Regret Minimization in Behaviorally-Constrained Zero-Sum Games[*]

Gabriele Farina, Christian Kroer, and Tuomas Sandholm

Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{gfarina,ckroer,sandholm}@cs.cmu.edu

Abstract. No-regret learning has emerged as a powerful tool for solving extensive-form games. This was facilitated by the counterfactual-regret minimization (CFR) framework, which relies on the instantiation of regret minimizers for simplexes at each information set of the game. We use an instantiation of the CFR framework to develop algorithms for solving behaviorally-constrained (and, as a special case, *perturbed* in the Selten sense) extensive-form games, which allows us to compute approximate Nash equilibrium refinements. Nash equilibrium refinements are motivated by a major deficiency in Nash equilibrium: it provides virtually no guarantees on how it will play in parts of the game tree that are reached with zero probability. Refinements can mend this issue, but have not been adopted in practice, mostly due to a lack of scalable algorithms. We show that, compared to standard algorithms, our method finds solutions that have substantially better refinement properties, while enjoying a convergence rate that is comparable to that of state-of-the-art algorithms for Nash equilibrium computation both in theory and practice.

1 Introduction

No-regret learning algorithms have become a powerful tool for solving large-scale zero-sum extensive-form games (EFGs) [3, 4]. This has largely been facilitated by the *counterfactual-regret minimization* (CFR) algorithm [36] and its newer variants [23, 29, 3, 6, 5, 7]. This framework works by defining a notion of regret local to an information set, and instantiating a standard regret minimizer at each information set in order to minimize local regret. Zinkevich et al. [36] prove that this scheme of local regret minimization leads to a Nash equilibrium in two-player zero-sum extensive-form games of perfect recall. The framework works with any regret-minimizing algorithm, but in practice variants of the *regret matching* algorithm have been dominant [13, 3, 6, 4]. We investigate the extension of regret-matching$^+$ (RM$^+$) [32], an even faster regret-matching algorithm, to more general regret-minimization problems over (finitely-generated) convex polytopes. We use these results to instantiate RM$^+$ for linearly constrained simplexes, which in turn allows us to model and solve behaviorally-constrained EFGs (which are EFGs with additional linear constraints on the

---

[*] A version of this manuscript appeared at ICML 2017.

simplexes at each information set). An important special case of this framework is behaviorally-perturbed EFGs, which can be used to compute Nash equilibrium refinements.[1]

Nash equilibrium refinements are motivated by major deficiencies in the Nash equilibrium solution concept: Nash equilibria provide no guarantees on performance in information sets that are reached with probability zero in equilibrium, beyond not giving up more utility than the value of the game. Thus, if an opponent makes a mistake, Nash equilibrium is not guaranteed to capitalize on it, but may instead give back up to all of that utility [24]. This is especially relevant when Nash equilibria are used as a solution concept for playing against imperfect opponents. Equilibrium refinements ameliorate this issue by introducing further constraints on behavior in information sets that are reached with probability zero. We will be interested in equilibrium concepts that achieve this through the notion of *perturbations* or *trembling hands* [30]. At each decision-point, a player is assumed to tremble with some small probability, and a Nash equilibrium is then computed in this perturbed game. A refinement is then a limit point of the sequence of Nash equilibria achieved as the probability of trembles is taken to zero. In *quasi-perfect equilibria*, players take into account only the trembles of their opponents [34], whereas in an *extensive-form perfect equilibrium* (EFPE), players take into account mistakes made both by themselves and opponents [30].

We compare our algorithm for perturbed EFGs to state-of-the-art large-scale zero-sum EFG-solving algorithms: the standard CFR$^+$ algorithm [32] and the excessive gap technique (EGT) [26] instantiated with a state-of-the art *smoothing function* [27, 14, 21, 20]. We find that our perturbed variant of CFR$^+$ converges (in the perturbed game) at the same rate as those algorithms converge while ours leads to orders of magnitude more refined strategies. Our algorithm also converges at the same rate in the unperturbed game, almost until the point where the imposed behavioral constraints necessarily prevent further convergence.

2   Related work

No-regret algorithms have a long history in EFG solving. Gordon [11] developed the *Lagrangian Hedging* algorithm, which can be used to find a Nash equilibrium in EFGs. However, it suffers from a drawback: it requires projection onto the strategy space at each iteration. Zinkevich et al. [36] developed CFR, which avoids projection while maintaining the same convergence rate. It has since been

---

[1] The idea of certain kinds of behavioral perturbations to CFR has been suggested by Neller and Lanctot [25]. They suggest that at every information set, with small probability ϵ, a player will make a random move. However, they provide no results of what the refinement consequences are (i.e., what kind of refinement this would lead to), and it is unclear whether the proposed method actually leads to refinements. In contrast, we establish a connection to (approximate) EFPEs in this paper. Furthermore, Neller and Lanctot cite Miltersen and Sørensen [24] which is about quasi-perfect equilibrium (where a player assumes she will not make errors in the future), while the ϵ modeling of Neller and Lanctot makes ϵ errors at future information sets as well.

extended in a number of ways. Lanctot et al. [23] showed how to incorporate sampling in CFR. Brown and Sandholm [6] showed how to achieve greater pruning in CFR, thereby reducing the iteration costs. CFR^+ is a state-of-the-art variant of CFR [32], which has vastly superior practical performance compared to standard CFR, though it is not known to be stronger from a theoretical perspective. Gordon et al. [12] shows how no-regret algorithms can also be utilized for computing extensive-form correlated equilibria in EFGs.

