Counterfactual Regret Minimization in Sequential Security Games

Viliam Lisý and **Trevor Davis** and **Michael Bowling**
Department of Computing Science
University of Alberta, Edmonton, AB, Canada T6G 2E8
{lisy,trdavis1,bowling}@ualberta.ca

Abstract

Many real world security problems can be modelled as finite zero-sum games with structured sequential strategies and limited interactions between the players. An abstract class of games unifying these models are the normal-form games with sequential strategies (NFGSS). We show that all games from this class can be modelled as well-formed imperfect-recall extensive-form games and consequently can be solved by counterfactual regret minimization. We propose an adaptation of the CFR$^+$ algorithm for NFGSS and compare its performance to the standard methods based on linear programming and incremental game generation. We validate our approach on two security-inspired domains. We show that with a negligible loss in precision, CFR$^+$ can compute a Nash equilibrium with five times less computation than its competitors.

Game theory has been recently used to model many real world security problems, such as protecting airports (Pita et al. 2008) or airplanes (Tsai et al. 2009) from terrorist attacks, preventing fare evaders form misusing public transport (Yin et al. 2012), preventing attacks in computer networks (Durkota et al. 2015), or protecting wildlife from poachers (Fang, Stone, and Tambe 2015). Many of these security problems are sequential in nature. Rather than a single monolithic action, the players' strategies are formed by sequences of smaller individual decisions. For example, the ticket inspectors make a sequence of decisions about where to check tickets and which train to take; a network administrator protects the network against a sequence of actions an attacker uses to penetrate deeper into the network.

Sequential decision making in games has been extensively studied from various perspectives. Recent years have brought significant progress in solving massive imperfect-information extensive-form games with a focus on the game of poker. Counterfactual regret minimization (Zinkevich et al. 2008) is the family of algorithms that has facilitated much of this progress, with a recent incarnation (Tammelin et al. 2015) essentially solving for the first time a variant of poker commonly played by people (Bowling et al. 2015). However, there has not been any transfer of these results to research on real world security problems.

We focus on an abstract class of sequential games that can model many sequential security games, such as games taking place in physical space that can be discretized as a graph. This class of games is called normal-form games with sequential strategies (NFGSS) (Bosansky et al. 2015) and it includes, for example, existing game theoretic models of ticket inspection (Jiang et al. 2013), border patrolling (Bosansky et al. 2015), and securing road networks (Jain et al. 2011).

In this work we formally prove that any NFGSS can be modelled as a slightly generalized chance-relaxed skew well-formed imperfect-recall game (CRSWF) (Lanctot et al. 2012; Kroer and Sandholm 2014), a subclass of extensive-form games with imperfect recall in which counterfactual regret minimization is guaranteed to converge to the optimal strategy. We then show how to adapt the recent variant of the algorithm, CFR$^+$, directly to NFGSS and present experimental validation on two distinct domains modelling search games and ticket inspection. We show that CFR$^+$ is applicable and efficient in domains with imperfect recall that are substantially different from poker. Moreover, if we are willing to sacrifice a negligible degree of approximation, CFR$^+$ can find a solution substantially faster than methods traditionally used in research on security games, such as formulating the game as a linear program (LP) and incrementally building the game model by double oracle methods.

Game Model

NFGSS is a class of two-player zero-sum sequential games in which each player i's strategy space has the structure of a finite directed acyclic Markov decision process (MDP) with the set of states S_i, set of actions A_i and a stochastic transition function $T : S_i \times A_i \to \Delta(S_i)$. The transition function defines the probability of reaching next states after an action is executed in the current state. We denote by $\Delta(\cdot)$ a set of probability distributions over a set; $A(s_i) \subseteq A_i$ the actions applicable in state s_i; and s_i^0 the initial state of the MDP.

A player's strategy is a probability distribution over actions applicable in each state. We denote $\delta_i(s_i, a_i)$ to be the probability that a player following strategy δ_i reaches state s_i and then executes action a_i. An important restriction in this class of games is that the utility is defined in terms of *marginal utilities* for simultaneously executed state-action pairs $(s_i, a_i) \in S_i \times A_i$ by the players. The first player's

(negative of second player's) expected utility has the form:

$$u(\delta_1, \delta_2) = \sum_{S_1 \times A_1} \sum_{S_2 \times A_2} \delta_1(s_1, a_1)\delta_2(s_2, a_2)U((s_1, a_1), (s_2, a_2))$$

for some $U : S_1 \times A_1 \times S_2 \times A_2 \to \mathbb{R}$.

Background

In this paper, we show that the games from NFGSS can be transformed to a previously studied subclass of extensive-form games with imperfect recall. This section defines the concepts required to understand this transformation.

Extensive-form games

Two-player extensive-form games model sequential decision making of *players* denoted $i \in N = \{1, 2\}$. In turn, players choose *actions* leading to sequences called *histories* $h \in H$. A history $z \in Z$, where $Z \subseteq H$, is called a *terminal history* if it represents a full game from start to end. At each terminal history z there is a payoff $u_i(z)$ to each player i. At each nonterminal history h, there is a single current player to act, determined by $P : H \backslash Z \to N \cup \{c\}$ where c is a special player called *chance* that plays with a fixed known stochastic strategy. For example, chance is used to represent rolls of dice and card draws. The game starts in the empty history $\emptyset$, and at each step, given the current history h, the current player chooses an action $a \in A(h)$ leading to successor history $h' = ha$. We call h a *prefix* of h', denoted $h \sqsubseteq h'$, and generalize this relation to its transitive and reflexive closure.

Set $\mathcal{I}_i$ is a partition over $H_i = \{h \in H : P(h) = i\}$ where each part is called an *information set*. An information set $I \in \mathcal{I}_i$ of player i is a set of histories that a player cannot tell apart (due to information hidden from that player). For all $h, h' \in I$, $A(h) = A(h')$ and $P(h) = P(h')$; hence, we extend the definition to $A(I)$, $P(I)$, and denote $I(h)$ the information set containing h. Let $X(h)$ be the set of information set and action pairs on the path from the root of the game to history $h \in H$; $X_i(h) \subseteq X(h)$ be only the pairs where player i chooses an action; and $X_i(h, z)$ only the pairs i chooses on the path from h to z. A game has *perfect recall*, if players remember all actions they took: $\forall i \in N\ \forall h \in H_i\ \forall h' \in I(h)\ X_i(h') = X_i(h)$. If this condition does not hold, the game has *imperfect recall*. In this paper we don't consider a form of imperfect recall often called absent mindedness, in which a history and a prefix of that history may be contained in the same information set.

A *behavioral strategy* for player i is a function mapping each information set $I \in \mathcal{I}_i$ to a probability distribution over the actions $A(I)$, denoted by $\sigma_i(I)$. For a *profile* $\sigma = (\sigma_1, \sigma_2)$, we denote the probability of reaching a terminal history z under σ as $\pi^\sigma(z) = \prod_{i \in N \cup \{c\}} \pi_i^\sigma(z)$, where each $\pi_i^\sigma(z) = \prod_{ha \sqsubseteq z, P(h)=i} \sigma_i(I(h), a)$ is a product of probabilities of the actions taken by player i along z. We use $\pi_i^\sigma(h, z)$ and $\pi^\sigma(h, z)$ for $h \sqsubseteq z$ to refer to the product of only the probabilities of actions along the sequence from the end of h to the end of z. We define Σ_i to be the set of behavioral strategies for player i and extend the utility function to strategy profiles as $u_i(\sigma) = \sum_{z \in Z} \pi^\sigma(z) u_i(z)$. By convention, $-i$ refers to player i's opponent and chance player, or just the opponent if the context does not admit chance.

