Computing Approximate Equilibria in Sequential Adversarial Games by Exploitability Descent

Edward Lockhart[1] , **Marc Lanctot**[1] , **Julien Pérolat**[1] , **Jean-Baptiste Lespiau**[1] ,
Dustin Morrill[2] , **Finbarr Timbers**[1] , **Karl Tuyls**[1] ,
[1]DeepMind
[2]University of Alberta
{locked, lanctot, perolat, jblespiau}@google.com, morrill@ualberta.ca,
{finbarrtimbers, karltuyls}@google.com,

Abstract

In this paper, we present *exploitability descent*, a new algorithm to compute approximate equilibria in two-player zero-sum extensive-form games with imperfect information, by direct policy optimization against worst-case opponents. We prove that when following this optimization, the exploitability of a player's strategy converges asymptotically to zero, and hence when both players employ this optimization, the joint policies converge to a Nash equilibrium. Unlike fictitious play (XFP) and counterfactual regret minimization (CFR), our convergence result pertains to the policies being optimized rather than the average policies. Our experiments demonstrate convergence rates comparable to XFP and CFR in four benchmark games in the tabular case. Using function approximation, we find that our algorithm outperforms the tabular version in two of the games, which, to the best of our knowledge, is the first such result in imperfect information games among this class of algorithms.

1 Introduction

Extensive-form games model sequential interactions between multiple agents, each of which maximize their own utility. Classic examples are perfect information games (e.g. chess and Go), which have served as milestones for measuring the progress of artificial intelligence [Campbell *et al.*, 2002; Silver *et al.*, 2016]. When there are simultaneous moves, such as in Markov games, the players may need stochastic policies to guarantee their worst-case expected utility, and must use linear programming at each state for value back-ups. Computing policies for imperfect information games is much more difficult: no Bellman operator exists, so approximate dynamic programming is not applicable; exact equilibrium solutions can be found by sequence-form linear programming [Koller *et al.*, 1994; Shoham and Leyton-Brown, 2009], but these techniques do not scale to very large games.

The challenge domain for imperfect information has been computer Poker, which has driven much of the progress in computational approaches to equilibrium-finding [Rubin and Watson, 2011]. While there are gradient descent techniques that can find an ϵ-Nash equilibrium in $O(\frac{1}{\epsilon})$ iterations [Hoda *et al.*, 2007], the dominant technique has been counterfactual regret minimization (CFR) [Zinkevich *et al.*, 2008]. Based on CFR, recent techniques have solved heads-up limit Texas Hold'em [Bowling *et al.*, 2015] and beat human professionals in no-limit Texas Hold'em [Moravčík *et al.*, 2017; Brown and Sandholm, 2017].

Other techniques have emerged in recent years, based first on fictitious play (XFP) [Heinrich *et al.*, 2015], and generalized to double oracle and any meta-game solver over sets of policies [Lanctot *et al.*, 2017]. Both require a subroutine that computes a best response (an "oracle"). Here, reinforcement learning can be used to compute approximate oracles, and function approximation can be used to generalize over the state space without domain-specific abstraction mechanisms. Hence, deep neural networks can trained from zero knowledge as in AlphaZero [Silver *et al.*, 2018]. Policy gradient techniques are also compatible with function approximation in this setting [Srinivasan *et al.*, 2018], but may require many iterations to converge. Combining data buffers with CFR using regression to predict regrets has also shown promise in medium-sized poker variants [Waugh *et al.*, 2015; Brown *et al.*, 2018].

In this paper, we introduce a new algorithm for computing approximate Nash equilibria. Like XFP, best responses are computed at each iteration. Unlike XFP, players optimize their policies directly against their worst-case opponent. When using tabular policies and ℓ_2 projections after policy updates, the sequence of policies will contain an ϵ-Nash equilibrium, unlike CFR and XFP that converge-in-average, adding the burden of computing average strategies. Our algorithm works well with function approximation, as the optimization can be expressed directly as a policy gradient algorithm. Our experiments show convergence rates comparable to XFP and CFR in the tabular setting, exhibit generalization over the state space using neural networks in four different common benchmark games.

2 Background and Terminology

An **extensive-form game** describes a sequential interaction between **players** $i \in \{1, 2, \cdots, n\} \cup \{c\}$, where c is considered a special player called **chance** with a fixed stochastic policy that determines the transition probabilities given states

and actions. We will often use $-i$ to refer to all the opponents of i. In this paper, we focus on the $n = 2$ player setting.

The game starts in the empty history $h = \emptyset$. On each turn, a player i chooses an action $a \in \mathcal{A}_i$, changing the history to $h' = ha$. Here h is called a prefix history of h', denoted $h \sqsubset h'$. The full history is sometimes also called a *ground* state because it uniquely identifies the true state, since chance's actions are included. In poker, for example, a ground state would include all the players' private cards. We define an **information state** $s \in \mathcal{S}$ for player i as the state as perceived by an agent which is consistent with its observations. Formally, each s is a set of histories, specifically $h, h' \in s \Leftrightarrow$ the sequence of of player i's observations along h and h' are equal. In poker, an information state groups together all the histories that differ only in the private cards of $-i$. Denote $\mathcal{Z}$ the set of terminal histories, each corresponding to the end of a game, and a utility to each player $u_i(z)$ for $z \in \mathcal{Z}$. We also define $\tau(s)$ as the player whose turn it is at s, and $\mathcal{Z}(h)$ the subset of terminal histories that share h as a prefix.

Since players cannot observe the ground state h, policies are defined as $\pi_i : \mathcal{S}_i \to \Delta(\mathcal{A}_i)$, where $\Delta(\mathcal{A}_i)$ is the set of probability distributions over $\mathcal{A}_i$. Each player tries to maximize its expected utility given the initial starting history $\emptyset$. We assume finite games, so every history h is bounded in length. The expected value of a joint policy $\boldsymbol{\pi}$ (all players' policies) for player i is defined as

$$v_{i,\boldsymbol{\pi}} = \mathbb{E}_{z\sim\boldsymbol{\pi}}[u_i(z)], \tag{1}$$

where the terminal histories $z \in \mathcal{Z}$ are composed of actions drawn from the joint policy. We also define state-action values for joint policies. The value $q_{i,\boldsymbol{\pi}}(s, a)$ represents the expected return starting at state s, taking action a, and playing $\boldsymbol{\pi}$:

$$\begin{aligned} q_{i,\boldsymbol{\pi}}(s, a) =& \mathbb{E}_{z\sim\boldsymbol{\pi}}[u_i(z) \mid h \in s, h \sqsubset z] \\ =& \sum_{h\in s, z\in\mathcal{Z}(h)} Pr(h|s) u_i(z) \\ =& \frac{\sum_{h\in s} \eta_{\boldsymbol{\pi}}(h) q_{i,\boldsymbol{\pi}}(h, a)}{\sum_{h\in s} \eta_{\boldsymbol{\pi}}(h)}, \end{aligned} \tag{2}$$

$$\begin{aligned} \text{where } q_{i,\boldsymbol{\pi}}(h, a) &= \mathbb{E}_{z\sim\boldsymbol{\pi}}[u_i(z) \mid ha \sqsubseteq z] \\ &= \sum_{h\in s, z\in\mathcal{Z}(h)} \eta_{\boldsymbol{\pi}}(h, z) u_i(z) \end{aligned}$$

is the expected utility of the ground state-action pair (h, a), and $\eta_{\boldsymbol{\pi}}(h)$ is the probability of reaching h under the policy $\boldsymbol{\pi}$. We make the common assumption that players have **perfect recall**, i.e. they do not forget anything they have observed while playing. Under perfect recall, the distribution of the states can be obtained only from the opponents' policies using Bayes' rule (see [Srinivasan *et al.*, 2018, Section 3.2]).