Polynomial-time algorithms have been proposed for computing certain equilibrium refinements in two-player zero-sum perfect-recall EFGs. Miltersen and Sørensen [24] develop a linear program (LP) for computing quasi-perfect equilibria by choosing a sufficiently small perturbation to realization plans. Farina and Gatti [9] develop a similar approach for EFPE computation, but rely on perturbations to behavioral strategies of players. These approaches rely on solving modified variants of the sequence-form LP for computing Nash equilibrium [35] in EFGs. These algorithms are of theoretical interest only and do not work in practice. They require rational numbers of precision $n \log n$ bits, where n is the number of sequences in the game. Another issue is that LP algorithms do not scale to large EFGs even when just finding Nash equilibria, and in practice CFR-based or EGT-based approaches are used to achieve scalability. Kroer et al. [18] recently showed how smoothing functions for first-order methods such as EGT can be extended to games with perturbations.

Johanson et al. [16] consider robust strategies that arise from assuming that the opponent will randomize between playing a Nash equilibrium and a strategy within some model of opponent behavior. Johanson and Bowling [15] consider a similar model-biased Nash equilibrium approach on games where an independent model is used at each information set. Ganzfried and Sandholm [10] develop an opponent modeling approach that adds opponent-modeled constraints across information sets. Our approach provides a principled framework for solving model-biased games that use general constraints on per-information set behavioral strategies. Constraints across information sets currently require the much less scalable LP approach.

3   Preliminaries

We briefly introduce several of the basic concepts we use in the rest of the paper. We denote by $\mathbb{R}_+$ and $\mathbb{R}_-$ the set of non-negative and non-positive reals, respectively.

3.1   Normal-Form Games

Definition 1. *A* two-player zero-sum normal-form game *(for the rest of the paper, simply* normal-form game *or* NFG*) is a tuple* (A_1, A_2, u) *where* A_1 *represents the finite set of actions that player 1 can play,* A_2 *represents the finite set of actions that player 2 can play, and* $u : A_1 \times A_2 \to \mathbb{R}$ *is the payoff function for player 1, mapping the pair of actions* (a_1, a_2) *of the players into the payoff for player 1. The corresponding payoff for player 2 is given by* $-u(a_1, a_2)$.

Usually, the payoff function u is given as a matrix U, called the *payoff matrix* of the game. The rows of U represent the actions $\{a_{1,1}, \ldots, a_{1,n}\} = A_1$ of player 1, while the columns of U represent the actions $\{a_{2,1}, \ldots, a_{2,m}\} = A_2$ of player 2. At the intersection of the i-th row and the j-th column is the payoff for the action pair $(a_{1,i}, a_{2,j})$.

Definition 2. *A mixed strategy π for player $i \in \{1,2\}$ is a probability mass function over the set A_i.*

When players play according to mixed strategies π_1 and π_2 respectively, the expected payoff is given by

$$\mathbb{E}_{\pi_1, \pi_2}(u) = \sum_{a_1 \in A_1} \sum_{a_2 \in A_2} \pi_1(a_1)\pi_2(a_2)u(a_1, a_2). \tag{1}$$

3.2 Generalized Normal-Form Games

Telgarsky [33] and Abernethy et al. [1] propose a generalization of the concept of normal-form games, which conveniently allows us to remove all expectation operators, making the notation lighter and more legible. In this generalization, players select deterministic strategies from a convex compact set. For a normal-form game, this set is the space of all mixed strategies.

Definition 3. *A two-player zero-sum generalized normal-form game $\Gamma = (\mathcal{X}, \mathcal{Y}, u)$ is a tuple defined by a pair of convex and compact action spaces $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{Y} \subseteq \mathbb{R}^m$, one for each player, as well as a* biaffine *utility function $u : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The utility function $u(x,y)$ maps the pair of actions $(x,y) \in \mathcal{X} \times \mathcal{Y}$ of the players into the payoff for player 1, while the corresponding payoff for player 2 is given by $-u(x,y)$.*

Observation 1. *Any normal-form game can be mapped to an instance of a generalized normal-form game. Given $\Gamma = (A_1, A_2, u)$, where $|A_1| = n$ and $|A_2| = m$, the set of all mixed strategies for player 1 forms the n-dimensional simplex $\mathcal{X} = \Delta_n$, while the set of all mixed strategies for player 2 forms the m-dimensional simplex $\mathcal{Y} = \Delta_m$. Let U be the payoff matrix associated with Γ. Using Equation 1, we conclude that Γ is equivalent to the generalized two-player zero-sum normal-form game $\Gamma^* = (\mathcal{X}, \mathcal{Y}, u^*)$, where $u^*(x,y) = x^\top U y$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$.*

3.3 Extensive-Form Games

Definition 4. *A two-player zero-sum extensive-form game with imperfect information and perfect recall Γ is a tuple $(H, Z, A, P, f_c, \mathcal{I}_1, \mathcal{I}_2, u)$ composed of:*
- *H: a finite set of possible sequences (or histories) of actions, such that the empty sequence $\varnothing \in H$, and every prefix z of h in H is also in H.*
- *$Z \subseteq H$: the set of terminal histories, i.e. those sequences that are not a proper prefix of any sequence.*

- *A: a function mapping $h \in H \setminus Z$ to the set of available actions at non-terminal history h.*
- *P: the player function, mapping each non-terminal history $h \in H \setminus Z$ to $\{1, 2, c\}$, representing the player who takes action after h. If $P(h) = c$, the player is chance.*
- *f_c: a function assigning to each $h \in H \setminus Z$ such that $P(h) = c$ a probability mass function over $A(h)$.*
- *$\mathcal{I}_i$, for $i \in \{1, 2\}$: partition of $\{h \in H : P(h) = i\}$ with the property that $A(h) = A(h')$ for each h, h' in the same set of the partition. For notational convenience, we will write $A(I)$ to mean $A(h)$ for any of the $h \in I$, where $I \in \mathcal{I}_i$. $\mathcal{I}_i$ is the information partition of player i, while the sets in $\mathcal{I}_i$ are called the information sets of player i.*
- *u: utility function mapping $z \in Z$ to the utility (a real number) gained by player 1 when the history is reached. The corresponding utility for player 2 is given by $-u(z)$.*

We further assume that all players can recall their previous actions and the corresponding information sets.

