Efficient Deviation Types and Learning for Hindsight Rationality in Extensive-Form Games

Dustin Morrill [1] **Ryan D'Orazio** [2] **Marc Lanctot** [3] **James R. Wright** [1] **Michael Bowling** [1 3] **Amy R. Greenwald** [4]

Abstract

Hindsight rationality is an approach to playing general-sum games that prescribes no-regret learning dynamics for individual agents with respect to a set of deviations, and further describes jointly rational behavior among multiple agents with mediated equilibria. To develop hindsight rational learning in sequential decision-making settings, we formalize behavioral deviations as a general class of deviations that respect the structure of extensive-form games. Integrating the idea of time selection into counterfactual regret minimization (CFR), we introduce the extensive-form regret minimization (EFR) algorithm that achieves hindsight rationality for any given set of behavioral deviations with computation that scales closely with the complexity of the set. We identify behavioral deviation subsets, the partial sequence deviation types, that subsume previously studied types and lead to efficient EFR instances in games with moderate lengths. In addition, we present a thorough empirical analysis of EFR instantiated with different deviation types in benchmark games, where we find that stronger types typically induce better performance.

1. Introduction

We seek more effective algorithms for playing multi-player, general-sum extensive-form games (EFGs). The hindsight rationality framework (Morrill et al., 2021) suggests a game playing approach that prescribes no-regret dynamics and describes jointly rational behavior with mediated equilibria (Aumann, 1974). Rationality within this framework is measured by regret in hindsight relative to strategy transformations, also called deviations, rather than as prospective optimality with respect to beliefs. Each deviation transforms the learner's behavior into a competitor that the learner must surpass, so a richer set of deviations pushes the learner to perform better.

While larger deviation sets containing more sophisticated deviations produce stronger competitors, they also tend to raise computational and storage requirements. For example, there is one external deviation (constant strategy transformation) for each strategy in a set of n, but there are n^2 internal deviations (Foster & Vohra, 1999) that transform one particular strategy into another, and the latter is fundamentally stronger. Though even achieving hindsight rationality with respect to external deviations appears intractable because the number of strategies in an EFG grows exponentially with the size of the game.

The counterfactual regret minimization (CFR) (Zinkevich et al., 2007) algorithm makes use of the EFG structure to be efficiently hindsight rational for external deviations. Modifications to CFR by Celli et al. (2020) and Morrill et al. (2021) are efficiently hindsight rational for other types of deviations as well. We generalize these algorithms as *extensive-form regret minimization* (*EFR*), a simple and extensible algorithm that is hindsight rational for any given deviation set where each deviation can be decomposed into action transformations at each decision point. It is generally intractable to run EFR with all such *behavioral deviations* so we identify four subsets that lead to efficient EFR instantiations that are hindsight rational for all previously studied tractable deviation types (external, causal (Forges & von Stengel, 2002; von Stengel & Forges, 2008; Gordon et al., 2008; Dudík & Gordon, 2009; Farina et al., 2020a), action (von Stengel & Forges, 2008; Morrill et al., 2021), and counterfactual (Morrill et al., 2021)) simultaneously. We provide EFR instantiations and sublinear regret bounds for each of these new *partial sequence deviation* types.

We present a thorough empirical analysis of EFR's performance with different deviation types in benchmark games from OpenSpiel (Lanctot et al., 2019). Stronger deviation types typically lead to better performance, and EFR with the strongest type of partial sequence deviation often per-

[1]Department of Computing Science, University of Alberta; Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada [2]DIRO, Université de Montréal; Mila, Montréal, Québec, Canada [3]DeepMind [4]Computer Science Department, Brown University, Providence, Rhode Island, United States. Correspondence to: Dustin Morrill <morrill@ualberta.ca>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

forms nearly as well as that with all behavioral deviations, in games where the latter is tractable.

2. Background

This work will continuously reference decision making from both the macroscopic, *normal-form* view, and the microscopic, *extensive-form* view. We first describe the normal-form view, which models simultaneous decision making, before extending it with the extensive-form view, which models sequential decision making.

2.1. The Normal-Form View

At the macro-scale, players in a game choose *strategies* that jointly determine the *utility* for each player. We assume a bounded utility function $u_i : \mathcal{Z} \to [-U, U]$ for each player i on a finite set of outcomes, $\mathcal{Z}$. Each player has a finite set of *pure strategies*, $s_i \in S_i$, describing their decision space. A set of results for entirely random events, *e.g.*, die rolls, is denoted S_c. A *pure strategy profile*, $s \in S = S_c \times \bigtimes_{i=1}^{N} S_i$, is an assignment of pure strategies to each player, and each strategy profile corresponds to a unique outcome $z \in \mathcal{Z}$ determined by the *reach function* $P(z; s) \in \{0, 1\}$.