An *ϵ-Nash equilibrium*, σ, is a strategy profile such that the benefit of switching to some alternative σ_i' is limited by ϵ, i.e., $\forall i \in N : \max_{\sigma_i' \in \Sigma_i} u_i(\sigma_i', \sigma_{-i}) - u_i(\sigma) \leq \epsilon$. When $\epsilon = 0$, the profile is called a Nash equilibrium. We focus on zero-sum games, where $u_2(z) = -u_1(z)$ and define *precision* of a profile σ as the sum of strategies' distances from an equilibrium, $\epsilon_\sigma = \max_{\sigma_1' \in \Sigma_1} u_1(\sigma_1', \sigma_2) + \max_{\sigma_2' \in \Sigma_2} u_2(\sigma_1, \sigma_2')$.

Counterfactual Regret Minimization

Counterfactual Regret is a notion of regret at the information set level for extensive-form games (Zinkevich et al. 2008). The Counterfactual Regret minimization algorithms iteratively learn strategies in self-play, converging to an equilibrium. The *counterfactual value* of information set I is the expected payoff given that player i played to reach I, the opponent played σ_{-i} and both players played σ after I:

$$v_i(I, \sigma) = \sum_{z \in Z_I} \pi_{-i}^\sigma(z[I]) \pi^\sigma(z[I], z) u_i(z), \tag{1}$$

where $Z_I = \{z \in Z : \exists h \in I, h \sqsubseteq z\}$, $z[I] = h$ such that $h \sqsubseteq z, h \in I$, and $\pi_{-i}^\sigma(h)$ is the reach probability due to the opponent and chance. Define $\sigma_{I \to a}^t$ to be a strategy identical to σ^t except at I action a is taken with probability 1. The counterfactual regret of not taking $a \in A(I)$ at time t is $r^t(I, a) = v_i(I, \sigma_{I \to a}^t) - v_i(I, \sigma^t)$. We use the most recent CFR$^+$ (Tammelin et al. 2015) algorithm to minimize these regrets. This algorithm maintains values $Q^t(I, a) = \max(0, Q^{t-1}(I, a) + r^t(I, a))$ with $Q^0(I, a) = 0$ for every action at every information set. The strategy for the next iteration $\sigma^{t+1}(I)$ is proportional to the maintained values:

$$\sigma^{T+1}(I, a) = \begin{cases} Q^T(I, a) / Q_{sum}^T(I) & \text{if } Q_{sum}^T(I) > 0 \\ 1/|A(I)| & \text{otherwise,} \end{cases} \tag{2}$$

where $Q_{sum}^T(I) = \sum_{a' \in A(I)} Q^T(I, a')$. Furthermore, the algorithm maintains the weighted average strategy profile:

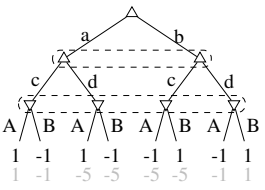
$$\bar{\sigma}^T(I, a) = \frac{\sum_{t=1}^T t \cdot \pi_i^{\sigma^t}(I) \sigma^t(I, a)}{\sum_{t=1}^T t \cdot \pi_i^{\sigma^t}(I)}, \tag{3}$$

where $\pi_i^{\sigma^t}(I) = \sum_{h \in I} \pi_i^{\sigma^t}(h)$. The combination of the counterfactual regret minimizers in individual information sets also minimizes the overall average regret, and hence the average profile is an ϵ-equilibrium, with $\epsilon \to 0$ as $T \to \infty$.

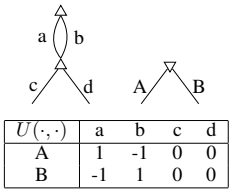
Well-formed Imperfect-Recall Games

Imperfect recall in general introduces complications in finding optimal solutions for games and the problem often becomes NP-hard (Koller and Megiddo 1992). Therefore Lanctot at al. (2012) introduce a subclass of imperfect-recall games which does not suffer from these problems. This subclass was further extended in (Kroer and Sandholm 2014).

For an imperfect-recall game Γ, we define its *perfect-recall refinement* Γ', which is the exact same game, but the information sets $\mathcal{I}_i$ are further partitioned to $\mathcal{I}_i'$ by splitting the histories that would violate the perfect-recall condition



(a) Imperfect-recall EFG



$U(\cdot,\cdot)$	a	b	c	d
A	1	-1	0	0
B	-1	1	0	0

(b) Equivalent NFGSS

Figure 1: Example of an imperfect-recall game and an equivalent NFGSS. With the grey utilities, the game cannot be represented as NFGSS.

to separate information sets. It means that histories h_1 and h_2 are in the same information set in the refinement if and only if they are in the same information set in the imperfect-recall game and $X_i(h_1) = X_i(h_2)$. We denote $D(I) \subset \mathcal{I}'_i$ the set of information sets $I \in \mathcal{I}_i$ is divided into.

Following (Kroer and Sandholm 2014), an extensive-form game with imperfect recall Γ is a *chance-relaxed skew well-formed* game with respect to its perfect-recall refinement Γ', if for all $i \in N, \bar{I} \in \mathcal{I}_i, I, J \in D(\bar{I})$, there exists a bijection $\phi : Z_I \to Z_J$, such that for all $z \in Z_I$:

1. The actions of the opponent and chance on the paths to the leaves mapped to each other are the same ($X_{-i}(z) = X_{-i}(\phi(z))$) or span the whole information sets on the same level if they differ ($\forall(\tilde{I}, a) \in X_{-i}(z) \setminus X_{-i}(\phi(z)), \forall z' \in Z_I : (\tilde{I}, a) \in X_{-i}(z')$). Since ϕ is a bijection, the roles of I and J are interchangeable.
2. The actions of player i after reaching $\bar{I}$ on paths to leaves which are mapped to each other are the same: $X_i(z[I], z) = X_i(\phi(z)[J], \phi(z))$.

Kroer and Sandholm (2014) show that in well-formed games, minimizing counterfactual regret in individual information sets also minimizes regret in its perfect-recall refinement. The relation between the regrets can be expressed in terms of the following error terms:

3. $|u_i(z) - \alpha_{I,J} u_i(\phi(z))| \leq \epsilon^R_{I,J}(z)$ for some fixed $\alpha_{I,J} \in \mathbb{R}$ is the reward error at z with respect to I, J;
4. $|\pi_0(z[I], z) - \pi_0(\phi(z)[J], \phi(z))| \leq \epsilon^0_{I,J}(z)$ is the leaf probability error at z with respect to I, J;
5. $\left| \frac{\pi_0(z[I])}{\pi_0(I)} - \frac{\pi_0(\phi(z)[J])}{\pi_0(J)} \right| \leq \epsilon^D_{I,J}(z[I])$ is the distribution error of $z[I]$.

For the clarity of exposition, we require all these errors to be 0. In that case, a direct consequence of Theorem 1 in (Kroer and Sandholm 2014) is that running a counterfactual regret minimization algorithm in the well-formed imperfect-recall game converges to a Nash equilibrium in the refinement.

NFGSS as CRSWF Game

This section proves that any NFGSS can also be represented as a well-formed imperfect-recall game. We start with a simple example showing the relation between the two classes of games and why the standard issues with imperfect recall do not occur in NFGSS. Figure 1a with the black utility values presents an imperfect-recall extensive-form game, which is equivalent to the NFGSS in Figure 1b. Even though the maximizing player $\triangle$ forgets whether she played a or b, her following decision does not influence the utilities. Therefore, it is a CRSWF game and it can be represented as an NFGSS. The same game structure with the grey utilities is a typical example of a problematic imperfect-recall game. It does not have a Nash equilibrium in behavioral strategies (Wichardt 2008), it has a non-zero reward error for any bijection between histories below a and b, and it cannot be represented as an NFGSS.

Before we state the main theorem, we need to extend the definition of the reward error in CRSWF games to allow forgetting the rewards accumulated in the past if they do not influence future observations, actions, or rewards.