In the rest of the paper, we will use the more relaxed term *extensive-form game*, or EFG, to mean a two-player zero-sum extensive-form game with imperfect information and perfect recall.

Observation 2. *Extensive-form games can be represented as generalized NFGs, for example, via the normal form representation or sequence form representation [28, 17, 35].*

3.4 Regret and Regret Minimization

Suppose there exists an iterative algorithm which, at each step $t = 1, \ldots, T$, computes a strategy $x_t \in \mathcal{X}$ for player 1, and plays a (generalized) normal-form game $(\mathcal{X}, \mathcal{Y}, u)$ against player 2 using such strategy. Let y_t be the strategy used by player 2 at step t. The average external regret of player 1 up to step T against action $\hat{x} \in \mathcal{X}$ is

$$\bar{R}_1^T(\hat{x}) = \frac{1}{T} \sum_{t=1}^{T} u(\hat{x}, y_t) - u(x_t, y_t).$$

The case for player 2 is symmetrical. A *regret-minimizing scheme* is a function that assigns, for each sequence of past actions $x_1, y_1, \ldots, x_{t-1}, y_{t-1}$, an action x_t such that $\limsup_{T \to \infty} \max_{\hat{x} \in \mathcal{X}} \bar{R}_1^T(\hat{x}) \leq 0$.

Regret-matching (RM) [13] is a regret-minimizing scheme for normal-form games, based on Blackwell's approachability theorem [2]. The following theorem, a proof of which is given by Cesa-Bianchi and Lugosi [8], characterizes the convergence rate of RM.

Theorem 1. *Given a normal-form game (A_1, A_2, u), the maximum average external regret for player 1 at iteration T, when player 1 plays according to the*

regret-matching algorithm, is

$$\max_{\hat{x}} \bar{R}_1^T(\hat{x}) \leq \gamma \frac{\sqrt{|A_1|}}{\sqrt{T}},$$

where $\gamma \doteq \max_{x,y} u(x,y) - \min_{x,y} u(x,y)$.

Regret-matching$^+$ is an extension of RM, and converges significantly faster in practice. Tammelin et al. (2015; Lemma 2) proved that the convergence rate of RM$^+$ is the same as that of RM above.

3.5   Nash Equilibria and Refinements

We now review the needed solution concepts from game theory. We mostly focus on generalized normal-form games, allowing a compact presentation of concepts pertaining both normal-form games and extensive-form games.

Definition 5 (Approximate best response). *Given a generalized normal-form game* $(\mathcal{X}, \mathcal{Y}, u)$ *and a strategy* $y \in \mathcal{Y}$*, we say that* $x \in \mathcal{X}$ *is an* ϵ*-best response to* y *for player 1 if* $u(x,y) + \epsilon \geq u(\hat{x}, y)$ *for all* $\hat{x} \in \mathcal{X}$*. Symmetrically, given* $x \in \mathcal{X}$*, we say that* $y \in \mathcal{Y}$ *is an* ϵ*-best response to* x *for player 2 if* $-u(x,y) + \epsilon \geq -u(x, \hat{y})$ *for all* $\hat{y} \in \mathcal{Y}$*.*

Definition 6 (Approximate Nash equilibrium). *Given a generalized normal-form game* $(\mathcal{X}, \mathcal{Y}, u)$*, the strategy pair* $(x,y) \in \mathcal{X} \times \mathcal{Y}$ *is a* ϵ*-Nash equilibrium for the game if* x *is an* ϵ*-best response to* y *for player 1, and* y *is an* ϵ*-best response to* x *for player 2.*

Definition 7 (Nash equilibrium). *Given a generalized normal-form game* $(\mathcal{X}, \mathcal{Y}, u)$*, a* Nash equilibrium *for the game is a* 0*-Nash equilibrium.*

There exists a well-known relationship between regret and approximate Nash equilibria (Definition 6), as summarized in the next theorem.

Theorem 2. *In a zero-sum game, if the average external regrets of the players up to step* T *are such that*

$$\bar{R}_1^T(\hat{x}) \leq \epsilon_1, \qquad \bar{R}_2^T(\hat{y}) \leq \epsilon_2$$

for all actions $\hat{x} \in \mathcal{X}, \hat{y} \in \mathcal{Y}$*, then the strategy pair*

$$(\bar{x}_T, \bar{y}_T) \doteq \left(\frac{1}{T} \sum_{i=1}^{T} x_i, \frac{1}{T} \sum_{i=1}^{T} y_i \right) \in \mathcal{X} \times \mathcal{Y}$$

is an $(\epsilon_1 + \epsilon_2)$*-Nash equilibrium.*

Theorem 2 basically says that if there exists an iterative algorithm able to progressively propose strategies so that the maximum average external regret go to zero, then recovering a Nash equilibrium is straightforward, and just a matter of averaging the individual strategies proposed.

We now turn to the class of *perturbed games* [30]. Intuitively, a perturbation restricts the set of playable strategies by enforcing a lower bound on the probability of playing each action. We recall the definition and some of the properties of game perturbations, starting from the normal-form case. We focus on player 1, but remark that the same definitions hold symmetrically for player 2 as well.