Each player i tries to find a policy that maximizes their own value $v_{i,\boldsymbol{\pi}}$. However, this is difficult to do independently since the value depends on the joint policy, not just player i's policy. A **best response** policy for player i is defined to be $b_i(\boldsymbol{\pi}_{-i}) \in \text{BR}(\boldsymbol{\pi}_{-i}) = \{\boldsymbol{\pi}_i \mid v_{i,(\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i})} = \max_{\boldsymbol{\pi}'_i} v_{i,(\boldsymbol{\pi}'_i,\boldsymbol{\pi}_{-i})}\}$. Given a joint policy $\boldsymbol{\pi}$, the **exploitability** of a policy $\boldsymbol{\pi}_{-i}$ is how much the other player could gain if they switched to a best response: $\delta_i(\boldsymbol{\pi}) = \max_{\boldsymbol{\pi}'_i} v_{i,(\boldsymbol{\pi}'_i,\boldsymbol{\pi}_{-i})} - v_{i,\boldsymbol{\pi}}$. In two-player zero-sum games, an ϵ-minmax (or ϵ-Nash equilibrium) policy is one where $\max_i \delta_i(\boldsymbol{\pi}) \leq \epsilon$. A Nash equilibrium is achieved when $\epsilon = 0$. A common metric to measure the distance to Nash is $\text{NashConv}(\boldsymbol{\pi}) = \sum_i \delta_i(\boldsymbol{\pi})$.

2.1 Extensive-Form Fictitious Play (XFP)

Extensive-form fictitious play (XFP) is equivalent to standard fictitious play, except that it operates in the extensive-form representation of the game [Heinrich *et al.*, 2015]. In fictitious play, the joint policy is initialized arbitrarily (e.g. uniform random distribution at each information state), and players learn by aggregating best response policies. The extensive-form

Algorithm 1: Fictitious Play

input : $\boldsymbol{\pi}^0$ — initial joint policy
for $t \in \{1, 2, \cdots\}$ **do**
  for $i \in \{1, \ldots, n\}$ **do**
    Compute a best response $\boldsymbol{b}_i^t(\boldsymbol{\pi}_{-i}^{t-1})$
    Update average policy $\boldsymbol{\pi}^t$ to include $\boldsymbol{b}_i^t$

version, XFP, requires game-tree traversals to compute the best responses and specific update rules that account for the reach probabilities to ensure that the updates are equivalent to the classical algorithm, as described in [Heinrich *et al.*, 2015, Section 3]. Fictitious play converges to a Nash equilibrium asymptotically in two-player zero-sum games. Sample-based approximations to the best response step have also been developed [Heinrich *et al.*, 2015] as well as function approximation methods to both steps [Heinrich and Silver, 2016]. Both steps have also been generalized to other best response algorithms and meta-strategy combinations [Lanctot *et al.*, 2017].

2.2 Counterfactual Regret Minimization (CFR)

CFR decomposes the full regret computation over the tree into per information-state regret tables and updates [Zinkevich *et al.*, 2008]. Each iteration traverses the tree to compute the local values and regrets, updating cumulative regret and average policy tables, using a local regret minimizer to derive the current policies at each information state.

The quantities of interest are **counterfactual values**, which are similar to Q-values, but differ in that they weigh only the opponent's reach probabilities, and are not normalized. Formally, let $\eta_{-i,\boldsymbol{\pi}}(h)$ be *only the opponents' contributions* to the probability of reaching h under $\boldsymbol{\pi}$. Then, similarly to equation 2, we define counterfactual values: $q^c_{i,\boldsymbol{\pi}}(s, a) = \sum_{h\in s} \eta_{-i,\boldsymbol{\pi}}(h) q_{i,\boldsymbol{\pi}}(h, a)$, and $v^c_{i,\boldsymbol{\pi}}(s) = \sum_{a\in\mathcal{A}_i} \pi_i(s, a) q^c_{i,\boldsymbol{\pi}}(s, a)$. On each iteration k, with a joint policy $\boldsymbol{\pi}^k$, CFR computes a counterfactual regret $r(s, a) = q^c_{i,\boldsymbol{\pi}^k}(s, a) - v^c_{i,\boldsymbol{\pi}^k}(s)$ for all information states s, and a new policy from the cumulative regrets of (s, a) over the iterations using regret-matching [Hart and Mas-Colell, 2000]. The average policies converge to an ϵ-Nash equilibrium in $O(|\mathcal{S}_i|^2|\mathcal{A}_i|/\epsilon^2)$ iterations.

CFR versus a Best Response Oracle (CFR-BR)

Instead of both players employing CFR (CFR-vs-CFR), each player can use CFR versus their worst-case (best response)

opponent, i.e. simultaneously running CFR-vs-BR and BR-vs-CFR. This is the main idea behind counterfactual regret minimization against a best response (CFR-BR) algorithm [Johanson *et al.*, 2012]. The combined average policies of the CFR players is also guaranteed to converge to an ϵ-Nash equilibrium. In fact, the current strategies also converge with high probability. Our convergence analyses are based on CFR-BR, showing that a policy gradient versus a best responder also converges to an ϵ-Nash equilibrium.

2.3 Policy Gradients in Games

We consider policies $\boldsymbol{\pi}_{\boldsymbol{\theta}} = (\pi_{i,\theta_i})_i$ each policy are parameterized by a vector of parameter $(\theta_i)_i = \boldsymbol{\theta}$. Using the likelihood ratio method, the gradient of $v_{i,\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ with respect to the vector of parameters θ_i is:

$$\nabla_{\theta_i} v_{i,\boldsymbol{\pi}_{\boldsymbol{\theta}}} = \sum_{s \in S_i} \left(\sum_{h \in s} \eta_{\boldsymbol{\pi}}(h) \right) \sum_a \nabla_{\theta_i} \pi_{i,\theta_i}(s,a) q_{i,\boldsymbol{\pi}_{\boldsymbol{\theta}}}(s,a) \tag{3}$$

This result can be seen as an extension of the policy gradient Theorem [Sutton *et al.*, 2000; Glynn and L'ecuyer, 1995; Williams, 1992a; Baxter and Bartlett, 2001] to imperfect information games and has been used under several forms: for a detailed derivation, see [Srinivasan *et al.*, 2018, Appendix D].

The critic ($q_{i,\boldsymbol{\pi}_{\boldsymbol{\theta}}}$) can be estimated in many ways (Monte Carlo Return [Williams, 1992a] or using a critic for instance in [Srinivasan *et al.*, 2018] in the context of games. Then:

$$\theta_i \quad \leftarrow \quad \theta_i + \alpha \sum_{l=0}^{K} \mathbb{1}_{i=\tau(s_l)} \sum_a \nabla_{\theta_i} \pi_{i,\theta_i}(s_l,a) \hat{q}_{i,\boldsymbol{\pi}_{\boldsymbol{\theta}}}(s_l,a),$$

where α is the learning rate used by the algorithm and $\hat{q}_{i,\boldsymbol{\pi}_{\boldsymbol{\theta}}}(s_l,a)$ is the estimation of the return used.

3 Exploitability Descent

We now describe our main contribution, Exploitability Descent (ED), followed by an analysis of its convergence guarantees. Conceptually, the algorithm is uncomplicated and shares the outline of fictitious play: on each iteration, there are two steps that occur for each player. The first step is identical to fictitious play: compute the best response to each player's policy. The second step then performs gradient ascent on the policy to increase each player's utility against the respective best responder (aiming to decrease each player's exploitability). The change in the second step is key and important for two reasons. First, it leads to a convergence of the policies that are being optimized without having to compute an explicit average, which is complex in the sequential setting. Secondly, the policies can now be easily approximated via parameterizations (i.e. using e.g. deep neural networks) and trained using policy gradient ascent without having to store a large buffer of previous data.

The general algorithm is outlined in Algorithm 2, where α^t the learning rate on iteration t. Two steps (lines 6 and 7) are intentionally unspecified: we will show properties for two specific instantiations of this general ED algorithm. The quantity $\mathbf{q}^{\boldsymbol{b}}$ refers to a set of expected values for player $i = \tau(s)$, one

Algorithm 2: Exploitability Descent (ED)

input : $\boldsymbol{\pi}^0$ — initial joint policy
1 **for** $t \in \{1, 2, \cdots\}$ **do**
2 **for** $i \in \{1, \cdots, n\}$ **do**
3 Compute a best response $\boldsymbol{b}_i^t(\boldsymbol{\pi}_{-i}^{t-1})$
4 **for** $i \in \{1, \cdots, n\}, s \in \mathcal{S}_i$ **do**
5 Define $\boldsymbol{b}_{-i}^t = \{\boldsymbol{b}_j^t\}_{j \neq i}$
6 Let $\mathbf{q}^{\boldsymbol{b}}(s) = \textsc{ValuesVsBRs}(\boldsymbol{\pi}_i^{t-1}(s), \boldsymbol{b}_{-i}^t)$
7 $\boldsymbol{\pi}_i^t(s) = \textsc{GradAscent}(\boldsymbol{\pi}_i^{t-1}(s), \alpha^t, \mathbf{q}^{\boldsymbol{b}}(s))$

for each action at s using $\boldsymbol{\pi}_i^{t-1}$ against a set of individual best responses. The GRADIENTASCENT update step unspecified for now as we will describe several forms, but the main idea is to increase/decrease the probability of higher/lower utility actions via the gradients of the value functions, and project back to the space of policies.