Proposition 1. *If we define the reward error as*

$$|u_i(z) - \beta_I - \alpha_{I,J}\left(u_i(\phi(z)) - \beta_J\right)| \leq \epsilon^R_{I,J}(z)$$

for some fixed $\alpha_{I,J}, \beta_I, \beta_J \in \mathbb{R}$, then Theorem 1 from (Kroer and Sandholm 2014) still holds. Consequently, converging to zero counterfactual regret in all information sets translates to converging to a Nash equilibrium in any perfect recall refinement of a game in case of zero error terms.

The complete statement and the proof of this proposition are included in the appendix available online. The main idea of the proof is that for any information set $\bar{I}$ in the imperfect-recall game, we can create a modified game with utilities

$$u^{\bar{I}}_i(z) = \begin{cases} u_i(z) - \beta_I & \text{if } z \in Z_I, I \in D(\bar{I}) \\ u_i(z) & \text{otherwise.} \end{cases} \quad (4)$$

In this game, we can use Proposition 1 from (Kroer and Sandholm 2014) to bound regret in $\bar{I}$ with the original definition of the reward error. Finally, we show that the regrets in the modified games are exactly the same as the regrets in the original game, since the shifts β_I cancel out.

For an NFGSS G=$(S_1, A_1, S_2, A_2, T, U)$, a *corresponding* extensive-form game $\Gamma(G) = (N, \mathcal{A}, H, Z, \mathcal{I}, u)$ is:

$\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_c$ The set of player's actions in the EFG is the set of state-action pairs from his MDP: $\mathcal{A}_i = \{(s, a) : s \in S_i, a \in A(s)\}$. Execution of each state-action pair (s, a) leads to a chance node with actions representing uncertainty in the transition using actions from $\mathcal{A}_c = \{(s, a, s') : s, s' \in S_i, a \in A(s), T(s, a, s') > 0\}$. The chance probabilities of these actions are naturally derived from T.

H A history $h \in H$ is a sequence of actions in which each player's action is followed by a corresponding chance action as defined above. The ordering of the actions of different players can be arbitrary. We assume that each history starts with actions of the first player until she reaches a terminal state of her MDP. The actions of the second player follow afterwards. For convenience, we refer to histories in these corresponding games as a pair (h_{1+c}, h_{2+c}) with each player's actions and consequent chance actions and do not deal with action ordering among different players explicitly.

Z Let $ls_i : H \to S_i$ be a function returning the MDP state of player i after last chance action consequent to her action in a history. The terminal histories are the histories in which both players reached the terminal state of their MDP: $Z = \{h \in H : A(ls_1(h)) = A(ls_2(h)) = \emptyset\}$.

$\mathcal{I}$ Each information set corresponds to a node in a player's MDP. Histories h^1, h^2 in which player i takes action ($P(h^1) = P(h^2) = i$), belong to the same information set ($I(h^1) = I(h^2)$) if and only if they end by the same state for the player: $ls_i(h^1) = ls_i(h^2)$.

u The utility of player 1 in a terminal history z is the sum of marginal utilities for all pairs of actions in the history $u_1(z) = \sum_{(s_1,a_1)\in z_1} \sum_{(s_2,a_2)\in z_2} U((s_1, a_1), (s_2, a_2))$.

Theorem 2. [1]*Let G be an NFGSS, then the corresponding extensive-form game $\Gamma(G)$ is a chance-relaxed skew well-formed imperfect-recall game with zero errors.*

Proof. Consider $\Gamma' = (N, \mathcal{A}, H, Z, \mathcal{I}', u)$, a perfect-recall refinement of $\Gamma(G)$, where each player's information sets are subdivided based on unique histories of her and subsequent chance actions (i.e., a path in the MDP). To show that Γ is well-formed with respect to Γ', we define for each $I, J \in D(\bar{I}), \bar{I} \in \mathcal{I}$ the bijection $\phi : Z_I \to Z_J$ to map to each other the terminal histories passing through I and J with the same opponent's actions and the same player $i = P(I)$'s actions after I or J is reached.

$$\forall z = (z_{i+c}, z_{-i+c}) \in Z_I \; \phi(z) = (h(J)_{i+c} z[I-]_{i+c}, z_{-i+c})$$

where $h(J)_{i+c}$ denotes the unique player $P(J)$'s and consequent chance history leading to J in the refinement Γ' and $z[I-]$ denotes the suffix of history z after reaching information set I. Subscript $i+c$ added to any (partial) history refers to only actions of player i and the consequent chance actions. The remaining actions are referred to by $-i+c$. This bijection is well-defined, because the possible future sequences of actions of player i are the same after reaching the same node in his MDP using different paths.

We follow by checking if all conditions required for a well-formed imperfect-recall game are satisfied with ϕ.

1) The actions of the opponent and their consequent chance actions are exactly the same in the mapping. The only complication could be the player i's chance actions, which generally differ for histories in I and J. However, the outcome of the chance nodes is always known to the player before he selects a next action. Therefore, in the perfect-recall refinement, all histories in each information set contain the exact same chance action on each level.

2) The second condition is trivially satisfied directly from the definition of ϕ.

3) The original definition of CRSWF games is not sufficient to guarantee zero reward error for all histories. Consider an example with the structure exactly the same as in Figure 1, but the marginal utilities $U(A, b) = 5, U(A, c) = 1$ and zero for all other action pairs. Let the information sets in the refinement be $I = \{a\}$ and $J = \{b\}$. Based on the definition of ϕ and u above:

$$\phi(acA) = bcA, \quad \phi(adA) = bdA$$

$$u(acA) = 1,\ u(bcA) = 6,\ u(adA) = 0,\ u(bdA) = 5.$$

There is no $\alpha_{I,J}$, such that $|1 - \alpha_{I,J}6| = 0 = |0 - \alpha_{I,J}5|$. However, this game can still be solved by counterfactual regret minimization thanks to Proposition 1. Past rewards can be forgotten and the future rewards are exactly the same.

4) The histories below information sets I and J are formed by the exact same chance actions. Therefore, the leaf probability error is zero.

5) We established above that all histories in any information set I in the refinement include the exact same chance actions of the player $i = P(I)$ on the way to I; therefore $\pi_0(z[I]_{i+c})/\pi_0(I) = 1/|I|$. There exist a bijection between I and J; hence, $|I| = |J|$. The distribution error is

$$\left| \frac{\pi_0(z[I]_{i+c})\pi_0(z[I]_{-i+c})}{\pi_0(I)} - \frac{\pi_0(\phi(z)[J]_{i+c})\pi_0(\phi(z)[J]_{-i+c})}{\pi_0(J)} \right| = \\ = \frac{1}{|I|} \left|\pi_0(z[I]_{-i+c}) - \pi_0(\phi(z)[J]_{-i+c})\right| = 0. \quad (5)$$

The last equality holds because the opponent's chance actions are copied in ϕ. ■

There are two main consequences of this transformation. First, we can run counterfactual regret minimization on the EFGs corresponding to NFGSS and we are guaranteed that these algorithms converge to a well defined Nash equilibrium solution. Second, since the computed equilibrium is an equilibrium in any perfect-recall refinement, it proves that the players gain no benefit from remembering the path that they take to individual states of an NFGSS.

CFR$^+$ for NFGSS

Previous works using CFR in imperfect-recall games use random sampling to update counterfactual regrets (Waugh et al. 2009). For NFGSS, we derive a more efficient version of the algorithm which does not require sampling.

The algorithm stores for each action-value pair the special form of regret Q defined in CFR$^+$ and the mean strategy $\bar{\sigma}$; and for each MDP state the probability $p(s)$ of reaching the state under the current strategy. The current strategies σ can be stored or always computed from Q to save memory. The pseudocode is presented in Figure 2. In the main part of the algorithm denoted by NFGSS-CFR$^+$, each iteration of the algorithm consists of separate updates for individual players (lines 4-6). First, the current strategy of the opponent is used to compute reach probabilities for all states and update the mean strategy in his MDP (line 5). The reach probabilities are used to compute the regrets Q and new strategies for all action-value pairs of the player (line 6). Finally, we need to normalize the mean strategies, since they are stored as sums.