Definition 8. *Let $\Gamma = (A_1, A_2, u)$ be an NFG and let $\Gamma^* = (\Delta_{|A_1|}, \Delta_{|A_2|}, u^*)$ be its generalized NFG representation (see Observation 1). A perturbation is a function $p : A_1 \cup A_2 \to \mathbb{R}_+$ such that $\sum_{a \in A_1} p(a) < 1$ and $\sum_{a \in A_2} p(a) < 1$. The corresponding perturbed NFG Γ_p is the generalized NFG where each action a must be played with probability at least $p(a)$. Formally, $\Gamma_p = (\tilde{\mathcal{X}}_p, \tilde{\mathcal{Y}}_p, u^*)$ where $\tilde{\mathcal{X}}_p = \{x \in \Delta^{|A_1|} : x_a \geq p(a) \forall a \in A_1\}$. $\tilde{\mathcal{Y}}_p$ is defined analogously.*

In the case of extensive-form games, a perturbation for player 1 assigns a lower-bound on each action playable by the player. More precisely:

Definition 9. *Let $\Gamma = (H, Z, A, P, f_c, \mathcal{I}_1, \mathcal{I}_2, u)$ be an extensive-form game. A perturbation is a function p mapping each pair (I, a) where $I \in \mathcal{I}_1 \cup \mathcal{I}_2$ and $a \in A(I)$ to a non-negative real, such that*

$$\sum_{a \in A(I)} p(I, a) < 1 \qquad \forall\, I \in \mathcal{I}_1 \cup \mathcal{I}_2.$$

The corresponding perturbed EFG Γ_p *is the analogous game where each action a at each information set I has to be played with probability at least $p(I, a)$.*

Perturbations play an important role in equilibrium refinement, as they form the basis for the concept of *equilibrium perfection* [30]. In this paper we only focus on the case of *extensive-form perfect equilibria (EFPEs)*.

Definition 10. *A strategy pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is an EFPE of Γ if it is a limit point of a sequence $\{(x_p, y_p)\}_{p \to 0}$ where (x_p, y_p) is a Nash equilibrium of the perturbed game Γ_p.*

Intuitively, an EFPE is an equilibrium refinement that takes into account an imperfect ability to deterministically commit to a single action.

4 Generalized Normal-Form Games over Finitely-Generated Convex Polytopes

In this section, we show how to adapt a regret-minimization algorithm to handle generalized normal-form games played on finitely-generated convex polytopes. The key insight is that when the action space is a finitely-generated convex polytope, the generalized game can be cast back as a normal-form game, i.e. a generalized normal-form game played over simplexes, and "solved" by a regret-minimization algorithm; subsequently, the solution for the normal-form game gets mapped back into the polytope. This is achieved by constructing new simplex action spaces for the players, where each point in a simplex denotes a convex combination of weights on the vertices of that players' finitely-generated convex polytopal action space.

Theorem 3. *Let $\Gamma = (\mathcal{X}, \mathcal{Y}, u)$ be a generalized normal-form game played on the finitely-generated convex polytopes $\mathcal{X}$ and $\mathcal{Y}$. There exists a regret-minimizing scheme for player 1 in Γ.*

Proof. Let $\{b_1, \ldots, b_n\}$ be a convex basis for $\mathcal{X}$, and $\{c_1, \ldots, c_m\}$ be a convex basis for $\mathcal{Y}$; also, let $B = (b_1 \mid \cdots \mid b_n)$ and $C = (c_1 \mid \cdots \mid c_m)$ be the basis matrices for $\mathcal{X}$ and $\mathcal{Y}$, respectively. We construct a generalized normal-form game $\Gamma^* = (\Delta_n, \Delta_m, u^*)$, where

$$u^*(x, y) \doteq u(Bx, Cy)$$

for all $x \in \Delta_n, y \in \Delta_m$. Of course $Bx \in \mathcal{X}, Cy \in \mathcal{Y}$ for all x and y, so the definition is valid. Let f^* be any of regret-minimizing schemes for normal-form games (e.g., RM or RM$^+$). We construct a regret-minimizing scheme f for Γ such that, at each iteration $t = 1, 2, \ldots,$

$$f(x_1, y_1, \ldots, x_{t-1}, y_{t-1}) = B f^*(x_1^*, y_1^*, \ldots, x_{t-1}^*, y_{t-1}^*),$$

where x^*, y^* denotes the coordinates of x, y with respect to the basis of $\mathcal{X}, \mathcal{Y}$, respectively; note that this definition is well-defined since the coordinates are guaranteed to belong to $\Delta_n, \Delta_m{}^2$. The regret induced by this scheme is

$$\bar{R}_1^T(\hat{x}) = \frac{1}{T} \sum_{t=1}^{T} u(\hat{x}, y_t) - u(x_t, y_t)$$

$$= \frac{1}{T} \sum_{t=1}^{T} u^*(\hat{x}^*, y_t^*) - u^*(x_t^*, y_t^*).$$

Notice that the last expression is exactly the average regret for player 1 up to iteration T against action $\hat{x}^*$ in Γ^*. Since f^* is a regret-minimizing scheme, the average regret against any action converges to zero, meaning that $\limsup_{T \to \infty} \bar{R}_1^T(\hat{x}) \leq 0$ for each $\hat{x}$, i.e. $\limsup_{T \to \infty} \max_{\hat{x} \in \mathcal{X}} \bar{R}_1^T(\hat{x}) \leq 0$. This proves that f is a regret-minimizing scheme for Γ, concluding the proof.

Another way to think about the construction above is that at each iteration, we compute the regret for not playing each of the "strategies" forming the vertices of the polytope, and updating the next strategy by taking a convex combination of the vertices, in a way proportional to the regret against them.

Algorithm 1 represents an instantiation of the construction given in the proof of Theorem 3, where the regret-minimizing scheme for the normal-form game was chosen to be RM$^+$. A careful analysis of the construction also reveals that the convergence bound for RM$^+$ carries over, as expressed by Theorem 4. At time t, RM$^+$ projects the cumulative regret r_{t-1} onto the non-negative orthant $\mathbb{R}_+^n$; the projection is equal to the vector $[r_{t-1}]^+$, where $[a]_i^+ \doteq \max\{0, a_i\}$.

---

^{2} Passage to coordinates might not be unique. In this case, any coordinate vector will do, as long as the choice is deterministic.

Algorithm 1 RM$^+$ algorithm for generalized normal-form games played over finitely-generated convex polytopes.