3.1 Tabular ED with q-values and ℓ_2 projection

For a vector of $|\mathcal{A}|$ real numbers $\boldsymbol{\theta}_s$, define the **simplex** as $\Delta_s = \{\boldsymbol{\theta}_{s,a} \mid \boldsymbol{\theta}_s \geq \mathbf{0}, \sum_a \theta_{s,a} = 1\}$, and the ℓ_2 projection as $\Pi_{\ell_2}(\boldsymbol{\theta}_s) = \text{argmin}_{\boldsymbol{\theta}'_s \in \Delta_s} ||\boldsymbol{\theta}'_s - \boldsymbol{\theta}_s||_2$.

Let $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ be a joint policy parameterized by $\boldsymbol{\theta}$, and $\boldsymbol{\pi}_{\boldsymbol{\theta}_i}$ refer to the portion of player i's parameters (*i.e.* in tabular form $\{\boldsymbol{\theta}_s \mid \tau(s) = i\}$). Here each parameter is a probability of an action at a particular state: $\theta_{s,a} = \pi_{\boldsymbol{\theta}}(s,a)$. We refer to TabularED($q$, ℓ_2) as an instance of exploitability descent with

$$\mathbf{q}^{\boldsymbol{b}}(s) = \{q_{i,(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{t-1}, \boldsymbol{b}_{-i}^t)}(s,a)\}_{a \in \mathcal{A}}, \tag{4}$$

and the policy gradient ascent update defined to be

$$\begin{aligned} \theta_s^t &= \Pi_{\ell_2}\left(\boldsymbol{\theta}_s^{t-1} + \alpha^t \langle \nabla_{\boldsymbol{\theta}_s} \boldsymbol{\pi}_{\boldsymbol{\theta}}^{t-1}(s), \mathbf{q}^{\boldsymbol{b}}(s) \rangle\right) \\ &= \Pi_{\ell_2}\left(\boldsymbol{\theta}_s^{t-1} + \alpha^t \mathbf{q}^{\boldsymbol{b}}(s)\right), \end{aligned} \tag{5}$$

where the Jacobian $\nabla_{\boldsymbol{\theta}_s} \boldsymbol{\pi}_{\boldsymbol{\theta}}^{t-1}(s)$ is an identity matrix because each parameter $\theta_{s,a}$ corresponds directly to the probability $\pi_{\boldsymbol{\theta}}(s,a)$, and $\langle \cdot, \cdot \rangle$ is the usual matrix inner product.

3.2 Tabular ED with counterfactual values and softmax projection

For some vector of real numbers, $\boldsymbol{\theta}_s$, define softmax$(\boldsymbol{\theta}_s) = \{\Pi_{\text{sm}}(\boldsymbol{\theta}_s)\}_a = \{\exp(\theta_{s,a}) / \sum_{a'} \exp(\theta_{s,a'})\}_a$. Re-using the tabular policy notation from the previous section, we now define a different instance of exploitability descent. We refer to TabularED(q^c, softmax) as the algorithm that specifies $\boldsymbol{\pi}_{\boldsymbol{\theta}}(s) = \Pi_{\text{sm}}(\boldsymbol{\theta}_s)$,

$$\mathbf{q}^{\boldsymbol{b}}(s) = \{q_{i,\boldsymbol{\pi}}^c((\boldsymbol{\pi}_{\boldsymbol{\theta}}^{t-1}, \boldsymbol{b}_{-i}^t), s, a)\}_{a \in \mathcal{A}}, \tag{6}$$

and the policy parameter update as

$$\boldsymbol{\theta}_s^t = \boldsymbol{\theta}_s^{t-1} + \alpha^t \langle \nabla_{\boldsymbol{\theta}_s} \boldsymbol{\pi}_{\boldsymbol{\theta}}^{t-1}(s), \mathbf{q}^{\boldsymbol{b}}(s) \rangle, \tag{7}$$

where $\nabla_{\boldsymbol{\theta}_s} \boldsymbol{\pi}_{\boldsymbol{\theta}}^{t-1}(s)$ represents the Jacobian of softmax.

3.3 Convergence Analyses

We now analyze the convergence guarantees of ED. We give results for two cases: first, in cyclical perfect information games and Markov games, and secondly imperfect information games.

Cyclical Perfect Information Games and Markov Games

The following result extends the policy gradient theorem [Sutton *et al.*, 2000; Glynn and L'ecuyer, 1995; Williams, 1992a; Baxter and Bartlett, 2001] to the zero-sum two-player case. It proves that a generalized gradient of the worst-case value function can be estimated from experience as in the single player case.

Theorem 1 (Policy Gradient in the Worst Case)**.** *The gradient of policy $\boldsymbol{\pi_{\theta_i}}$'s value, $v_{i,(\boldsymbol{\pi_{\theta_i}},\boldsymbol{b})}$, against a best response, $\boldsymbol{\beta} \doteq \boldsymbol{b_{-i}(\pi_{\theta_i})} \in \mathrm{BR}(\pi_{\theta_i})$ is a generalized gradient (see [Clarke, 1975]) of $\boldsymbol{\pi_{\theta_i}}$'s worst-case value function,*

$$\nabla_{\theta_i} v_{i,(\pi_{\theta_i}, b_{-i}(\pi_{\theta_i}))} \in \partial \min_{\pi_{-i}} v_{i,(\pi_{\theta_i},\pi_{-i})}.$$

Proof. The proof uses tools from the non-smooth analysis to properly handle gradients of a non-smooth function. We use the notion of generalized gradients defined in [Clarke, 1975]. The generalized gradient of a Lipschitz function is the convex hull of the limits of the form $\lim \nabla f(x + h_i)$ where $h_i \to 0$. The only assumption we will require is that the parameters of our policy θ_i remains in a compact set Θ_i and that $v_{i,(\pi_{\theta_i},\pi_{-i})}$ is differentiable with respect to θ_i for all π_{-i}.

More precisely we use [Clarke, 1975, Theorem 2.1] to state our result. The theorem requires the function to be uniformly semi-continuous which is the case if the policy is differentiable since the dependence of $v_{i,(\pi_{\theta_i},\pi_{-i})}$ on π_{-i} is polynomial. The function $v_{i,(\pi_{\theta_i},\pi_{-i})}$ is Lipschitz with respect to (θ_i, π_{-i}). The uniform continuity of $\nabla_{\theta_i} v_{i,(\pi_{\theta_i},\pi_{-i})}$ comes from the fact that Θ_i is compact.

Using [Clarke, 1975, Theorem 2.1], we have that $\partial \min_{\pi_{-i}} v_{i,(\pi_{\theta_i},\pi_{-i})}$ is the convex hull of $\{\nabla_{\theta_i} v_{i,\pi_{\theta_i},\beta} | \beta \in \mathrm{BR}(\pi_{\theta_i}))\}$, so $\nabla_{\theta_i} v_{i,\pi_{\theta_i},\beta} \in \partial \min_{\pi_{-i}} v_{i,(\pi_{\theta_i},\pi_{-i})}$.

The proof follows by applying the policy gradient theorem [Baxter and Bartlett, 2001] to $\nabla_{\theta_i} v_{i,\pi_{\theta_i},\beta}$. □

This theorem is a natural extension of the policy gradient theorem to the zero-sum two-player case. As in policy gradient, this process is only guaranteed to converge to a local maximum of the worst case value $\min_{\pi_{-i}} v_{i,(\pi_{\theta_i},\pi_{-i})}$ of the game but not necessarily to an equilibrium of the game. An equilibrium of the game is reached when the two following conditions are met simultaneously: (1) if the policy is tabular and (2) if all states are visited with at least some probability for all policies. This statement is proven in Appendix B.