A *mixed strategy*, $\pi_i \in \Pi_i = \Delta^{|S_i|}$, is a probability distribution over pure strategies. In general, we assume that strategies are mixed where pure strategies are point masses. The probability of a chance outcome, $s_c \in S_c$, is determined by the "chance player" who plays the fixed strategy π_c. A *mixed strategy profile*, $\pi \in \Pi = \{\pi_c\} \times \bigtimes_{i=1}^{N} \Pi_i$, is an assignment of mixed strategies to each player. The probability of sampling a pure strategy profile, s, is the product of sampling each pure strategy individually, *i.e.*, $\pi(s) = \pi_c(s_c) \prod_{i=1}^{N} \pi_i(s_i)$. For convenience, we denote the tuple of mixed strategies for all players except i as $\pi_{-i} \in \Pi_{-i} = \{\pi_c\} \times \bigtimes_{j \neq i} \Pi_j$. We overload the reach function to represent the probability of realizing outcome z according to mixed profile π, *i.e.*, $P(z; \pi) = \mathbb{E}_{s \sim \pi}[P(z; s)]$, allowing us to express player i's expected utility as $u_i(\pi_i, \pi_{-i}) \doteq u_i(\pi) = \mathbb{E}_{z \sim P(\cdot; \pi)}[u_i(z)]$.

The *regret* for playing strategy π_i instead of deviating to an alternative strategy π_i' is their difference in expected utility $u_i(\pi_i', \pi_{-i}) - u_i(\pi)$. We construct alternative strategies by transforming π_i. Let $\Phi_{\mathcal{X}}^{\text{SW}} = \{\phi : \mathcal{X} \to \mathcal{X}\}$ be the set of transformations to and from a given finite set $\mathcal{X}$. The pure strategy transformations in $\Phi_{S_i}^{\text{SW}}$ are known as *swap deviations* (Greenwald et al., 2003). Given a mixed strategy π_i, the transformed mixed strategy under deviation $\phi \in \Phi_{S_i}^{\text{SW}}$ is the pushforward measure of π_i, denoted as $\phi(\pi_i)$ and defined by $[\phi\pi_i](s_i') = \sum_{s_i \in \phi^{-1}(s_i')} \pi_i(s_i)$ for all $s_i' \in S_i$, where $\phi^{-1} : s_i' \mapsto \{s_i \mid \phi(s_i) = s_i'\}$ is the pre-image of ϕ. The regret for playing strategy π_i instead of deviating according to ϕ is then $\rho(\phi; \pi) = u_i(\phi(\pi_i), \pi_{-i}) - u_i(\pi)$.

In an online learning setting, a learner repeatedly plays a game with unknown, dynamic, possibly adversarial players. On each round $1 \leq t \leq T$, the learner who acts as player i chooses a strategy, π_i^t, simultaneously with the other players who in aggregate choose π_{-i}^t. The learner is evaluated on their strategies, $(\pi_i^t)_{t=1}^T$, against a deviation, ϕ, with the cumulative regret $\rho^{1:T}(\phi) \doteq \sum_{t=1}^{T} \rho(\phi; \pi^t)$. A learner is rational in hindsight with respect to a set of deviations, $\Phi \subseteq \Phi_{S_i}^{\text{SW}}$, if the maximum positive regret, $\max_{\phi \in \Phi} \left(\rho^{1:T}(\phi)\right)^+$ where $\cdot^+ = \max\{\cdot, 0\}$, is zero. A *no-regret* or *hindsight rational* algorithm ensures that average maximum positive regret vanishes as $T \to \infty$.

The *empirical distribution of play*, $\mu^T \in \Delta^{|S|}$, is the distribution that summarizes online correlated play, *i.e.*, $\mu^T(s) = \frac{1}{T} \sum_{t=1}^{T} \pi^t(s)$, for all pure strategy profiles, s. The distribution μ^T can be viewed as a source of "strategy recommendations" distributed to players by a neutral "mediator". The incentive for player i to deviate from the mediator's recommendations, sampled from μ^T, to behavior chosen by ϕ is then player i's average regret $\mathbb{E}_{s \sim \mu^T}[\rho(\phi; s)] = \frac{1}{T} \rho^{1:T}(\phi)$. Jointly hindsight rational play converges toward a *mediated equilibrium* (Aumann, 1974) where no player has an incentive to deviate from the mediator's recommendations, since hindsight rational players ensure that their average regret vanishes.

The deviation set influences what behaviors are considered rational and the difficulty of ensuring hindsight rationality. For example, the *external deviations*, $\Phi_{S_i}^{\text{EX}} = \{\phi^{\to s_i} : s_i \mapsto s_i\}_{s_i \in S_i}$, are the constant strategy transformations, a set which is generally limited in strategic power compared to the full set of swap deviations. However, it is generally intractable to directly minimize regret with respect to the external deviations in sequential decision-making settings because the set of pure strategies grows exponentially with the number of decision points.