1: **procedure** REGRET-MATCHING$^+$(Γ)
   ▷ $\Gamma = (\mathcal{X}, \mathcal{Y}, u)$, and B is a fixed convex basis for $\mathcal{X}$
   ▷ **note**: this reflects the point of view of player 1
2:     $r_0 \leftarrow (0, \ldots, 0)^\top \in \mathbb{R}^n$
3:     $\bar{x} \leftarrow (0, \ldots, 0)^\top \in \mathbb{R}^n$
4:     **for** $t = 1, 2, 3, \ldots$ **do**
5:         **if** $r_{t-1} \in \mathbb{R}^n_-$ **then**
6:             $x_t \leftarrow$ any action $\in \tilde{\mathcal{X}}_p$
7:         **else**
8:             $\Lambda_{t-1} \leftarrow \sum_{i=1}^{n} [r_{t-1}]^+_i$
9:             $x_t \leftarrow B \dfrac{[r_{t-1}]^+}{\Lambda_{t-1}}$
10:        play action x_t
11:        observe $y_t \in \mathcal{Y}$ played by opponent
12:        $r_t \leftarrow \left[r_{t-1} + \begin{pmatrix} u(b_1, y_t) - u(x_t, y_t) \\ \vdots \\ u(b_n, y_t) - u(x_t, y_t) \end{pmatrix} \right]^+$
13:        $\bar{x} \leftarrow \dfrac{t-1}{t}\bar{x} + \dfrac{1}{t}x_t$
        ▷ $\bar{x}$ contains the average strategy for player 1

Theorem 4. *Given a generalized normal-form game $(\mathcal{X}, \mathcal{Y}, u)$ with finitely generated $\mathcal{X}$ and $\mathcal{Y}$, the maximum average external regret for player 1 at iteration T, when player 1 plays according to Algorithm 1, is bounded by*

$$\max_{\hat{x} \in \mathcal{X}} \bar{R}_1^T(\hat{x}) \leq \gamma \frac{\sqrt{|\mathcal{X}|}}{\sqrt{T}}$$

where $\gamma \doteq \max_{x,y} u(x,y) - \min_{x,y} u(x,y)$, and $|\mathcal{X}|$ denotes the number of vertices of $\mathcal{X}$.

5 Behavioral Constraints and Perturbations

Behavioral constraint are linear constraints on the simplexes at each information set. In order to obtain a regret minimizer for a behaviorally-constrained EFG, we could try to cast the game as a generalized NFG by means of the normal-form or sequence form representation (see Observation 2). However, the number of vertices of this representation is exponential, and therefore it does not work well with Theorem 4. Counterfactual Regret (CFR, Zinkevich et al. 2007) solves this problem, by defining a regret-minimizing scheme that runs in polynomial time in the size of the game. Intuitively, CFR minimizes a variant of instantaneous regret, called *immediate counterfactual regret*, at each information set separately, and later combines the strategies computed at each information set. It requires

simplex regret minimizers for each information set. If we have a finite number of them, each information set can be modeled as a finitely-generated convex polytope. We can then use Theorem 4 to get regret minimizers for each information set. Perturbations can be handled as a special case.

Theorem 5 below shows that CFR$^+$ instantiated with such regret minimizers for each behaviorally-constrained information set converges to an equilibrium of the constrained EFG. For this approach to be practical, we need the set of vertices for each information set to be of manageable size, as reflected in the dependence on $\max_{I \in \mathcal{I}} \sqrt{|\mathcal{Q}^I|}$ in Theorem 5, where $|\mathcal{Q}^I|$ is the number of vertices in the behaviorally-constrained simplex at information set I.

Theorem 5. *Let $(H, Z, A, P, f_c, \mathcal{I}_1, \mathcal{I}_2, u)$ be an extensive-form game; let $\mathcal{Q}^I \subseteq \Delta_{|A(I)|}$ represent the behaviorally-constrained strategy space at information set I, for all $I \in \mathcal{I}_1 \cup \mathcal{I}_2$. The maximum average external regret for player 1 in the constrained game at iteration T, when player 1 plays according to CFR$^+$, is bounded by*

$$\bar{R}_1^T \leq \gamma |\mathcal{I}_1| \frac{\sqrt{\max_{I \in \mathcal{I}_1} |\mathcal{Q}^I|}}{\sqrt{T}},$$

where $\gamma \doteq \max_{x,y} u(x, y) - \min_{x,y} u(x, y)$.

6   Perturbed Normal-Form Games

Section 4 established that, in general, the problem of finding an approximate Nash equilibrium for player 1 in the generalized normal-form game $\Gamma = (\mathcal{X}, \mathcal{Y}, u)$, where $\mathcal{X}$ is a convex polytope generated by n vectors, can be solved via regret-matching.

We now specialize this result for the specific case of perturbed normal-form games. The following holds:

Proposition 1. *Let $\Gamma = (A_1, A_2, u)$ be a normal-form game, where $A_1 = \{a_1, \ldots, a_n\}$, and let p be a perturbation for player 1. Let $\Gamma_p = (\tilde{\mathcal{X}}_p, \tilde{\mathcal{Y}}_p, u^*)$ be the generalized normal-form game corresponding to the perturbation (Definition 8). Then the perturbed action space $\tilde{\mathcal{X}}_p$ is a finitely generated convex polytope of dimension n, a basis of which is given by the columns of the following invertible matrix:*

$$B_p \doteq \begin{pmatrix} \tau_p + p(a_1) & p(a_1) & \cdots & p(a_1) \\ p(a_2) & \tau_p + p(a_2) & \cdots & p(a_2) \\ \vdots & \vdots & \ddots & \vdots \\ p(a_n) & p(a_n) & \cdots & \tau_p + p(a_n) \end{pmatrix}$$

where $\tau_p \doteq 1 - p(a_1) - \cdots - p(a_n)$.