The method is called exploitability descent because policy gradient in the worst case minimizes exploitability. In a two-player, zero-sum game, if both players generate their policies by independently running ED, NASHCONV is locally minimized.

Lemma 1. *In the two-player zero-sum case, simultaneous policy gradient in the worst case locally minimizes* NASHCONV.

Proof. In a two-player, zero-sum game, NASHCONV reduces to the sum of exploitabilities:

$$\begin{aligned}
&\textsc{NashConv}(\pi_{\theta_i}, \pi_{\theta_{-i}}) = \sum_i \max_{\pi'_i} v_{i,(\pi'_i,\pi_{\theta_{-i}})} - v_{i,\boldsymbol{\pi}} \\
&= \sum_i \max_{\pi'_i} v_{i,(\pi'_i,\pi_{\theta_{-i}})} = \sum_i -\min_{\pi'_i} -v_{i,(\pi'_i,\pi_{\theta_{-i}})} \\
&= -(\min_{\pi'_1} v_{2,(\pi_{\theta_2},\pi'_1)} + \min_{\pi'_2} v_{1,(\pi_{\theta_1},\pi'_2)})
\end{aligned}$$

so doing policy gradient in the worst case independently for all players locally minimizes the sum of exploitabilities and therefore NASHCONV. Formally we have[1]:

$$\begin{aligned}
&\partial_{\boldsymbol{\theta}} \sum_i \max_{\pi'_i} v_{i,(\pi'^i,\pi_{\theta_i}^{-i})} - v_{i,\boldsymbol{\pi}} \qquad (8) \\
&= -\left(\partial_{\theta_2} \min_{\pi'_1} v_{2,(\pi_{\theta_2},\pi'_1)} + \partial_{\theta_2} \min_{\pi'_2} v_{1,(\pi_{\theta_1},\pi'_2)}\right).
\end{aligned}$$

□

Imperfect Information Games

We now examine convergence guarantees in the imperfect information setting. There have two main techniques used to solve adversarial games in this case: the first is to rely on the sequence-form representation of policies which makes the optimization problem convex [Koller *et al.*, 1994; Hoda *et al.*, 2007]. The second is to weight the values by the appropriate reach probabilities, and employ local optimizers [Zinkevich *et al.*, 2008; Johanson *et al.*, 2012]. Both techniques take into account the probability of reaching information states, but the latter allows direct policy update rules and a convenient tabular policy representation.

We prove finite time exploitability bounds for TabularED(q, ℓ_2), and we relate TabularED(q^c, softmax) to a similar algorithm that also has finite time bounds.

The convergence analysis is built upon two previous results: the first is CFR-BR [Johanson *et al.*, 2012]. The second is a very recent result that relates policy gradient optimization in imperfect information games to CFR [Srinivasan *et al.*, 2018]. The result here is also closely related to the optimization against a worst-case opponent [Waugh and Bagnell, 2014, Theorem 4], except our policies are expressed in tabular (*i.e.* behavioral) form rather than the sequence form.

Case: TabularED(q, ℓ_2)

Recall that the parameters $\boldsymbol{\theta} = \{\theta_{s,a}\}_{s\in\mathcal{S},a\in\mathcal{A}}$ correspond to the tabular policy, i.e. one parameter per information state and action. For convenience, let $\boldsymbol{\theta}_s = \{\theta_{s,a}\}_{a\in\mathcal{A}}$.

Definition 1. *Recall Δ_s and Π_{ℓ_2} from Section 3.1. Define **strong policy gradient policy iteration against specific opponent(s)**, SPGPI($\boldsymbol{\pi}_{-i}$), to be the process where player i updates their policy using state-localized gradient ascent while all other players use $\boldsymbol{\pi}_{-i}$, according to Eq 5.*

[1] Usually one would have $\partial_{\boldsymbol{\theta}}(f_1 + f_2) \subset \partial_{\boldsymbol{\theta}} f_1 + \partial_{\boldsymbol{\theta}} f_2$. But since in our case the functions are defined on two different sets of parameters, we have an equality.

This is a form of strong policy gradient policy iteration (SPGPI) defined in [Srinivasan *et al.*, 2018, Theorem 2] that separates the optimization for each player. Also notice that an iteration of TabularED(q, ℓ_2) is equivalent to n simultaneous applications of SPGPI($b_{-i}(\boldsymbol{\pi}_i)$).

Definition 2. *Suppose all players use a sequence of joint policies $\{\boldsymbol{\pi}^1, \cdots, \boldsymbol{\pi}^T\}$ over T iterations. Define player i's* ***regret*** *after T iterations to be the difference in expected utility between the best possible policy in hindsight and the expected utility given the sequence of policies:*

$$R_i^T = \max_{\boldsymbol{\pi}_i} \sum_{t=1}^{T} \left(v_{i,(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^t)} - v_{i,\boldsymbol{\pi}^t} \right)$$

Theorem 2. *[Srinivasan* et al.*, 2018, Theorem 2] Suppose players play a finite game using joint policies $\{\boldsymbol{\pi}^1, \cdots, \boldsymbol{\pi}^T\}$ over T iterations. In a two-player zero-sum game, if $\forall s, a \in \mathcal{S} \times \mathcal{A} : \theta_{s,a} > 0$ and $\alpha^t = t^{-\frac{1}{2}}$, then the regret of SPGPI($\boldsymbol{\pi}_{-i}$) after T iterations is $R_i^T \leq \epsilon$, where $\epsilon = |\mathcal{S}_i| \left(\sqrt{T} + (\sqrt{T} - \frac{1}{2}) |\mathcal{A}_i| (\Delta_{u_i})^2 \right)$, and $\Delta_u = \max_{z,z' \in \mathcal{Z}} (u_i(z) - u_i(z'))$.*

Note that, despite the original application of policy gradients in self-play, it follows from the original proof that the statement about the regret incurred by player i does not require a specific algorithm generate the opponents' policies: it is only a function of the specific sequence of opponent policies. In particular, they could be best response policies, and so SPGPI($b_{-i}(\boldsymbol{\pi}_i)$) has the same regret guarantee.

We need one more lemma before we prove the convergence guarantee of Tabular ED. The following lemma states an optimality bound of the best iterate under time-independent loss functions equal to the average regret. The best strategy in a no-regret sequence of strategies then approaches an equilibrium strategy over time without averaging and without probabilistic arguments.

Lemma 2. *Denote $[[T]] = \{1, 2, \cdots, T\}$. For any sequence of T iterates, $\{\boldsymbol{\theta}^t \in \Theta\}_{t \in [[T]]}$, from decision set Θ, the regret of this sequence under loss, $\ell(\boldsymbol{\theta}) : \Theta \to \Re$, is*

$$R^T = \sum_{t=1}^{T} \ell(\boldsymbol{\theta}^t) - \inf_{\boldsymbol{\theta}^* \in \Theta} \sum_{t=1}^{T} \ell(\boldsymbol{\theta}^*).$$

Then, the iterates with the lowest loss, $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{t \in [[T]]} \ell(\boldsymbol{\theta}^t)$, has an optimality gap bounded by the average regret:

$$\ell(\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta}^* \in \Theta} \ell(\boldsymbol{\theta}^*) \leq \frac{R^T}{T}.$$

Proof. Since ℓ is fixed (not varying with t),

$$\begin{aligned} R^T &= \sum_{t=1}^{T} \ell(\boldsymbol{\theta}^t) - T \inf_{\boldsymbol{\theta}^* \in \Theta} \ell(\boldsymbol{\theta}^*) \\ &\geq T(\min_{t \in [[T]]} \ell(\boldsymbol{\theta}^t) - \inf_{\boldsymbol{\theta}^* \in \Theta} \ell(\boldsymbol{\theta}^*)) \\ &= T(\ell(\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta}^* \in \Theta} \ell(\boldsymbol{\theta}^*)), \end{aligned}$$

and dividing by T yields the result. □

In finite games, we can replace the $\inf$ operation in Lemma 2 with $\min$ when the decision set is the set of all possible strategies, since this set is closed.

We now show that if all players optimize their policy in this way, the combined joint policy converges to an approximate Nash equilibrium.