2.2. The Extensive-Form View

Actions, histories, and information sets. An *extensive-form game* (*EFG*) models player behavior as a sequence of decisions. Outcomes, here called *terminal histories*, are constructed incrementally from the empty history, $\varnothing \in \mathcal{H}$. At any history h, one player determined by the *player function* $\mathcal{P} : \mathcal{H} \setminus \mathcal{Z} \to \{1, \ldots, N\} \cup \{c\}$ plays an *action*, $a \in \mathcal{A}(h)$, from a finite set, which advances the current history to ha. We write $h \sqsubset ha$ to denote that h is a predecessor of ha. We denote the maximum number of actions at any history as $n_{\mathcal{A}}$.

Histories are partitioned into *information sets* to model imperfect information, *e.g.*, private cards. The player to act in each history $h \in I$ in information set $I \in \mathcal{I}$ must do so knowing only that the current history is in I. The unique

information set that contains a given history is returned by $\mathbb{I}$ ("blackboard I") and an arbitrary history of a given information set is returned by $\mathbb{h}$ ("blackboard h"). Naturally, the action sets of each history within an information set must coincide, so we overload $\mathcal{A}(I) = \mathcal{A}(\mathbb{h}(I))$.

Each player i has their own *information partition*, denoted $\mathcal{I}_i$. We restrict ourselves to *perfect-recall* information partitions that ensure players never forget the information sets they encounter during play and their information set transition graphs are forests (not trees since other players may act first). We write $I \prec I'$ to denote that I is a predecessor of I' and $\mathbb{p}(I')$ to reference the unique parent (immediate predecessor) of I'. Let d_I be the number of I's predecessors representing I's depth and $d_* = \max_{I \in \mathcal{I}_i} d_I$ be the depth of player i's deepest information set.We use $a_h^{\rightarrow I'}$ or $a_I^{\rightarrow I'}$ to reference the unique action required to play from $h \in I$ to a successor history in $I' \succ I$.

Strategies and reach probabilities. From the extensive-form view, a pure strategy is an assignment of actions to each of a player's information sets, *i.e.*, $s_i(I)$ is the action that player i plays in information set I according to pure strategy s_i. A natural generalization is to randomize at each information set, leading to the notion of a *behavioral strategy* (Kuhn, 1953). A behavioral strategy is defined by an assignment of *immediate strategies*, $\pi_i(I) \in \Delta^{|\mathcal{A}(I)|}$, to each of player i's information sets, where $\pi_i(a \mid I)$ is the probability that i plays action a in I. Perfect recall ensures *realization equivalence* between the set of mixed and behavioral strategies where there is always a behavioral strategy that applies the same weight to each terminal history as a mixed strategy and *vice-versa*. Thus, we treat mixed and behavioral strategies (and by extension pure strategies) as interchangeable representations.

Since histories are action sequences and behavioral strategies define conditional action probabilities, the probability of reaching a history under a profile is the joint action probability that follows from the chain rule of probability. We overload $P(h; \pi)$ to return the probability of a non-terminal history h. Furthermore, we can look at the joint probability of actions played by just one player or a subset of players, denoted, for example, as $P(h; \pi_i)$ or $P(h; \pi_{-i})$. We can use this and perfect recall to define the probability that player i plays to their information set $I \in \mathcal{I}_i$ as $P(\mathbb{h}(I); \pi_i)$. Additionally, we can exclude actions taken before some initial history h to get the probability of playing from h to history h', written as $P(h, h'; \cdot)$, where it is 1 if $h = h'$ and 0 if $h \not\sqsubseteq h'$.

2.3. Extensive-Form Correlated Equilibrium

Extensive-form correlated equilibrium (*EFCE*) is defined by Definition 2.2 of von Stengel & Forges (2008) as a mediated equilibrium with respect to deviations that are constructed according to the play of a *deviation player*. At the beginning of the game, the mediator samples a pure strategy profile (strategy recommendations), s, and the game plays out according to this profile until it is player i's turn to act. Player i's decision at this information set I is determined by the deviation player who observes $s_i(I)$, which is the action recommended to player i by the mediator at I, and then chooses an action by either following this recommendation or deviating to a different action. After choosing an action and waiting for the other players to move according to their recommended strategies, the deviation player arrives at i's next information set. Knowing the actions that were previously recommended to i, they again choose to follow the next recommendation or to deviate from it. This process continues until the game ends.