This means that Algorithm 1 is applicable and provides a regret-minimizing scheme. We remark that when computing the instantaneous regrets (Algorithm 1, Line 12), it is important to remember these values have to be computed against the basis $\{b_1, \ldots, b_n\}$ of $\tilde{\mathcal{X}}_p$. However, computing the (expected) utility of the

game when player 1 plays according to a mixed strategy is usually more expensive than the same task when player 1 plays a deterministic action. For this reason, we express the instantaneous regret calculation against $\{b_1, \ldots, b_n\}$ in terms of instantaneous regrets against the pure actions $\{e_1, \ldots, e_n\}$ in the unperturbed game. In particular, by using the fact that the utility function is biaffine, we can write, for each $i \in \{1, \ldots, n\}$,

$$u(b_i, y_t) = u\left(\tau_p e_i + \sum_{j=1}^n p(a_j)e_j, y_t\right)$$

$$= \tau_p u(e_i, y_t) + \sum_{j=1}^n p(a_j)u(e_j, y_t),$$

so that, by introducing $\phi_{t,i} \doteq u(e_i, u_t) - u(x_t, y_t)$ and the corresponding vector $\phi_t \doteq (\phi_{t,1}, \ldots, \phi_{t,n})^\top$, we have

$$r_t = r_{t-1} + \tau_p \phi_t + \mathbf{1}\begin{pmatrix} p(a_1) \\ \vdots \\ p(a_n) \end{pmatrix}^\top \phi_t.$$

This allows us to compute the regret update in terms of the instantaneous regret against $\{e_1, \ldots, e_n\}$ in the unperturbed game, without introducing any overhead from an asymptotic point of view.

The maximum average external regret for player 1 at iteration T is given by Theorem 4; in this case $|\tilde{\mathcal{X}}_p| = |A_1|$.

7 Perturbed Extensive-Form Games

As discussed in Section 5, it is possible to use CFR in conjunction with any regret-minimizing scheme for generalized NFGs, in order to define a regret-minimizing scheme able to support any behaviorally-perturbed EFG (thus including the restricted case of perturbed EFGs). In Algorithm 2, we propose an implementation of CFR$^+$, i.e. CFR instantiated with the RM$^+$ algorithm able to handle perturbed EFGs. Algorithm 2 assumes that we are given a perturbation p of the extensive-form game, $\mathcal{X}_p^I$ denotes the perturbed simplex for information set I, and $\tau_p(I)$ is as in Proposition 1.

The following theorem characterizes the convergence guarantee of the proposed algorithm.

Theorem 6. *Let $(H, Z, A, P, f_c, \mathcal{I}_1, \mathcal{I}_2, u)$ be an extensive-form game; let p be the perturbation applied to the game. The maximum average external regret for player 1 in the perturbed game at iteration T, when player 1 plays according to Algorithm 2, is bounded by*

$$\bar{R}_1^T \leq \gamma |\mathcal{I}_1| \frac{\sqrt{\max_{I \in \mathcal{I}_1} |A(I)|}}{\sqrt{T}},$$

Algorithm 2 Regret minimization algorithm for perturbed extensive-form games.

1: **procedure** REGRET-MATCH$^+$-INFOSET(I, t)
    ▷ we assume $A(I) = \{a_1, \ldots, a_n\}$.
2:      **if** $r_{t-1}^I \in \mathbb{R}_-^n$ **then**
3:          $x_t^I \leftarrow$ any action $\in \tilde{\mathcal{X}}_p^I$
4:      **else**
5:          $\Lambda^I \leftarrow \sum_{i=1}^n \left[r_{t-1}^I \right]_i^+$
6:          $x_t^I \leftarrow \begin{pmatrix} p(I, a_1) \\ \vdots \\ p(I, a_n) \end{pmatrix} + \tau_p(I) \dfrac{\left[r_{t-1}^I \right]^+}{\Lambda^I}$

1: **procedure** TRAVERSE(h, i, t, π_1, π_2)
    ▷ assume h belongs to information set I
2:      **if** $h \in Z$ **then**
3:          **return** $u(h)$
4:      **if** $P(h) = c$ **then**   ▷ chance node
5:          sample $a \sim f_c(h)$
6:          **return** TRAVERSE(ha, t, π_1, π_2)
7:      **else if** $P(h) = 2$ **then**
8:          $v_t^I \leftarrow \begin{pmatrix} \text{TRAVERSE}(ha_1, t, \pi_1, y_{t,1}^I \pi_2) \\ \vdots \\ \text{TRAVERSE}(ha_n, t, \pi_1, y_{t,n}^I \pi_2) \end{pmatrix}$
9:      **else**   ▷ player 1's turn
10:         REGRET-MATCH-INFOSET(I, t)
11:         $v_t^I \leftarrow \begin{pmatrix} \text{TRAVERSE}(ha_1, t, x_{t,1}^I \pi_1, \pi_2) \\ \vdots \\ \text{TRAVERSE}(ha_n, t, x_{t,n}^I \pi_1, \pi_2) \end{pmatrix}$
12:         $\bar{v} \leftarrow (x_t^I)^\top v_t^I$
13:         $\phi_t^I \leftarrow \pi_2 (v_t^I - \mathbf{1}\,\bar{v})$
14:         $r_t^I \leftarrow \left[r_{t-1}^I + \tau_p(I)\phi_t^I + \mathbf{1} \begin{pmatrix} p(I, a_1) \\ \vdots \\ p(I, a_n) \end{pmatrix}^\top \phi_t^I \right]^+$
15:         $\bar{x}^I \leftarrow \bar{x}^I + \pi_1 x_t^I$
16:         **return** $\bar{v}$

1: **procedure** CFR$^+(\Gamma)$ ▷ $\Gamma = (H, Z, A, P, f_c, \mathcal{I}_1, \mathcal{I}_2, u)$
2:      **for all** $I \in \mathcal{I}_1$ **do**
3:          $r_0^I \leftarrow (0, \ldots, 0)^\top \in \mathbb{R}^{|A(I)|}$
4:          $\bar{x}^I \leftarrow (0, \ldots, 0)^\top \in \mathbb{R}^{|A(I)|}$
5:      **for** $t = 1, 2, 3, \ldots$ **do**
6:          play according to strategy x_t
7:          observe strategy y_t played by opponent
8:          TRAVERSE$(\varnothing, t, 1, 1)$
9:      **for all** $I \in \mathcal{I}_1$ **do**
10:         $\bar{x}^I \leftarrow \left(\bar{x}^I / \sum_{i=1}^{|A(I)|} \bar{x}_i^I \right)$
    ▷ $\bar{x}$ contains the average strategy for player 1.

where $\gamma \doteq \max_{x,y} u(x,y) - \min_{x,y} u(x,y)$.