The update of the probabilities on lines 1-6 of the second function in Figure 2 is straightforward. Since the MDPs are assumed to be acyclic, we can store the states in topological order and easily traverse them from the root to leaves, always having the predecessors resolved before reaching the successors. Once a new probability is computed, the mean strategy can be updated immediately on line 7. For updating regrets and strategies in the third function, we process the state in the reverse order. For each state, we compute the expected value of the game after it reaches the state $v(s_i)$ and the expected value of playing each action available in

[1]This claim does not hold; see the appendix for details.

NFGSS-CFR$^+$

1: $\forall i \in N\ \sigma_i :=$ uniform strategy
2: $\forall i \in N\ Q_i := 0, \bar{\sigma}_i = 0$
3: **for** iteration $t \in (1, 2, \dots)$ **do**
4: **for** $i \in N$ **do**
5: UpdateStateProbabilitiesMeanStrategies($-i, t$)
6: UpdateActionRegretsCurStrategies(i)
7: $\forall i \in N$ Normalize($\bar{\sigma}_i$)

UpdateStateProbabilitiesMeanStrategies(i, t)

1: $\forall s_i \in S_i\ p(s_i) = 0$
2: $p(s_i^0) := 1$
3: **for** $s_i \in S_i$ in topological order **do**
4: **for** $a_i \in A(s_i)$ **do**
5: **for** $s' \in T(s_i, a_i)$ **do**
6: $p(s')$ += $p(s_i)\sigma(s_i, a_i)T(s_i, a_i, s')$
7: $\bar{\sigma}_i(s_i, a_i)$ += $tp(s_i)\sigma_i(s_i, a_i)$

UpdateActionRegretsCurStrategies(i)

1: $\forall s_i \in S_i\ v(s_i) = 0$
2: **for** $s_i \in S_i$ in reverse topological order **do**
3: **for** $a_i \in A(s_i)$ **do**
4: $v(a_i) := 0$
5: **for** $s' \in T(s_i, a_i)$ **do**
6: $v(a_i)$ += $T(s_i, a_i, s')v(s')$
7: **for** $s_{-i}, a_{-i} \in S_{-i} \times A_{-i}$ **do**
8: $v(a_i)$ += $p(s_{-i})\sigma(s_{-i}, a_{-i})U_i(s_i, a_i, s_{-i}, a_{-i})$
9: $v(s_i)$ += $\sigma(s_i, a_i)v(a_i)$
10: **for** $a_i \in A(s_i)$ **do**
11: $Q(s_i, a_i) := \max(0, Q(s_i, a_i) + v(a_i) - v(s_i))$
12: $\sigma_i(s_i) :=$ RegretMatching($Q(s_i)$)

Figure 2: Adaptation of CFR$^+$ for NFGSS

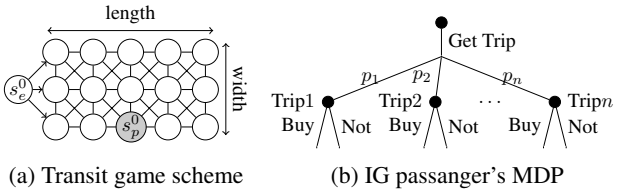
the state (lines 4-9). The value of the action is the sum of the expected value of the successors (lines 5-6) and the marginal utility of executing the action, with respect to the probabilities that the opponent executes her actions (lines 7-8). We implement this step efficiently using sparse representation of U. The expected value of the state is the sum of the expected values of executing individual actions weighted by their probability in the current strategy (line 9). The regrets are updated as prescribed by CFR$^+$ (lines 10-11) and the new strategy is computed by regret matching (Equation 2).

Proposition 3. *The NFGSS-CFR$^+$ algorithm converges to an ϵ-Nash equilibrium of the game after*

$$T \leq \max_{i \in N} \frac{\Delta^2 |S_i|^2 |A_i|}{\epsilon^2}$$

iterations, where Δ is the overall range of (not marginal) utility values in the game.

This proposition holds for CFR (and therefore also CFR$^+$) in CRSWF extensive-form games with zero error terms (Lanctot et al. 2012). We only have to show that NFGSS-CFR$^+$ performs exactly the same updates as it would perform in the equivalent EFG defined above. Focus on a specific information set I in iteration t and assume all updates in previous iterations were equivalent. On iteration t, the current strategy is the same and therefore, the values $v(a_i)$ and $v(s_i)$ computed on lines 3-9 of the algorithm represent the same counterfactual values as the values computed on the equivalent EFG based on Equation 1. Since the update is defined only in terms of counterfactual values and the current strategy (in case of mean strategy), all updates are exactly the same.



(a) Transit game scheme (b) IG passanger's MDP

Figure 3: Evaluation domain schemas.

Experimental Evaluation

Transit game (TG) is the game used for evaluation in (Bosansky et al. 2015). For fair comparison, we have implemented CFR$^+$ within their publicly available code base. The game is a search game in Euclidean space discretized as an eight-connected rectangular grid, such as in Figure 3a. The evader attempts to cross this grid from left to right without meeting the patroller. Each move incurs a small penalty (0.02) to the evader. The patroller starts in his base marked in grey and moves on the graph along the edges (or stays at its current node) for d time steps. If he is not back at the base at the end of the game, he suffers a penalty (20). The movement of each player fails with probability 0.1, which causes the player to stay at its previous position. Every time the players meet, the evader looses one point of utility to the patroller. The authors in (Bosansky et al. 2015) approximate this game as zero-sum; hence, all the penalties suffered by one player are considered to be gains of the other player. In the evaluation below, we consider the transit game of size w to be played on $w \times 2w$ grid for $d = 2w + 4$ time steps.

Ticket inspection game (IG) is based on (Jiang et al. 2013). It models scheduling a three hours long shift of ticket inspectors on a single train line. Jiang at al. use real world data from the LA metro system and schedule for multiple patrols. We consider a single patrol and generate synthetic data to create problem instances of varying size. The data include (1) the train schedules, which are in both directions generated non-uniformly, starting with approximately 5 times longer intervals at the beginning of the shift than in the peak time close to the end of hour two of the shift; (2) the passenger trips defined by the number of passengers that take each train between each pair of stations generated pseudo-randomly based on non-uniform station popularity values; and (3) the time intervals in which a train reaches a following station. The patrol starts in the middle of the line. In a station, the patrol can take the next train in either direction or check the tickets of the passengers in the station for 15 minutes. On a train, she can either check the tickets of the passengers on the train until the next stop, or exit the train. The patrol checks 5 passengers per minute and the passengers

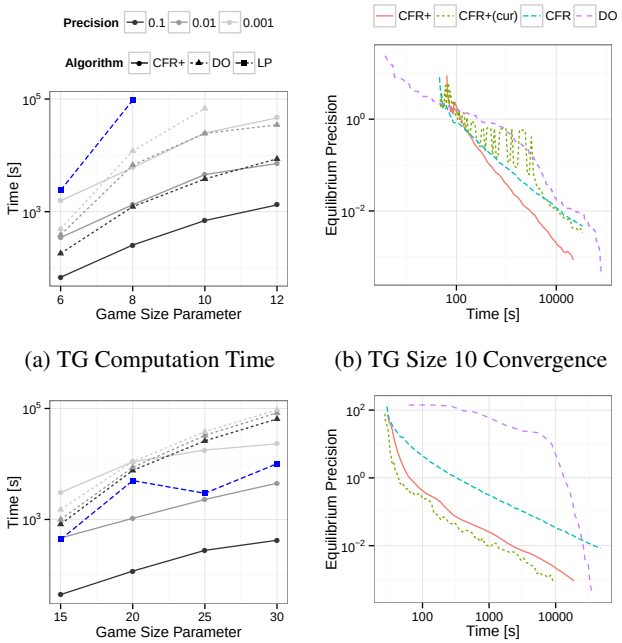
are assumed to be present in the stations 3 minutes before their train departs and 3 minutes after it arrives. Because of unexpected delays, the patrol can with probability 0.1 miss the train it intended to take and stay at the station, or fail to exit the train at the intended station and stay in the train until the next one. The game has originally been modelled as a Bayesian game with passengers taking the same trip as types, but it can be also seen as NFGSS. Instead of types, we model the passengers' strategy by the MDP in Figure 3b. The MDP starts with a dummy action with one outcome for each trip, occurring with the probability that a random passenger takes the trip. In each of these outcomes, the passenger chooses an action representing either buying a ticket or not buying a ticket. The MDP of the patroller starts with a dummy action of collecting the fare money, which gives him a reward of \$1.50 for each passenger that buys a ticket. Afterwards, the MDP describes her movement on the line and the patrol gets a reward of \$100 for each checked passenger which did not buy a ticket. The game is zero-sum, with the patrol maximizing its revenue per passenger and each passenger type minimizing its cost. The ticket inspection game of size s has s stations, s trains in each direction and approximately s thousand passengers.