Theorem 3. *Let TabularED(q, ℓ_2) be described as in Section 3.1 using tabular policies and the update rule in Definition 1. In a two-player zero-sum game, if each player updates their policy simultaneously using TabularED(q, ℓ_2), under the conditions on $\theta_{s,a}$ and α^t in Theorem 2, then for each player i: after T iterations, a policy $\boldsymbol{\pi}_i^* \in \{\boldsymbol{\pi}_i^1, \cdots, \boldsymbol{\pi}_i^T\}$ will have been generated such that $\boldsymbol{\pi}_i^*$ is i's part of a $\frac{2\epsilon}{T}$-Nash equilibrium, where ϵ is defined as in Theorem 2.*

Proof. The first part of the proof follows the logic of the CFR-BR proofs [Johanson *et al.*, 2012]. Unlike CFR-BR, we then use Lemma 2 to bound the quality of the best iterate.

SPGPI($b_{-i}(\boldsymbol{\pi}_i)$), has bounded regret sublinear in T for player i by Theorem 2. Define loss function, $\ell_i(\boldsymbol{\pi}_i) = -\min_{\boldsymbol{\pi}_{-i}} v_{i,(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i})}$, as the negated worst-case value for player i, like that described by [Waugh and Bagnell, 2014, Theorem 4]. Then by Lemma 2 and Theorem 2 we have, for the best iterate:

$$\ell_i(\boldsymbol{\pi}^*) - \min_{\boldsymbol{\pi}_i'} \ell_i(\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}) \leq \frac{R^T}{T} \leq \frac{\epsilon}{T} \Rightarrow \delta_i(\boldsymbol{\pi}^*) \leq \frac{\epsilon}{T},$$

where $\delta_i(\boldsymbol{\pi})$ is the exploitability defined in Section 2.

The Nash equilibrium approximation bound is just the sum of the explioitabilities, so when both $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_{-i}$ are returned from ED, they form a $\frac{2\epsilon}{T}$-equilbrium.

□

ED is computing best responses each round already, so it is easy to track the best iterate: it will simply be the one with the highest expected value versus the opponent's best response.

The proof can also be applied to the original CFR-BR theorem, so we now present an improved guarantee, whereas the original CFR-BR theorem made a probabilistic guarantee.

Corollary 1. *(Improved [Johanson* et al.*, 2012, Theorem 4]) If player i plays T iterations of CFR-BR, then it will have generated a $\boldsymbol{\pi}_i^* \in \{\boldsymbol{\pi}^1, \boldsymbol{\pi}^2, \cdots, \boldsymbol{\pi}^T\}$, where $\boldsymbol{\pi}_i^*$ is a 2ϵ-equilibrium, where ϵ is defined as in [Johanson* et al.*, 2012, Theorem 3].*

The best iterate can be tracked in the same way as ED, and the convergence is guaranteed.

Remark 1. *There is a caveat when using q-values: the values are normalized by a quantity, $\mathcal{B}_{-i}(\boldsymbol{\pi}, s) = \sum_h \eta_{-i}^{\boldsymbol{\pi}}(h)$, that depends on the opponents' policies [Srinivasan* et al.*, 2018, Section 3.2]. The convergence guarantee of TabularED(q, ℓ_2) relies on the regret bound of SPGPI, whose proof includes a division by $\mathcal{B}_{-i}(\boldsymbol{\pi}, s)$ [Srinivasan* et al.*, 2018, Appendix E.2]. Therefore, the regret bound is undefined when $\mathcal{B}_{-i}(\boldsymbol{\pi}, s) = 0$, which can happen when an opponent no longer plays to reach s with positive probability.*

Case: TabularED(q^c, ℓ_2)

Instead of using q-values, we can implement ED with counterfactual values. In this case, TabularED with the ℓ_2 projection becomes CFR-BR(GIGA):

Theorem 4. *Let TabularED(q^c, ℓ_2) be described as in Section 3.1 using tabular policies and the following update rule:*

$$\boldsymbol{\pi}_i^t(s) = \Pi_{\ell_2}\left(\boldsymbol{\pi}_i^{t-1}(s) + \alpha^t \boldsymbol{q}^{c,\boldsymbol{b}}(s)\right).$$

This update rule is identical to that of generalized infinitesimal gradient ascent (GIGA) [Zinkevich, 2003] at each information state with best response counterfactual values. CFR-BR(GIGA) therefore performs the same updates and the two algorithms coincide.

With step sizes $\alpha^t = t^{-\frac{1}{2}}, 0 < t \leq T$, each local GIGA instance has regret after T iterations upper bounded by

$$\sqrt{T} + (\sqrt{T} - \frac{1}{2})|\mathcal{A}_i|(\Delta_{u_i})^2,$$

where $\Delta_u = \max_{z,z' \in \mathcal{Z}}(u_i(z) - u_i(z'))$ [Srinivasan et al.*, 2018, Lemma 5]. By the CFR Theorem [Zinkevich* et al.*, 2008], the total regret of CFR-BR(GIGA) (and thus TabularED(q^c, ℓ_2)) is then*

$$R_i^T \leq |\mathcal{S}_i|\left(\sqrt{T} + (\sqrt{T} - \frac{1}{2})|\mathcal{A}_i|(\Delta_{u_i})^2\right).$$

.

This change allows us to avoid the issues discussed in Remark 1.

Case: TabularED(q^c, softmax)

We now relate TabularED with counterfactual values and softmax policies closely to an algorithm with known finite time convergence bounds. We present here only a high-level overview; for details, see Appendix A.

TabularED(q^c, softmax) is still a policy gradient algorithm: it differentiates the policy (*i.e.* softmax function) with respect to its parameters, and updates in the direction of higher value. With two subtle changes to the overall process, we can show that the algorithm would become CFR-BR using hedge [Freund and Schapire, 1997] as a local regret minimizer. CFR with hedge is known to have a better bound, but has typically not performed as well as regret matching in practice, though it has been shown to work better when combined with pruning based on dynamic probability thresholding [Brown *et al.*, 2017].

Instead of policy gradient, one can use a softmax projection over the the sum of action values (or regrets) over time, which are the gradients of the *value function* with respect to the policy. Accumulating the gradients in this way, the algorithm can be recognized as Mirror Descent [Nemirovsky and Yudin, 1983], which also coincides with hedge given the softmax projection [Beck and Teboulle, 2003]. When using the counterfactual values, ED then turns into CFR-BR(hedge). Then CFR-BR(hedge) converges for the same reasons as CFR-BR(regret-matching).

We do not have a finite time bound of the exploitability of TabularED(q^c, softmax) as we do for the same algorithm with an ℓ_2 projection or CFR-BR(hedge). But since TabularED(q^c, softmax) is a policy gradient algorithm, its policy will be adjusted toward a local optimum upon each update and will converge at that point when the gradient is zero. We use this algorithm because the policy gradient formulation allows for easily-applicable general function approximation inspired by reinforcement learning.

4 Experimental Results

We now present our experimental results. We start by comparing empirical convergence rates to XFP and CFR in the tabular setting, following by convergence behavior when training neural network functions to approximate the policy.

In our initial experiments, we found that using q-values led to plateaus in convergence in some cases, possibly due to numerical instability caused by the problem outlined in Remark 1. Therefore, we present results only using TabularED(q^c, softmax), which for simplicity we refer to as TabularED for the remainder of this section. We also found that the algorithm converged faster with slightly higher learning rates than the ones suggested by Section 3.3.

4.1 Experiment Domains

Our experiments are run across four different imperfect information games (see [Kuhn, 1950], [Southey *et al.*, 2005], and [Lanctot, 2013, Chapter 3]).

Kuhn poker is a simplified poker game first proposed by Harold Kuhn [Kuhn, 1950]. Each player antes a single chip, and gets a single private card from a totally-ordered 3-card deck, e.g.$\{J, Q, K\}$. There is a single betting round limited to one raise of 1 chip, and two actions: pass (check/fold) or bet (raise/call). If a player folds, they lose their commitment (2 if the player made a bet, otherwise 1). If neither player folds, the player with the higher card wins the pot (2, 4, or 6 chips). The utility for each player is defined as the number of chips after playing minus the number of chips before playing.