Proof. Follows as a corollary of Theorem 5.

Notice that the bound provided by Theorem 5 is the same provided by the original CFR algorithm proposed by Zinkevich et al. [36]. In other words, our modification does not impair the convergence and speed guarantees given by the original algorithm.

8 Experimental Evaluation

We conducted experiments to investigate the practical performance of our perturbed-regret-minimization approach when used to instantiate the CFR and CFR+ algorithms for computing approximate EFPE in EFGs. We compare these algorithms to state-of-the-art Nash-equilibrium-finding algorithms: EGT [26] on an unperturbed polytope using the state-of-the-art smoothing technique by Kroer et. al. [20], CFR [36] and CFR+ [32]. We conducted the experiments on Leduc hold'em poker [31], a widely-used benchmark in the imperfect-information game-solving community. In our variant, *Leduc k*, the deck consists of k pairs of cards $1 \ldots k$, for a total deck size of $2k$. We experiment on the standard Leduc game where $k = 3$ and a larger game where $k = 5$. Each player initially pays one chip to the pot, and is dealt a single private card. After a round of betting, a community card is dealt face up. After a subsequent round of betting, if neither player has folded, both players reveal their private cards. If either player pairs their card with the community card they win the pot. Otherwise, the player with the highest private card wins. In the event that both players have the same private card, they draw and split the pot. We consider $k \in \{3, 5\}$. We test our approach on games subject to different uniform perturbations $p(I, a) = \xi$ for all information sets I and actions $a \in A(I)$, for $\xi \in \{0.1, 0.05, 0.01, 0.005, 0.001\}$.

Figure 1 reports on convergence to Nash equilibrium. The x-axis shows the number of tree traversals performed. We use tree traversals rather than iterations because EGT requires more tree traversals than CFR+ per iteration. The y-axis shows the sum of player regrets in the full (unperturbed) game. For both Leduc 3 and 5, we find that the ξ perturbations have only a small effect on overall convergence rate until convergence within the perturbed polytope, at which point the regret in the unperturbed game stops decreasing, as expected. Until bottoming out, the convergence is almost identical for all CFR+ algorithms. This shows that our approach can be utilized in practice: there is no substantial loss of convergence rate. Later in the run once the perturbed algorithms have bottomed out, there is a tradeoff between exploitability in the full game and refinement (i.e., better performance in low-probability information sets).

The second set of experiments, Figure 2, investigates the improvement that our perturbation approach achieves compared to standard Nash equilibrium solutions in terms of equilibrium refinement. Our measure of refinement is the maximum regret at any information set, conditioned on reaching that information set. As discussed, convergence to a Nash Equilibrium does not guarantee that this measure goes to zero. Again, the x-axis shows the number of tree traversals
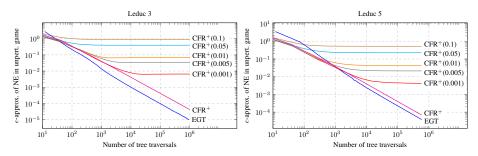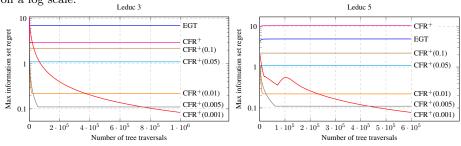
Fig. 1. Regret in Leduc 3 and Leduc 5 as a function of the number of iterations for EGT and CFR+ with various ξ perturbations (denoted in parentheses). Both axes are on a log scale.

Fig. 2. Maximum regret at any individual information set in Leduc 3 and Leduc 5, as a function of the number of iterations, for standard EGT as well as with various ξ perturbations (denoted EGT(ξ)) and CFR+. The y-axis is on a log scale.

performed. The y-axis shows the maximum regret at any individual information set. Both unperturbed CFR+ and EGT perform badly in both games with respect to this measure of refinement. In Leduc 5, both have maximum regret two orders of magnitude worse than the perturbed approach. In Leduc 3, EGT still does as poorly. CFR+ does slightly better, but is still worse than our stronger refinements by more than an order of magnitude. The maximum regret one can possibly cause in an information set in either Leduc game is 23, so CFR+ and unperturbed EGT also do poorly in that sense.

In contrast to this, we find that our ξ-perturbed solution concepts converge to a strategy with low regret at every information set. The choice of ξ is important: for $\xi = 0.001$, the smallest perturbation, we see that it takes a long time to converge at low-probability information sets, whereas we converge reasonably quickly for $\xi = 0.01$ or $\xi = 0.005$; for $\xi = 0.1$ and $\xi = 0.05$ the perturbations are too large, and we end up converging with relatively high regret (due to being forced to play every action with probability ξ). Thus, within this set of experiments, $\xi \in [0.005, 0.01]$ seems to be the ideal amount of perturbation.

9 Discussion

We extended the RM and RM$^+$ regret minimization algorithms to finitely-generated convex polytopes, and specialized our results to linearly-constrained

simplexes and behaviorally-perturbed EFGs. We then showed how this allows us to compute an approximate EFPE. Our experiments showed that this approach leads to much stronger strategies for information sets reached with low probability, while maintaining the strong convergence rate of CFR^+.

