE scales significantly better, but it typically achieves lower utility than QSE and might be even worse than the worst Nash equilibrium. Therefore, we propose a variant of CFR which, based on our experimental evaluation, scales to large games, and computes strategies that outperform QNE against both quantal response opponents and perfectly rational opponents.

	COMB	RQR	QNE	NE	GA
Scalability	-	0	0	0	- - -
Gain	++	++	+	0	+++
Exploitability	- -	-	- - -	0	- -

Table 1: Visual comparison of the algorithms

Acknowledgements

This research is supported by Czech Science Foundation (grant no. 18-27483Y) and the SIMTech-NTU Joint Laboratory on Complex Systems. Bo An is partially supported by Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU), which is a collaboration between Singapore Telecommunications Limited (Singtel) and Nanyang Technological University (NTU) that is funded by the Singapore Government through the Industry Alignment Fund – Industry Collaboration Projects Grant. Computation resources were provided by the OP VVV MEYS funded project CZ.02.1.01/0.0/0.0/16 019/0000765 Research Center for Informatics.

References

Bard, N.; Johanson, M.; Burch, N.; and Bowling, M. 2013. Online implicit agent modelling. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, 255–262.

Basak, A.; Černý, J.; Gutierrez, M.; Curtis, S.; Kamhoua, C.; Jones, D.; Bošanský, B.; and Kiekintveld, C. 2018. An initial study of targeted personality models in the flipit game. In *International Conference on Decision and Game Theory for Security*, 623–636. Springer.

Boyd, S.; and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press.

Brown, N.; and Sandholm, T. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* 359(6374): 418–424.

Černý, J.; Lisý, V.; Bošanský, B.; and An, B. 2020a. Computing Quantal Stackelberg Equilibrium in Extensive-Form Games. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*.

Černý, J.; Lisý, V.; Bošanský, B.; and An, B. 2020b. Dinkelbach-Type Algorithm for Computing Quantal Stackelberg Equilibrium. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 246–253. Main track.

Conitzer, V.; and Sandholm, T. 2006. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, 82–90.

Daskalakis, C.; Goldberg, P. W.; and Papadimitriou, C. H. 2009. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing* 39(1): 195–259.

Davis, T.; Burch, N.; and Bowling, M. 2014. Using response functions to measure strategy strength. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Delle Fave, F. M.; Jiang, A. X.; Yin, Z.; Zhang, C.; Tambe, M.; Kraus, S.; and Sullivan, J. P. 2014. Game-theoretic patrolling with dynamic execution uncertainty and a case study on a real transit system. *Journal of Artificial Intelligence Research* 50: 321–367.

Fang, F.; Nguyen, T. H.; Pickles, R.; Lam, W. Y.; Clements, G. R.; An, B.; Singh, A.; Schwedock, B. C.; Tambe, M.; and Lemieux, A. 2017. PAWS-A Deployed Game-Theoretic Application to Combat Poaching. *AI Magazine* 38(1): 23–36.

Farina, G.; Kroer, C.; and Sandholm, T. 2019. Online convex optimization for sequential decision processes and extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1917–1925.

Haile, P. A.; Hortaçsu, A.; and Kosenok, G. 2008. On the empirical content of quantal response equilibrium. *American Economic Review* 98(1): 180–200.

Hart, S.; and Mas-Colell, A. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68(5): 1127–1150.

Johanson, M.; and Bowling, M. 2009. Data biased robust counter strategies. In *Artificial Intelligence and Statistics*, 264–271.

Johanson, M.; Zinkevich, M.; and Bowling, M. 2008. Computing robust counter-strategies. In *Advances in neural information processing systems*, 721–728.

Letchford, J.; and Conitzer, V. 2010. Computing optimal strategies to commit to in extensive-form games. In *Proceedings of the 11th ACM conference on Electronic commerce*, 83–92.

McFadden, D. L. 1976. Quantal choice analaysis: A survey. In *Annals of Economic and Social Measurement, Volume 5, number 4*, 363–390. NBER.

McKelvey, R. D.; and Palfrey, T. R. 1995. Quantal response equilibria for normal form games. *Games and economic behavior* 10(1): 6–38.

McKelvey, R. D.; and Palfrey, T. R. 1998. Quantal response equilibria for extensive form games. *Experimental economics* 1(1): 9–41.

Moravčík, M.; Schmid, M.; Burch, N.; Lisý, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; and Bowling, M. 2017. Deepstack: Expert-level artificial intelligence in Heads-Up No-Limit Poker. *Science* 356(6337): 508–513.

Pita, J.; Jain, M.; Marecki, J.; Ordóñez, F.; Portway, C.; Tambe, M.; Western, C.; Paruchuri, P.; and Kraus, S. 2008. Deployed ARMOR protection: The application of a game theoretic model for security at the Los Angeles International Airport. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, 125–132.

Sandholm, T.; Gilpin, A.; and Conitzer, V. 2005. Mixed-integer programming methods for finding Nash equilibria. In *AAAI*, 495–501.

Tambe, M. 2011. *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. New York, NY, USA: Cambridge University Press. ISBN 1107096421, 9781107096424.

Turocy, T. L. 2005. A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. *Games and Economic Behavior* 51(2): 243–263.

Yang, R.; Ordonez, F.; and Tambe, M. 2012. Computing optimal strategy against quantal response in security games.

In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 847–854.

Zinkevich, M.; Johanson, M.; Bowling, M.; and Piccione, C. 2008. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, 1729–1736.