Leduc poker is significantly larger game with two rounds and a 6-card deck in two suits, e.g. {JS,QS,KS, JH,QH,KH}. Like Kuhn, each player initially antes a single chip to play and obtains a single private card and there are three actions: fold, call, raise. There is a fixed bet amount of 2 chips in the first round and 4 chips in the second round, and a limit of two raises per round. After the first round, a single public card is revealed. A pair is the best hand, otherwise hands are ordered by their high card (suit is irrelevant). Utilities are defined similarly to Kuhn poker.

Liar's Dice(1,1) is dice game where each player gets a single private die in $\{\boxdot, \boxdot, \cdots, \boxplus\}$, rolled at the beginning of the game. The players then take turns bidding on the outcomes of both dice, i.e. with bids of the form q-f referring to quantity and face, or calling "Liar". The bids represent a claim that there are at least q dice with face value f among both players. The highest die value, $\boxplus$, counts as a wild card matching any value. Calling "Liar" ends the game, then both players reveal their dice. If the last bid is not satisfied, then the player who called "Liar" wins. Otherwise, the other player wins. The winner receives +1 and loser -1.

Goofspiel, or the Game of Pure Strategy, is a bidding card game where players are trying to obtain the most points. shuffled and set face-down. Each turn, the top point card is revealed, and players simultaneously play a bid card; the point card is given to the highest bidder or discarded if the bids are equal. In this implementation, we use a fixed deck of decreasing points. In this paper, we use $K = 4$ and an imperfect information variant where players are only told whether they have won or lost the bid, but not what the other player played.

4.2 Tabular Convergence Results

We now present empirical convergence rates to ϵ-Nash equilibria. The main results are depicted in Figure 1.

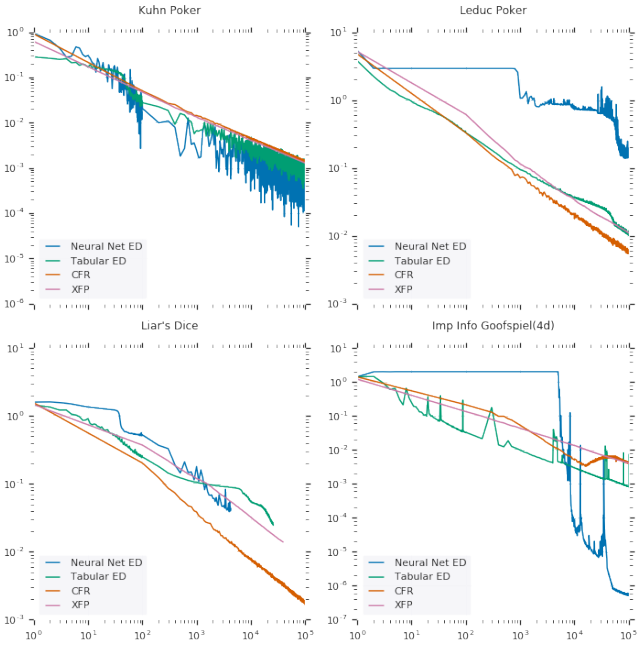

Figure 1: Extensive-form fictitious play (XFP), CFR, tabular and neural-net ED. The y-axis is NASHCONV defined in Section 2, and the x-axis is number of iterations, both in log-scale.

4.3 Neural Network Policies

For the neural network experiments, we use a single policy network for both players, which takes as input the current state of the game and whose output is a softmax distribution over the actions of the game.

The state of the game is represented in a game-dependent fashion as a fixed-size vector of between 11 and 52 binary bits, encoding public information, private information, and the game history.

The neural network consists of a number of fully-connected hidden layers, each with the same number of units and with rectified linear activation functions after each layer. A linear output layer maps from the final hidden layer to a value per action. The values for the legal actions are selected and mapped to a policy using the softmax function.

At each step, we evaluate the policy for every state of the game, compute a best response to it, and evaluate each action against the best response. We then perform a single gradient descent step on the loss function: $-\sum_s \boldsymbol{\pi}_i(s) \cdot \big(\mathbf{q}^b(s) - B(s)\big) + w_r \frac{1}{n} \sum_i \boldsymbol{\theta}_i^2$, where the final term is a regularization for all the neural network weights, and the baseline $B(s)$ is a computed constant (i.e. it does not contribute to the gradient calculation) with $B(s) = \boldsymbol{\pi}_i(s) \cdot \mathbf{q}^{\boldsymbol{b}}(s)$. We performed a sweep over the number of hidden layers (from 1 to 5), the number of hidden units (64, 128 or 256), the regularization weight ($10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}$), and the initial learning rate (powers of 2). The plotted results show the best values from this sweep for each game.

4.4 Discussion

There are several interesting observations to make about the results. First, it appears that the convergence of the neural network policies is more erratic than the tabular counterparts. This is to be expected, as the network policies are generalizing over the entire state space. However, in two of the four games, the neural network policies have learned *more accurate* approximate equilibria than any of the tabular algorithms given the same number of iterations. This is a promising result: the network could be generalizing across the state space (discovering patterns) in a way that is not possible in the tabular case, despite raw input features. To the best of our knowledge, this is the first such result of its form in imperfect information games among this class of algorithms.

We observe that although Tabular ED and XFP have roughly the same convergence rate, the respective function approximation versions of the two algorithms have an order of magnitude difference in speed, with Neural ED reaching an exploitability of 0.08 in Leduc Poker after 10^5 iterations, a level which NFSP reaches after approximately 10^6 iterations [Heinrich and Silver, 2016]. Neural ED and NFSP are not directly comparable as NFSP is computing an approximate equilibrium using sampling and RL while ED uses true best response. However, NFSP uses a reservoir buffer dataset of 2 million samples, whereas this is not necessary in ED. The convergence speed difference might indicate that there is a worthwhile trade-off to consider (in space and convergence time) between fewer iterations with better/best responses and more iterations with approximate responses.

5 Conclusion

We introduced a policy gradient ascent algorithm, exploitability descent (ED), that optimizes its policy directly against worst-case opponents. In cyclical perfect information and Markov games, we prove that ED policies converge to strong policies that are unexploitable in the tabular case. In imperfect information games, we also present finite time exploitability bounds for tabular policies, which imply Nash equilibrium approximation bounds for a complete profile of ED polices. While the empirical convergence rates using tabular policies are comparable to fictitious play and CFR, the policies themselves provably converge. So, unlike XFP and CFR, there is no need to compute the average policy. In addition, neural network function approximation is applicable via direct

policy gradient ascent (whereas computing an average policy is difficult with neural networks), also avoiding the need for domain-specific abstractions, or the need to store large replay buffers of past experience, as in neural fictitious self-play [Heinrich and Silver, 2016], or a set of past networks, as in PSRO [Lanctot *et al.*, 2017].

In some of our experiments, the neural networks learned lower-exploitability policies than the tabular counterparts given the same number of iterations, which could be an indication of strong generalization potential by recognizing similar patterns across states.

There are interesting directions for future work: for example, using approximate best responses and sampling trajectories for the policy optimization in larger games where enumerating the trajectories is not feasible.

Acknowledgments

We would like to thank Neil Burch and Johannes Heinrich for helpful feedback on early drafts of this paper. This research was supported by The Alberta Machine Intelligence Institute (Amii) and Alberta Treasury Branch (ATB).

References

[Baxter and Bartlett, 2001] Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[Beck and Teboulle, 2003] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, May 2003.

[Bowling *et al.*, 2015] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up Limit Hold'em Poker is solved. *Science*, 347(6218):145–149, January 2015.

[Brown and Sandholm, 2017] Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 360(6385), December 2017.

[Brown *et al.*, 2017] Noam Brown, Christian Kroer, and Tuomas Sandholm. Dynamic thresholding and pruning for regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[Brown *et al.*, 2018] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. *CoRR*, abs/1811.00164, 2018.

[Campbell *et al.*, 2002] M. Campbell, A. J. Hoane, and F. Hsu. Deep blue. *Artificial Intelligence*, 134:57–83, 2002.

[Clarke, 1975] Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

[Freund and Schapire, 1997] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[Glynn and L'ecuyer, 1995] Peter W Glynn and Pierre L'ecuyer. Likelihood ratio gradient estimation for stochastic recursions. *Advances in applied probability*, 1995.

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

[Hazan, 2015] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325, 2015.

[Heinrich and Silver, 2016] Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. *CoRR*, abs/1603.01121, 2016.

[Heinrich *et al.*, 2015] Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *ICML 2015*, 2015.