Our experiments raise an interesting question. Across both games, we see that the maximum information set regret goes down much faster for larger amounts of perturbation, but then it bottoms out earlier than for smaller perturbations (as expected). To get the best of both large and small perturbations, it may be possible to decrease the perturbations over time, leading to faster convergence rate, while never bottoming out. However, this has a number of challenges associated with it. Most importantly, we need a variant of RM or RM^+ that can handle a slowly expanding feasible set within the simplex. This would also require decreasing the perturbations at the correct rate; if decreased too quickly, it is unlikely that we will converge to a refinement, and if decreased too slowly, we might still bottom out.

The CFR algorithms have been shown to work well with a number of other techniques, notably sampling [23] and abstraction [22, 19]. It would be both theoretically and practically interesting to see how well our refinement approach works in conjunction with these techniques.

References

[1] Abernethy, J., Bartlett, P.L., Hazan, E.: Blackwell approachability and no-regret learning are equivalent. In: COLT. pp. 27–46 (2011)
[2] Blackwell, D.: An analog of the minmax theorem for vector payoffs. Pacific Journal of Mathematics 6, 1–8 (1956)
[3] Bowling, M., Burch, N., Johanson, M., Tammelin, O.: Heads-up limit hold'em poker is solved. Science 347(6218) (January 2015)
[4] Brown, N., Ganzfried, S., Sandholm, T.: Hierarchical abstraction, distributed equilibrium computation, and post-processing, with application to a champion no-limit Texas Hold'em agent. In: AAMAS (2015)
[5] Brown, N., Kroer, C., Sandholm, T.: Dynamic thresholding and pruning for regret minimization. In: AAAI (2017)
[6] Brown, N., Sandholm, T.: Regret-based pruning in extensive-form games. In: NIPS (2015)
[7] Brown, N., Sandholm, T.: Reduced space and faster convergence in imperfect-information games via pruning. In: ICML (2017)
[8] Cesa-Bianchi, N., Lugosi, G.: Prediction, learning, and games. Cambridge university press (2006)
[9] Farina, G., Gatti, N.: Extensive-form perfect equilibrium computation in two-player games. In: AAAI (2017)
[10] Ganzfried, S., Sandholm, T.: Game theory-based opponent modeling in large imperfect-information games. In: AAMAS (2011)
[11] Gordon, G.J.: No-regret algorithms for online convex programs. NIPS 19, 489 (2006)
[12] Gordon, G.J., Greenwald, A., Marks, C.: No-regret learning in convex games. In: ICML (2008)
[13] Hart, S., Mas-Colell, A.: A simple adaptive procedure leading to correlated equilibrium. Econometrica 68, 1127–1150 (2000)

[14] Hoda, S., Gilpin, A., Peña, J., Sandholm, T.: Smoothing techniques for computing Nash equilibria of sequential games. Mathematics of Operations Research 35(2) (2010)

[15] Johanson, M., Bowling, M.: Data biased robust counter strategies. In: AISTATS (2009)

[16] Johanson, M., Zinkevich, M., Bowling, M.: Computing robust counter-strategies. In: NIPS (2007)

[17] Koller, D., Megiddo, N., von Stengel, B.: Efficient computation of equilibria for extensive two-person games. Games and Economic Behavior 14(2) (1996)

[18] Kroer, C., Farina, G., Sandholm, T.: Smoothing method for approximate extensive-form perfect equilibrium. In: IJCAI (2017)

[19] Kroer, C., Sandholm, T.: Imperfect-recall abstractions with bounds in games. In: EC (2016)

[20] Kroer, C., Waugh, K., Kilinc-Karzan, F., Sandholm, T.: Theoretical and practical advances on smoothing for extensive-form games. In: EC (2017)

[21] Kroer, C., Waugh, K., Kılınç-Karzan, F., Sandholm, T.: Faster first-order methods for extensive-form game solving. In: EC (2015)

[22] Lanctot, M., Gibson, R., Burch, N., Zinkevich, M., Bowling, M.: No-regret learning in extensive-form games with imperfect recall. In: ICML (2012)

[23] Lanctot, M., Waugh, K., Zinkevich, M., Bowling, M.: Monte Carlo sampling for regret minimization in extensive games. In: NIPS (2009)

[24] Miltersen, P.B., Sørensen, T.B.: Computing a quasi-perfect equilibrium of a two-player game. Economic Theory 42(1) (2010)

[25] Neller, T.N., Lanctot, M.: An introduction to counterfactual regret minimization. Tutorial (2013), available from http://modelai.gettysburg.edu/2013/cfr/cfr.pdf

[26] Nesterov, Y.: Excessive gap technique in nonsmooth convex minimization. SIAM Journal of Optimization 16(1) (2005)

[27] Nesterov, Y.: Smooth minimization of non-smooth functions. Mathematical Programming 103 (2005)

[28] Romanovskii, I.: Reduction of a game with complete memory to a matrix game. Soviet Mathematics 3 (1962)

[29] Sandholm, T.: The state of solving large incomplete-information games, and application to poker. AI Magazine (2010), special issue on Algorithmic Game Theory

[30] Selten, R.: Reexamination of the perfectness concept for equilibrium points in extensive games. IJGT (1975)

[31] Southey, F., Bowling, M., Larson, B., Piccione, C., Burch, N., Billings, D., Rayner, C.: Bayes' bluff: Opponent modelling in poker. In: UAI (2005)

[32] Tammelin, O., Burch, N., Johanson, M., Bowling, M.: Solving heads-up limit Texas hold'em. In: IJCAI (2015)

[33] Telgarsky, M.: Blackwell Approachability and Minimax Theory. ArXiv e-prints (Oct 2011)

[34] van Damme, E.: A relation between perfect equilibria in extensive form games and proper equilibria in normal form games. IJGT (1984)

[35] von Stengel, B.: Efficient computation of behavior strategies. Games and Economic Behavior 14(2) (1996)

[36] Zinkevich, M., Bowling, M., Johanson, M., Piccione, C.: Regret minimization in games with incomplete information. In: NIPS (2007)