Results

We compare the run time of the proposed NFGSS-CFR$^+$ algorithm to the full compact strategy LP formulation and the double oracle (DO) algorithms proposed in (Bosansky et al. 2015). Since CFR$^+$ is generally used as an approximative algorithm, we compare the run time of the algorithms with various target precision: 0.1, 0.01 and 0.001 in absolute utility value. For solving LPs, we used IBM CPLEX 12.51. The simplex algorithm is substantially faster than the barrier method on large instances of these problems, so we use simplex. The precision of CPLEX is by default set to 10^{-6}. Setting it to a higher value increased the run time, most likely because of "Harris' method"-like feasibility bounds shifts. The limit on these shifts is also controlled by the precision parameter. When these shifts are eventually removed, the solution is in a substantially more infeasible state, which is harder to correct and the correction has a larger negative effect on optimality (Klotz and Newman 2013). Therefore, we evaluate only the default setting. The results presented in Figures 4(a,c) are means of 10 runs with small a variance and Figures 4(b,d) are each a single representative run.

The precision of 0.01 in the transit game is half the penalty the evader pays for each move. We compute the utility in IG in cents; hence, the error of 0.01 per passenger means a loss of \$3 of overall revenue in the largest game with 30 thousand passengers. Since the models are always an approximation of the real world problems, we consider error of 0.1 to be perfectly acceptable and the error smaller than 0.01 irrelevant from the domain perspective.

Figure 4a presents the computation times required by the algorithms to reach the given precision in TG. This domain is suitable for the double oracle algorithm; hence, in games of all evaluated sizes, DO outperforms LP for any precision. LP was not able to solve the larger games within 30 hours. For precision of 0.1 and 0.01, CFR$^+$ clearly outperforms DO



(a) TG Computation Time (b) TG Size 10 Convergence

(c) IG Computation Time (d) IG Size 25 Convergence

Figure 4: Computation times and convergence curves for Transit (a,b) and Ticket inspection (c,d) games. CFR+(cur) is the current strategy of CFR$^+$.

for all game sizes by up to factor of 5 (note the log scale). For the largest game with width 12 (5041 and 6118 states in the MDPs), CFR$^+$ finds the solution with precision 0.01 in 119 minutes, while for DO, it takes almost 10 hours. Two hours are enough for DO to reach precision of 0.01 in game of size 8 with 1729 and 2036 MDP states. For very high precision, DO already outperforms CFR$^+$ on the smallest game instance. Figure 4b presents the progress of the convergence in time on a log-log plot. Besides CFR$^+$ and DO, we present also the adaptation of standard CFR to NFGSS analogous to the presented adaptation of CFR$^+$, and the performance of the current strategy of CFR$^+$. The current strategy has no formal guarantees, but has been observed to converge to an equilibrium in some poker games. We can see that the mean strategy of CFR$^+$ converges the fastest, but the current strategy also converges, which allows reducing the memory requirements of the algorithm by not storing of the average strategy. DO initially converges slowly, but eventually starts converging very quickly, which allows it to reach higher precisions before CFR$^+$.

We present the exact same experiments for the ticket inspection game in Figures 4c and 4d. The LP formulation of the game always outperforms DO. With the exception of the smallest game, CFR$^+$ reaches precision of 0.1 and 0.01 before the LP finishes. In the largest game (4699 and 15362 MDP states), CFR$^+$ reaches the precision of 0.01 after 74 minutes, while the LP requires almost 226 minutes. Even in this domain, CFR$^+$ converges faster than CFR. The current strategy converges even faster than the average strategy.

Conclusions

We study an abstract class of games called NFGSS, suitable for modelling a variety of real world security problems. We transfer several key poker research results to this class. We extend the previously studied notion of well-formed imperfect-recall games and show that after this extension, it can model any game from NFGSS. We propose an adaptation of CFR$^+$ for this class and provide a formal guarantee that it converges to a Nash equilibrium. We empirically show that with a small loss of precision, it allows solving larger problem instances with up to five times less computation time than the currently used approaches. With CFR$^+$, we can solve the game models more frequently (e.g., if they are not entirely static) or solve substantially larger game instances in the same time as with standard methods.

This paper opens two natural directions of future research. First, since incremental game generation in double oracle algorithms is, to a large extent, orthogonal to the actual equilibrium solving algorithm, combining it with CFR$^+$ might lead to an additional speedup. Second, the link we established between NFGSS and CRSWF games can help to further extend these classes of games to allow modelling more complex interactions between players' strategies.

Acknowledgments

We thank Branislav Bosansky for providing the source codes and support for the LP based methods. This research was supported by Alberta Innovates Technology Futures through the Alberta Innovates Centre for Machine Learning and Reinforcement Learning and AI Lab, and used the computing resources of Compute Canada and Calcul Quebec.

References

Bosansky, B.; Jiang, A. X.; Tambe, M.; and Kiekintveld, C. 2015. Combining compact representation and incremental generation in large games with sequential strategies. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Bowling, M.; Burch, N.; Johanson, M.; and Tammelin, O. 2015. Heads-up limit holdem poker is solved. *Science* 347(6218):145–149.

Durkota, K.; Lisy, V.; Bosansky, B.; and Kiekintveld, C. 2015. Optimal network security hardening using attack graph games. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 526–532.

Fang, F.; Stone, P.; and Tambe, M. 2015. Defender strategies in domains involving frequent adversary interaction. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, 1663–1664. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Jain, M.; Korzhyk, D.; Vaněk, O.; Conitzer, V.; Pěchouček, M.; and Tambe, M. 2011. A double oracle algorithm for zero-sum security games on graphs. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 327–334. International Foundation for Autonomous Agents and Multiagent Systems.

Jiang, A. X.; Yin, Z.; Zhang, C.; Tambe, M.; and Kraus, S. 2013. Game-theoretic randomization for security patrolling with dynamic execution uncertainty. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '13, 207–214. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Klotz, E., and Newman, A. M. 2013. Practical guidelines for solving difficult linear programs. *Surveys in Operations Research and Management Science* 18(12):1 – 17.

Koller, D., and Megiddo, N. 1992. The complexity of two-person zero-sum games in extensive form. *Games and Economic Behavior* 4(4):528 – 552.

Kroer, C., and Sandholm, T. 2014. Extensive-form game imperfect-recall abstractions with bounds. *arXiv preprint arXiv:1409.3302*.

Lanctot, M.; Gibson, R.; Burch, N.; Zinkevich, M.; and Bowling, M. 2012. No-regret learning in extensive-form games with imperfect recall. In Langford, J., and Pineau, J., eds., *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, 65–72. New York, NY, USA: Omnipress.

Pita, J.; Jain, M.; Marecki, J.; Ordóñez, F.; Portway, C.; Tambe, M.; Western, C.; Paruchuri, P.; and Kraus, S. 2008. Deployed armor protection: the application of a game theoretic model for security at the los angeles international airport. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: industrial track*, 125–132. International Foundation for Autonomous Agents and Multiagent Systems.