[Hoda *et al.*, 2007] S. Hoda, A. Gilpin, and J. Pe na. A gradient-based approach for computing Nash equilibria of large sequential games. *Optimization Online*, July 2007. http://www.optimization-online.org/DB_HTML/2007/07/1719.html.

[Johanson *et al.*, 2012] M. Johanson, N. Bard, N. Burch, and M. Bowling. Finding optimal abstract strategies in extensive form games. In *Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI)*, pages 1371–1379, 2012.

[Koller *et al.*, 1994] D. Koller, N. Megiddo, and B. von Stengel. Fast algorithms for finding randomized strategies in game trees. In *Proceedings of the 26th ACM Symposium on Theory of Computing (STOC '94)*, pages 750–759, 1994.

[Kuhn, 1950] H. W. Kuhn. Simplified two-person Poker. *Contributions to the Theory of Games*, 1:97–103, 1950.

[Lanctot *et al.*, 2017] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.

[Lanctot, 2013] Marc Lanctot. *Monte Carlo Sampling and Regret Minimization for Equilibrium Computation and Decision-Making in Large Extensive Form Games*. PhD thesis, Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada, June 2013.

[Moravčík *et al.*, 2017] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 358(6362), October 2017.

[Nemirovsky and Yudin, 1983] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[Rubin and Watson, 2011] J. Rubin and I. Watson. Computer poker: A review. *Artificial Intelligence*, 175(5–6):958–987, 2011.

[Shalev-Shwartz and others, 2012] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[Shoham and Leyton-Brown, 2009] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484—-489, 2016.

[Silver *et al.*, 2018] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 632(6419):1140–1144, 2018.

[Southey *et al.*, 2005] Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: Opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertaintyin Artificial Intelligence (UAI)*, pages 550–558, 2005.

[Srinivasan *et al.*, 2018] Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Pérolat, Karl Tuyls, Rémi Munos, and Michael Bowling. Actor-critic policy optimization in partially observable multiagent environments. In *Advances in Neural Information Processing Systems*, 2018.

[Sutton and Barto, 2018] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

[Sutton *et al.*, 2000] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 2000.

[Waugh and Bagnell, 2014] Kevin Waugh and J. Andrew Bagnell. A unified view of large-scale zero-sum equilibrium computation. *CoRR*, abs/1411.5007, 2014.

[Waugh *et al.*, 2015] Kevin Waugh, Dustin Morrill, J. Andrew Bagnell, and Michael Bowling. Solving games with functional regret estimation. In *Proceedongs of the AAAI Conference on Artificial Intelligence*, 2015.

[Williams, 1992a] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[Williams, 1992b] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992.

[Zinkevich *et al.*, 2008] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2008.

[Zinkevich, 2003] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of Twentieth International Conference on Machine Learning (ICML-2003)*, 2003.

Appendices

A Connections and Differences Between Gradient Descent, Mirror Descent, Policy Gradient, and Hedge

There is a broad class of algorithms that attempt to make incremental progress on an optimization task by moving parameters in the direction of a gradient. This elementary idea is intuitive, requiring only basic knowledge of calculus and functions to understand in abstract. One way to more formally justify this procedure comes from the field of online convex optimization [Zinkevich, 2003; Hazan, 2015; Shalev-Shwartz and others, 2012]. The linearization trick [Shalev-Shwartz and others, 2012] reveals how simple parameter adjustments based on gradients can be used to optimize complicated non-linear functions. Perhaps the most well known learning rule is that of **gradient descent**: $\boldsymbol{x} \leftarrow \boldsymbol{x} - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x})$, where the goal is to minimize function f.

Often problems will include constraints on $\boldsymbol{x}$, such as the probability simplex constraint required of decision policies. One often convenient way to approach this problem is project unconstrained parameters, $\boldsymbol{\theta}$, to the nearest point in the feasible set, $\boldsymbol{x} \doteq \Pi(\boldsymbol{\theta})$, with projection function Π. This separation between the unconstrained and constrained space produces some ambiguity in the way optimization algorithms are adapted to handle constraints. Do we adjust the projected parameters or the unconstrained parameters with the gradient? And do we take the gradient with respect to the projected parameters or the unconstrained parameters?

Projected gradient descent (PGD) [Zinkevich, 2003] resolves these ambiguities by adjusting the projected parameters with the gradient of the projected parameters. For PGD, the unconstrained parameters are not saved, they are only produced temporarily before they can be projected into the feasible set, $\boldsymbol{x} \leftarrow \Pi(\boldsymbol{x} - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}))$. **Mirror descent (MD)** [Nemirovsky and Yudin, 1983; Beck and Teboulle, 2003], broadly, makes adjustments exclusively in the unconstrained space, and projects to the feasible set on-demand. However, like PGD, MD uses the gradient with respect to the projected parameters. E.g. A MD-based update is $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x})$, and projection is done on-demand, $\boldsymbol{x} \doteq \Pi(\boldsymbol{\theta})$.

Further difficulties are encountered when function approximation is involved, that is, when $\boldsymbol{x} \doteq \Pi(g(\boldsymbol{\theta}))$, for an arbitrary function g. Now PGD's approach of making adjustments in the transformed space where $\boldsymbol{x}$ resides is untenable because the function parameters, $\boldsymbol{\theta}$, might reside in a very different space. E.g. $\boldsymbol{x}$ may be a complete strategy while $\boldsymbol{\theta}$ may be a vector of neural network parameters with many fewer dimensions. But the gradient with respect to the $\boldsymbol{x}$ is also in the transformed space, so MD's update cannot be done exactly either.

A simple fix is to apply the chain rule to find the gradient with respect to $\boldsymbol{\theta}$. This is the approach taken by **policy gradient (PG)** methods [Williams, 1992b; Sutton *et al.*, 2000; Sutton and Barto, 2018] (the "all actions" versions rather than sample action versions). A consequence of this choice, however, is that PG updates in the tabular setting (when g is the identity function) generally do not reproduce MD updates.

E.g. **hedge**, **exponentially weighted experts**, or **entropic mirror descent** [Freund and Schapire, 1997; Beck and Teboulle, 2003], is a celebrated algorithm for approximately solving games. It is a simple no-regret algorithm that achieves the optimal regret dependence on the number of actions, $\log |\mathcal{A}|$, and it can be used in the CFR and CFR-BR framework instead of the more conventional regret-matching to solve sequential imperfect information games. It is also an instance of MD, which we show here.

Hedge accumulates values associated with each action (e.g. counterfactual values or regrets) and projects them into the probability simplex with the softmax projection to generate a policy. Formally, given a sequence of values $[\boldsymbol{\upsilon}^t]_{t=1}^T, \boldsymbol{\upsilon}^t \doteq [\upsilon_a^t]_{a \in \mathcal{A}}$ and temperature, $\tau > 0$, hedge plays

$$\pi^T_{\text{hedge}} = \frac{e^{\frac{1}{\tau} \sum_{t=1}^T \boldsymbol{\upsilon}^t}}{\sum_{a \in \mathcal{A}} e^{\frac{1}{\tau} \sum_{t=1}^T \upsilon_a^t}}.$$

We now show how to recover this policy creation rule with MD.

Given a vector of bounded action values, $\boldsymbol{\upsilon} \in \Upsilon^{|\mathcal{A}|} \subseteq \mathbb{R}^{|\mathcal{A}|}, \max \Upsilon - \min \Upsilon \leq \Delta$, the expected value of policy $\boldsymbol{\pi}_{\text{MD}}$ interpreted as a probability distribution is just the weighted sum of values, $\bar{\upsilon} \doteq \sum_{a \in \mathcal{A}} \pi_{\text{MD}}(a) \upsilon_a$.

The gradient of the expected value of $\boldsymbol{\pi}_{\text{MD}}$'s value is then just the vector of action values, $\boldsymbol{\upsilon}$. MD accumulates the gradients on each round,

$$\begin{aligned}
\boldsymbol{\theta}^t &\doteq \boldsymbol{\theta}^{t-1} + \alpha \nabla_{\boldsymbol{\pi}^{t-1}} \bar{\upsilon}^t \\
&= \boldsymbol{\theta}^{t-1} + \alpha \boldsymbol{\upsilon}^t \\
&= \boldsymbol{\theta}^0 + \alpha \sum_{i=1}^t \boldsymbol{\upsilon}^i,
\end{aligned}$$

where $\alpha > 0$ is a step-size parameter. If $\boldsymbol{\theta}^0$ is zero, then the current parameters, $\boldsymbol{\theta}^t$, are simply the step-size weighted sum of the action values.