Tammelin, O.; Burch, N.; Johanson, M.; and Bowling, M. 2015. Solving heads-up limit texas hold'em. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*.

Tsai, J.; Rathi, S.; Kiekintveld, C.; Ordóñez, F.; and Tambe, M. 2009. IRIS - A Tool for Strategic Security Allocation in Transportation Networks Categories and Subject Descriptors. In *Proc. of the 8th Int. Conf. on Autonomous Agents and Multiagent Systems*, 37–44.

Waugh, K.; Zinkevich, M.; Johanson, M.; Kan, M.; Schnizlein, D.; and Bowling, M. 2009. A practical use of imperfect recall. In *Proceedings of the 8th Symposium on Abstraction, Reformulation and Approximation (SARA)*, 175–182.

Wichardt, P. C. 2008. Existence of nash equilibria in finite extensive form games with imperfect recall: A counterexample. *Games and Economic Behavior* 63(1):366–369.

Yin, Z.; Jiang, A. X.; Johnson, M. P.; Tambe, M.; Kiekintveld, C.; Leyton-Brown, K.; Sandholm, T.; and Sullivan, J. P. 2012. TRUSTS: Scheduling Randomized Patrols for Fare Inspection in Transit Systems. In *Proceedings of 24th Conference on Innovative Applications of Artificial Intelligence (IAAI)*.

Zinkevich, M.; Johanson, M.; Bowling, M.; and Piccione, C. 2008. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 1729–1736.

Proof of Proposition 1

Proposition 1 from (Kroer and Sandholm 2014) says that for any CRSWF game Γ, its refinement Γ', strategy profile σ and information set $\bar{I}$ in $\mathcal{I}$ such that player i has bounded regret $r(\bar{I}, a)$ for all $a \in A(\bar{I})$, the regret $r(I, a')$ at any information set $I \in D(\bar{I})$ and action $a' \in A_I$ is bounded by

$$r(I, a') \leq \max_{J \in D(\bar{I})} \alpha_{I,J} r(\bar{I}, a') + 2 \sum_{h \in I} \frac{\pi^\sigma(h)}{\pi^\sigma(I)} \left(\epsilon^0_{I,J}(h) + \epsilon^R_{I,J}(h) \right) + \epsilon^D_{I,J} \tag{6}$$

Let Γ be a CRSWF game with perfect recall refinement Γ'. Let $\bar{I} \in \mathcal{I}$ be an information set in the CRSWF game. We define a modified game $\Gamma_{\bar{I}}$ by substituting a new utility function $u^{\bar{I}}$ for the utility function u of Γ. $u^{\bar{I}}$ is defined such that

$$u_i^{\bar{I}}(z) = \begin{cases} u_i(z) & \text{if } z \notin Z_{\bar{I}} \\ u_i(z) - \beta_I & \text{if } z \in Z_I \text{ where } I \in \mathcal{P}(\bar{I}) \end{cases} \tag{7}$$

Note that $u^{\bar{I}}$ is well-defined: if $z \in Z_{\bar{I}}$, then there is some unique $I \in \mathcal{P}(\bar{I})$ such that $z[\bar{I}] \in I$. Because Γ is CRSWF and the CRSWF conditions are not dependent on the utility structure of the game, $\Gamma_{\bar{I}}$ is also CRSWF. Let $\Gamma'_{\bar{I}}$ be the perfect recall refinement of $\Gamma_{\bar{I}}$. Fix σ to be a strategy profile for Γ (and thus also for $\Gamma_{\bar{I}}$). Let $a \in A(\bar{I})$ be any action. Then we can use Proposition 1 from Kroer and Sandholm to bound the regret of a' in $\Gamma'_{\bar{I}}$:

$$r_{\bar{I}}(I, a') \leq \max_{J \in \mathcal{P}(\bar{I})} \alpha_{I,J} r_{\bar{I}}(\bar{I}, a') + 2 \sum_{h \in I} \frac{\pi^\sigma(h)}{\pi^\sigma(I)} \left(\epsilon^0_{I,J}(h) + \epsilon^{R,\bar{I}}_{I,J}(h) \right) + \epsilon^D_{I,J} \tag{8}$$

In this equation $r_{\bar{I}}(I, a')$ and $r_{\bar{I}}(\bar{I}, a')$ use the utility structure of $\Gamma_{\bar{I}}$, as does the reward error $\epsilon^{R,\bar{I}}_{I,J}(h)$. The other terms are independent of utility and thus also apply to Γ. Let $Z_h = \{z \in Z \mid h \sqsubseteq z\}$ be the set of terminal histories which can be reached from h. We now show that regret in $\Gamma_{\bar{I}}$ is the same as in Γ.

$$\begin{aligned}
r_{\bar{I}}(\bar{I}, a) =& \sum_{h \in \bar{I}} \frac{\pi^\sigma(h)}{\pi^\sigma(\bar{I})} \sum_{z \in Z_{(ha)}} \pi^\sigma(ha, z) u_i^{\bar{I}}(z) - \sum_{h \in \bar{I}} \frac{\pi^\sigma(h)}{\pi^\sigma(\bar{I})} \sum_{a' \in A(I)} \pi^\sigma(\bar{I}, a') \sum_{z \in Z_{(ha')}} \pi^\sigma(ha', z) u_i^{\bar{I}}(z) \\
=& \sum_{I \in \mathcal{P}(I)} \sum_{h \in I} \frac{\pi^\sigma(h)}{\pi^\sigma(\bar{I})} \left(\sum_{z \in Z_{(ha)}} \pi^\sigma(ha, z) u_i^{\bar{I}}(z) - \sum_{a' \in A(I)} \pi^\sigma(\bar{I}, a') \sum_{z \in Z_{(ha')}} \pi^\sigma(ha', z) u_i^{\bar{I}}(z) \right) \\
=& \sum_{I \in \mathcal{P}(I)} \sum_{h \in I} \frac{\pi^\sigma(h)}{\pi^\sigma(\bar{I})} \left(\sum_{z \in Z_{(ha)}} \pi^\sigma(ha, z)(u_i(z) - \beta_I) - \sum_{a' \in A(I)} \pi^\sigma(\bar{I}, a') \sum_{z \in Z_{(ha')}} \pi^\sigma(ha', z)(u_i(z) - \beta_I) \right) \\
=& \sum_{I \in \mathcal{P}(I)} \sum_{h \in I} \frac{\pi^\sigma(h)}{\pi^\sigma(\bar{I})} \left(\sum_{z \in Z_{(ha)}} \pi^\sigma(ha, z) u_i(z) - \sum_{a' \in A(I)} \pi^\sigma(\bar{I}, a') \sum_{z \in Z_{(ha')}} \pi^\sigma(ha', z) u_i(z) \right. \\
& \qquad \left. - \beta_I \left(\sum_{z \in Z_{(ha)}} \pi^\sigma(ha, z) - \sum_{a' \in A(I)} \pi^\sigma(\bar{I}, a') \sum_{z \in Z_{(ha')}} \pi^\sigma(ha', z) \right) \right) \\
=& \sum_{I \in \mathcal{P}(I)} \sum_{h \in I} \frac{\pi^\sigma(h)}{\pi^\sigma(\bar{I})} \left(\sum_{z \in Z_{(ha)}} \pi^\sigma(ha, z) u_i(z) - \sum_{a' \in A(I)} \pi^\sigma(\bar{I}, a') \sum_{z \in Z_{(ha')}} \pi^\sigma(ha', z) u_i(z) \right) = r(\bar{I}, a)
\end{aligned}$$

Where $r(\bar{I}, a)$ is the regret of action a with strategy σ at $\bar{I}$ in Γ. A similar proof shows that $r_{\bar{I}}(I, a) = r(I, a)$ for each $I \in \mathcal{P}(\bar{I})$. Combining these facts with (6) lets us bound the regret in Γ'.

$$\begin{aligned}
r(I, a) = r_{\bar{I}}(I, a) &\leq \max_{J \in \mathcal{P}(\bar{I})} \alpha_{I,J} r_{\bar{I}}(\bar{I}, a) + 2 \sum_{h \in I} \frac{\pi^\sigma(h)}{\pi^\sigma(I)} \left(\epsilon^0_{I,J}(h) + \epsilon^{R,\bar{I}}_{I,J}(h) \right) + \epsilon^D_{I,J} \\
&= \max_{J \in \mathcal{P}(\bar{I})} \alpha_{I,J} r(\bar{I}, a) + 2 \sum_{h \in I} \frac{\pi^\sigma(h)}{\pi^\sigma(I)} \left(\epsilon^0_{I,J}(h) + \epsilon^{R,\bar{I}}_{I,J}(h) \right) + \epsilon^D_{I,J}
\end{aligned}$$

And here $\epsilon^{R,\bar{I}}_{I,J}(h)$ is defined as a function of terminal reward errors $\epsilon^{R,\bar{I}}_{I,J}(z)$ such that

$$\epsilon^{R,\bar{I}}_{I,J}(z) \geq |u_i^{\bar{I}}(z) - \alpha_{I,J} u_i^{\bar{I}}(\phi(z))| = |(u_i(z) - \beta_I) - \alpha_{I,J}(u_i(\phi(z)) - \beta_J)|.$$

So we have shown that Proposition 1 still applies in Γ' when we shift the utility of each information set before defining the reward error.

Mistake in Theorem 2

Theorem 2 in the paper is unfortunately wrong. The proposed extension of CRSWF games is not sufficient to include all NFGSS. The mistake in the proof is in reward error. The rewards realized after an information set may depend on a (possibly forgotten) decision before the information set, which was overlooked in the proof. A simple counterexample is the game from Figure 1 with marginal utility U(A,a)=2 and all other utilities identical. However, CFR$^+$ is still guaranteed to converge in NFGSS and we present an alternative proof below.

Convergence of CFR$^+$ in NFGSS

Consider an imperfect recall EFG constructed from an NFGSS as outlined in the paper. In this proof, we will show that when CFR$^+$ is used to select strategies in the imperfect recall game, the average regret of these strategies converges to 0 in the perfect recall refinement. As a result, CFR$^+$ safely converges to a Nash equilibrium in the perfect recall refinement.

The general strategy of this proof is to show that a player's utility at an information set can be decomposed into utility generated by actions leading to the information set and utility generated by actions taken after the information set, with the former cancelling out in the counterfactual regret calculation. This in turn means that information sets which are distinct in the perfect recall refinement but combined in the imperfect recall game must have proportional regrets, as they differ only in the actions leading to the information set. Regret in the imperfect recall game is the sum of regrets in the perfect recall refinement, so when imperfect recall regret is minimized, each of the perfect recall regrets must be proportionately minimized.

For any history h in the EFG, we can decompose it into partial histories as $h = (h_1, h_2)$ where each h_i contains only actions of player i and the chance actions from the MDP of player i. Note that we further use only index i, but for the rest of this appendix, we always mean $i{+}c$ with partial histories. In a perfect recall game a player cannot forget any of his previous actions (or the corresponding chance events, which are immediately observed in the MDP), so if there are $h, h' \in I$ for a perfect recall information set I where $P(I) = i$, then it must be the case that $h_i = h'_i$. Thus for any perfect recall information set I where $P(I) = i$, each history in I contains a unique partial history for player i, which we denote h_i^I.

For any terminal history z, recall $z[I]$ is the unique $h \in I$ such that $h \sqsubseteq z$ and $z[I-]$ to be the suffix of z after $z[I]$ (so $z = z[I]z[I-]$). We extend these definitions to player partial histories: we decompose $z = (z_1, z_2)$, and when $P(I) = i$ we define $z_i[I]$ to be the unique partial history h_i such that $h_i \sqsubseteq z_i$; and $h_i \in h$ for some $h \in I$,[2] and define $z_i[I-]$ to be the suffix of z_i after $z_i[I]$.

We can also decompose the the utility of a terminal history. If $z \in Z, P(I) = i$ and we decompose z as $z = (z_i[I]z_i[I-], z_{-i})$, then by the properties of the NFGSS

$$
\begin{aligned}
u(z) &= u(z_i, z_{-i}) && (9)\\
&= \sum_{(s_i,a_i)\in z_i} \sum_{(s_{-i},a_{-i})\in z_{-i}} U((s_i,a_i),(s_{-i},a_{-i})) && (10)\\
&= \sum_{(s_i,a_i)\in z_i[I]} \sum_{(s_{-i},a_{-i})\in z_{-i}} U((s_i,a_i),(s_{-i},a_{-i})) + \sum_{(s_i,a_i)\in z_i[I-]} \sum_{(s_{-i},a_{-i})\in z_{-i}} U((s_i,a_i),(s_{-i},a_{-i})) && (11)
\end{aligned}
$$

To reflect this property, we overload the u function and write $u(z_i, z_{-i}) = u(z_i[I], z_{-i}) + u(z_i[I-], z_{-i})$.

Define $\pi_{ci}(h)$ to be the contribution to $\pi(h)$ only from the chance nodes that follow actions of player i (chance events from the MDP of player i). Thus $\pi(h) = \pi_i(h)\pi_{ci}(h)\pi_{-i+c}(h)$. We are now ready to show that regrets in the perfect recall refinement are proportional:

Lemma 4. *Let $\bar{I}$ be an information set in an EFG corresponding to a NFGSS, and let $I, J \in D(\bar{I})$ be information sets in the perfect recall refinement of $\bar{I}$. Then for any fixed strategy profile in the imperfect recall game, the counterfactual regrets are proportional in I and J:*

$$r(I,a) = \frac{\pi_{ci}(h_i^I)}{\pi_{ci}(h_i^J)} r(J,a) \tag{12}$$

Proof. Define Z_i to be the set of partial histories for player i that occur in any terminal history: $Z_i = \{h_i \colon (\exists h_{-i})(h_i, h_{-i}) \in Z\}$. We have already defined Z_I to be the set of terminal histories which can be reached from I; we can extend this definition to player partial histories as $Z_I^i = \{z_i \colon (\exists z_{-i} \in Z_{-i})((z_i, z_{-i}) \in Z \wedge (\exists h \in I) h \sqsubseteq (z_i, z_{-i}))\}$ where $i = P(I)$. Because each player never observes the actions of the other player, there is a clear bijection between Z_I and $Z_I^i \times Z_{-i}$.

Because I is from a perfect recall abstraction, $z_i[I] = h_i^I$ for any $z_i \in Z_I^i$, and thus we can decompose $z_i \in Z_I^i$ as $z_i = h_i^I z_i[I-]$. We define a bijection $\phi \colon Z_I^i \to Z_J^i$ as $\phi(z_i) = h_i^J z_i[I-]$. We know $\phi(z_i) \in Z_J$ because $\bar{I}$ corresponds to an MDP state for player i in the NFGSS, and thus the same series of actions must be possible from both I and J. The function is clearly a bijection because it can be inverted by reversing the role of I and J.

[2] We know $z[I]$ and $z_i[I]$ are unique because the EFG is not absent-minded; no prefix of a history can be in the same information set as the history.