If $\boldsymbol{\pi}^t_{\text{MD}}$ on each round, $t \geq 0$ is chosen to be $\boldsymbol{\pi}^t_{\text{MD}} = \Pi_{\text{sm}}(\boldsymbol{\theta}^t)$, then we can rewrite this policy in terms of the action values alone:

$$\boldsymbol{\pi}^t_{\text{MD}} = \frac{e^{\alpha \sum_{t=1}^T \boldsymbol{\upsilon}^t}}{\sum_{a \in \mathcal{A}} e^{\alpha \sum_{t=1}^T \upsilon_a^t}},$$

which one can recognize as hedge with $\tau \doteq 1/\alpha$. This shows how hedge fits into the ED framework. When counterfactual values are used for action values, ED with MD gradient-updates and softmax projection at every information state is identical to CFR-BR with hedge at every information state.

In comparison, PG, using the same projection and tabular parameterization, generates policies according to

$$\begin{aligned}
\boldsymbol{\theta}^t &\doteq \boldsymbol{\theta}^{t-1} + \alpha \langle \nabla_{\boldsymbol{\theta}^{t-1}} \boldsymbol{\pi}^{t-1}_{\text{PG}}, \boldsymbol{\upsilon}^t \rangle \\
\pi^t_{\text{PG}} &\doteq \Pi_{\text{sm}}(\boldsymbol{\theta}^t).
\end{aligned}$$

$\frac{\partial \pi(a')}{\partial \theta_a} = \pi(a') \left[\mathbb{1}_{a'=a} - \pi(a) \right]$ [Sutton and Barto, 2018, Section 2.8], so the update direction, $\langle \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}, \boldsymbol{\upsilon} \rangle$, is actually the

regret scaled by $\boldsymbol{\pi}$:

$$\begin{aligned}
\frac{\partial \boldsymbol{\pi}}{\partial \theta_a} \cdot \boldsymbol{\upsilon} &= \sum_{a' \in \mathcal{A}} \frac{\partial \pi(a')}{\partial \theta_a} \upsilon_{a'} \\
&= \sum_{a' \in \mathcal{A}} \pi(a') \left[\mathbb{1}_{a'=a} - \pi(a)\right] \upsilon_{a'} \\
&= \pi(a) \left[1 - \pi(a)\right] \upsilon_a - \sum_{a' \in \mathcal{A}, a' \neq a} \pi(a') \pi(a) \upsilon_{a'} \\
&= \pi(a) \left[\left[1 - \pi(a)\right] \upsilon_a - \sum_{a' \in \mathcal{A}, a' \neq a} \pi(a') \upsilon_{a'} \right] \\
&= \pi(a) \left[\upsilon_a - \pi(a) \upsilon_a - \sum_{a' \in \mathcal{A}, a' \neq a} \pi(a') \upsilon_{a'} \right] \\
&= \pi(a) \left[\upsilon_a - \sum_{a' \in \mathcal{A}} \pi(a') \upsilon_{a'} \right].
\end{aligned}$$

Knowing this, we can write the $\boldsymbol{\pi}^t_{\text{PG}}$ in concrete terms:

$$\begin{aligned}
\theta^t_a &\doteq \theta^{t-1}_a + \alpha \pi^{t-1}_{\text{PG}}(a) \left[\upsilon_a - \sum_{a' \in \mathcal{A}} \pi^{t-1}_{\text{PG}}(a') \upsilon_{a'} \right] \\
\boldsymbol{\pi}^t_{\text{PG}} &\doteq \Pi_{\text{sm}}(\boldsymbol{\theta}^t).
\end{aligned}$$

The fact that the PG parameters accumulate regret instead of action value is inconsequential because the difference between action values and regrets is a shift that is shared between each action, and the softmax projection is shift-invariant. But there is a substantive difference in that updates are scaled by the current policy.

B Global Minimum Conditions

In this section we will suppose that the policy $\pi_i(s,a) = \theta_{s,a}$ under the simplex constraints $\forall s, a,\ \theta_{s,a} - \psi^2_{s,a} = 0$ and $\forall s,\ \sum_a \theta_{s,a} = 1$ (where psi is a slack variable to enforce the inequality constrain $\forall s, a,\ \theta_{s,a} \geq 0$)

$$\begin{aligned}
\underset{x}{\text{maximize}} \quad & \min_{\pi_{-i}} v_{i,(\pi_{\theta_i}, \pi_{-i})} \\
\text{subject to} \quad & \sum_a \theta_{s,a} = 1,\ \forall s \\
& \theta_{s,a} - \psi^2_{s,a} = 0,\ \forall s, a
\end{aligned}$$

The Lagrangian is:

$$\begin{aligned}
L(\theta, \psi, \lambda) = & \min_{\pi_{-i}} v_{i,(\pi_{\theta_i}, \pi_{-i})} - \sum_{s \in S^i} \lambda_s (\sum_a \theta_{s,a} - 1) \\
& - \sum_{s \in S^i, a} \lambda_{s,a} (\theta_{s,a} - \psi^2_{s,a})
\end{aligned}$$

Knowing that for all br_{-i} :

$$\frac{\partial v_{i,(\pi_{\theta_i}, br_{-i})}}{\theta_{s,a}} = \eta_{(\pi_{\theta_i}, br_{-i})}(s) q_{i,(\pi_{\theta_i}, br_{-i})}(s,a)$$

The gradient of the Lagrangian with respect to θ, ψ, λ is:

$$\begin{aligned}
&\left(\eta_{(\pi_{\theta_i}, br_{-i})}(s) q_{i,(\pi_{\theta_i}, br_{-i})}(s,a) - (\lambda_s + \lambda_{s,a}) \right)_{s,a} \\
&\qquad \in \partial_{\theta_i} L(\theta, \psi, \lambda) \qquad (9) \\
&\frac{\partial L(\theta, \psi, \lambda)}{\partial \lambda_s} = -(\sum_a \theta_{s,a} - 1) \qquad (10) \\
&\frac{\partial L(\theta, \psi, \lambda)}{\partial \lambda_{s,a}} = -(\theta_{s,a} - \psi^2_{s,a}) \qquad (11) \\
&\frac{\partial L(\theta, \psi, \lambda)}{\partial \psi_{s,a}} = 2 \psi_{s,a} \lambda_{s,a} \qquad (12) \\
& \qquad (13)
\end{aligned}$$

Suppose that there exists a best response br_{-i} such that $\frac{\partial L(\theta,\psi,\lambda)}{\partial \theta_{s,a}} = \frac{\partial L(\theta,\psi,\lambda)}{\partial \lambda_s} = \frac{\partial L(\theta,\psi,\lambda)}{\partial \lambda_{s,a}} = \frac{\partial L(\theta,\psi,\lambda)}{\partial \psi_{s,a}} = 0$ (*i.e.* if 0 is in the set of generalized gradients). Two cases can appear:

-If $\theta_{s,a} > 0$ then $\lambda_{s,a} = 0$ and then:

$$\forall s, a | \theta_{s,a} > 0\ \eta_{(\pi_{\theta_i}, br_{-i})}(s) q_{i,(\pi_{\theta_i}, br_{-i})}(s,a) = \lambda_s$$

-If $\theta_{s,a} = 0$ then $\psi_{s,a} = 0$ and then:

$$\eta_{(\pi_{\theta_i}, br_{-i})}(s) q_{i,(\pi_{\theta_i}, br_{-i})}(s,a) = \lambda_s + \lambda_{s,a}$$

Two cases (one stable and one unstable):

- $\lambda_{s,a} \leq 0$ then we have a stable fixed point,
- $\lambda_{s,a} > 0$ is not stable as then we could increase the value by switching to that action.

We conclude by noticing that π_{θ_i} is greedy with respect to the value of the joint policy $(\pi_{\theta_i}, br_{-i})$, thus π_{θ_i} is a best response to br_{-i}. Since both policies are best responses to each other, $(\pi_{\theta_i}, br_{-i})$ is a Nash equilibrium. π_{θ_i} is also therefore unexploitable.