We now analyze the counterfactual regret in I.

$$
\begin{aligned}
r_i(I,a) &= \sum_{z\in Z_I} \pi_{-i}(z[I])\pi(z[I]a,z)u(z) - \sum_{z\in Z_I} \pi_{-i}(z[I])\pi(z[I],z)u(z) \\
&= \sum_{z\in Z_I} \pi_{-i+c}(z)\pi_{ci}(z[I])\left(\pi_{i+c}(z[I]a,z) - \pi_{i+c}(z[I],z)\right)u(z) \\
&= \sum_{z_i\in Z_I^i}\sum_{z_{-i}\in Z_{-i}} \pi_{-i+c}(z_{-i})\pi_{ci}(z_i[I])\left(\pi_{i+c}(z_i[I]a,z_i) - \pi_{i+c}(z_i[I],z_i)\right)u(z_i,z_{-i}) \\
&\quad \text{because there is a bijection between } Z_I \text{ and } Z_I^i \times Z_{-i} \\
&= \sum_{z_i\in Z_I^i}\sum_{z_{-i}\in Z_{-i}} \pi_{-i+c}(z_{-i})\pi_{ci}(h_i^I)\left(\pi_{i+c}(h_i^I a,z_i) - \pi_{i+c}(h_i^I,z_i)\right)\left(u(h_i^I,z_{-i}) + u(z_i[I-],z_{-i})\right) \\
&= \sum_{z_{-i}\in Z_{-i}} \pi_{-i+c}(z_{-i})\pi_{ci}(h_i^I)u(h_i^I,z_{-i})\left(\sum_{z_i\in Z_I^i}\pi_{i+c}(h_i^I a,z_i) - \sum_{z_i\in Z_I^i}\pi_{i+c}(h_i^I,z_i)\right) \\
&\quad + \sum_{z_i\in Z_I^i}\sum_{z_{-i}\in Z_{-i}} \pi_{-i+c}(z_{-i})\pi_{ci}(h_i^I)\left(\pi_{i+c}(h_i^I a,z_i) - \pi_{i+c}(h_i^I,z_i)\right)u(z_i[I-],z_{-i}) \\
&= \sum_{z_i\in Z_I^i}\sum_{z_{-i}\in Z_{-i}} \pi_{-i+c}(z_{-i})\pi_{ci}(h_i^I)\left(\pi_{i+c}(h_i^I a,z_i) - \pi_{i+c}(h_i^I,z_i)\right)u(z_i[I-],z_{-i}) \\
&= \frac{\pi_{ci}(h_i^I)}{\pi_{ci}(h_i^J)}\sum_{z_i\in Z_I^i}\sum_{z_{-i}\in Z_{-i}} \pi_{-i+c}(z_{-i})\pi_{ci}(h_i^J)\left(\pi_{i+c}(h_i^J a,\phi(z_i)) - \pi_{i+c}(h_i^J,\phi(z_i))\right)u(\phi(z_i)[J-],z_{-i}) \\
&\quad \text{because the sequence of actions from } I \text{ to } z_i \text{ is the same as the sequence of actions from } J \text{ to } \phi(z_i) \\
&= \frac{\pi_{ci}(h_i^I)}{\pi_{ci}(h_i^J)}\sum_{z_i\in Z_J^i}\sum_{z_{-i}\in Z_{-i}} \pi_{-i+c}(z_{-i})\pi_{ci}(h_i^J)\left(\pi_{i+c}(h_i^J a,z_i) - \pi_{i+c}(h_i^J,z_i)\right)u(z_i[J-],z_{-i}) \\
&\quad \text{because } \phi \text{ is a bijection} \\
&= \frac{\pi_{ci}(h_i^I)}{\pi_{ci}(h_i^J)} r(J,a)
\end{aligned}
$$

□

We have shown that the regret in each perfect recall information set is proportional to the regret in any information set that it is combined with in the imperfect recall game. Thus if we minimize the regret in any one of these information sets, we necessarily minimize regret in each of the others. All that remains to show is that minimizing regret in the imperfect recall game achieves this goal.

The *overall average regret* for player i of a series of strategy profiles $\sigma^1, ..., \sigma^T$ in a perfect recall game Γ' is

$$\frac{1}{T}R_i^T = \frac{1}{T}\max_{\sigma_i^*\in\Sigma_i}\sum_{t=1}^{T}\left(u_i(\sigma_i^*,\sigma_{-i}^t) - u_i(\sigma^t)\right) \tag{13}$$

This is a measure of how much player i would have rather played a fixed strategy against the actual series of opponent strategies.

Theorem 5. *Let $\mathcal{I}_i$ be the information sets for player i in an EFG Γ corresponding to an NFGSS and let $\mathcal{I}'_i$ be the information sets in the corresponding perfect recall refinement Γ'. If CFR$^+$ is used to select a series of strategies in Γ, then we can bound the average regret of these strategies in Γ' for each player i:*

$$\frac{1}{T}R_i^T \le \frac{\Delta_i|\mathcal{I}_i|\sqrt{|A_i|}}{\sqrt{T}} \tag{14}$$

Proof. We begin by showing that counterfactual regrets in an information sets $\bar{I}$ in the imperfect recall game are sums of counterfactual regrets over $D(\bar{I})$ in the refinement, and thus we can use Lemma 4 to show that the regrets are proportional. Fix

some $\bar{I} \in \mathcal{I}_i$ and $I \in D(\bar{I})$.

$$\begin{aligned}
r_i(\bar{I}, a) &= \sum_{z \in Z_{\bar{I}}} \pi_{-i}(z[\bar{I}])\pi(z[\bar{I}]a, z)u(z) - \sum_{z \in Z_{\bar{I}}} \pi_{-i}(z[\bar{I}])\pi(z[\bar{I}], z)u(z) \\
&= \sum_{J \in D(\bar{I})} \left(\sum_{z \in Z_J} \pi_{-i}(z[J])\pi(z[J]a, z)u(z) - \sum_{z \in Z_J} \pi_{-i}(z[J])\pi(z[J], z)u(z) \right) \\
&= \sum_{J \in D(\bar{I})} r_i(J, a) \\
&= \sum_{J \in D(\bar{I})} \frac{\pi_{ci}(h_i^J)}{\pi_{ci}(h_i^I)} r_i(I, a)
\end{aligned}$$

We can now bound the average regret in Γ'.

$$\begin{aligned}
R_i^T &\leq \sum_{I \in \mathcal{I'}_i} \max_a \left(\sum_{t=1}^{T} r_i^t(I, a) \right)^+ \\
&\quad \text{by Theorem 3 of (Zinkevich et al. 2008)} \\
&= \sum_{I \in \mathcal{I}_i} \sum_{I \in D(\bar{I})} \max_a \left(\sum_{t=1}^{T} r_i^t(I, a) \right)^+ \\
&= \sum_{I \in \mathcal{I}_i} \sum_{I \in D(\bar{I})} \max_a \left(\sum_{t=1}^{T} \frac{\pi_{ci}(h_i^I)}{\sum_{J \in D(\bar{I})} \pi_{ci}(h_i^J)} r_i^t(\bar{I}, a) \right)^+ \\
&= \sum_{I \in \mathcal{I}_i} \max_a \left(\sum_{t=1}^{T} r_i^t(\bar{I}, a) \right)^+ \sum_{I \in D(\bar{I})} \frac{\pi_{ci}(h_i^I)}{\sum_{J \in D(\bar{I})} \pi_{ci}(h_i^J)} \\
&= \sum_{I \in \mathcal{I}_i} \max_a \left(\sum_{t=1}^{T} r_i^t(\bar{I}, a) \right)^+ \\
&\leq \sum_{I \in \mathcal{I}_i} \max_a Q(\bar{I}, a) \\
&\quad \text{by Lemma 1 of (Tammelin et al. 2015)} \\
\therefore \frac{1}{T} R_i^T &\leq \frac{\Delta_i |\mathcal{I}_i| \sqrt{|A_i|}}{\sqrt{T}} \\
&\quad \text{by Lemma 2 of (Tammelin et al. 2015)}
\end{aligned}$$

□

Thus minimizing regret in each information set of the imperfect recall game minimizes overall average regret in the perfect recall refinement for each player. It is well known (see, e.g., (Zinkevich et al. 2008)) that when average regret is minimized by a series of strategy profiles, the profiles must converge to a Nash equilibrium. Thus we can use CFR$^+$ in the EFG to find a Nash equilibrium for the NFGSS. Consequently, NFGSS-CFR$^+$ is guaranteed to find a Nash equilibrium in any NFGSS.