The number of different states that the deviation player's memory could be in upon reaching information set I at depth d_I is $n_{\mathcal{A}}^{d_I}$ corresponding to the number of action combinations across I's predecessors. One way to avoid this exponential growth is to assume that recommended strategies are *reduced*, that is, they do not assign actions to information sets that could not be reached according to actions assigned to previous information sets. Thus, the action recommendation that the deviation player would normally observe after a previous deviation does not exist to observe. This assumption effectively forces the deviation player to behave according to an *informed causal deviation* (Gordon et al., 2008; Dudík & Gordon, 2009) defined by a "trigger" action and information set pair, along with a strategy to play after triggering, and the number of possible memory states grows linearly with depth. Defining EFCE as a mediated equilibrium with respect to informed causal deviations allows them to be computed efficiently, which has led to this becoming the conventional definition of EFCE.

3. Behavioral Deviations

Instead of achieving tractability by limiting the amount of information present in strategy recommendations, what if we intentionally hide information from the deviation player? At each information set, I, we now provide the deviation player with three options: (i) follow the action recommendation at information set I, $s_i(I)$, sight unseen, (ii) choose a new action without ever seeing $s_i(I)$, or (iii) observe $s_i(I)$ and then choose an action.

If $\mathcal{A}_* = \bigcup_{I \in \mathcal{I}_i} \mathcal{A}(I)$ is the union of player i's action sets, then we can describe the deviation player's memory, $g \in G_i \subseteq (\{*\} \cup \mathcal{A}_*)^{d_*}$, as a string that begins empty and gains a character after each of player i's actions. The recommendation, $s_i(I)$, at information set I where option one or three is chosen must be revealed to the deviation player, either as a consequence of play (option one) or as a prerequisite (option three), thus resulting in the next memory

state $gs_i(I)$. Otherwise, the next memory state is formed by appending the "$*$" character to indicate that $s_i(I)$ remains hidden. Limiting the options available to the deviation player thus limits the number of memory states that they can realize. Given a memory state g, there is only one realizable child memory state at the next information set if the deviation player is allowed either option one or two, or two memory states if both options one and two are allowed. If all three options are allowed, the number of realizable child memory states at the next information set is equal to the number of actions at the current information set plus one.

Formally, these three options are executed at each information set I with an *action transformation*, $\phi_I : \mathcal{A}(I) \to \mathcal{A}(I)$, chosen from one of three sets: (i) the singleton containing the *identity transformation*, $\{\phi^1 : a \mapsto a\}$, (ii) the external transformations, $\Phi^{\text{EX}}_{\mathcal{A}(I)}$, or (iii) the *internal transformations* (Foster & Vohra, 1999)

$$\Phi^{\text{IN}}_{\mathcal{A}(I)} = \left\{ \phi^{a^! \to a} \quad : a \mapsto \begin{cases} a & \text{if } a = a^! \\ a & \text{o.w.} \end{cases} \right\}_{a^!, a \ \in \mathcal{A}(I)}.$$

While internal transformations can only swap one action with another, there is no loss in generality because every multi-action swap can be a represented as the combination of single swaps (Dudík & Gordon, 2009; Greenwald et al., 2003). Thus, any strategy sequence that can be improved upon by a swap deviation can also be improved upon by at least one internal deviation.

A complete assignment of action transformations to each information set and realizable memory state represents a complete strategy for the deviation player. We call such an assignment a *behavioral deviation* in analogy with behavioral strategies and denote them as $\Phi^{\text{IN}}_{\mathcal{I}_i}$ since the behavioral deviations are a natural analog of the internal transformations in EFGs.

All previously described EFG deviation types can be represented as sets of behavioral deviations:

von Stengel & Forges (2008)'s deviations. Any strategy that von Stengel & Forges (2008)'s deviation player could employ[1] is an assignment of internal transformations to every information set and memory state so the set of all such behavioral deviations represents all possible deviation player strategies. A mediated equilibrium with respect to the behavioral deviations could thus perhaps be called a "full strategy EFCE", though "behavioral correlated equilibrium" may lead to less confusion with the conventional EFCE definition.

Causal deviations. An informed causal deviation is defined by trigger information set $I^!$, trigger action $a^!$, and strategy π'_i. The following behavioral deviation reproduces any such deviation: assign (i) the internal transformation $\phi^{a^! \to a}$ to the sole memory state at $I^!$, (ii) external transformations to all successors $I' \succ I^!$ where $a^!$ is in the deviation player's memory to reproduce π'_i, and (iii) identity transformations to every other information set and memory state. The analogous *blind causal deviation* (Farina et al., 2020a) always triggers in $I^!$, which is reproduced with the same behavioral deviation except that the external transformation $\phi^{\to \pi'_i(I^!)}$ is assigned to $I^!$.

Action deviations. An *action deviation* (von Stengel & Forges, 2008) modifies the immediate strategy at $I^!$, $\pi_i(I^!)$, only, either conditioning on $\pi_i(I^!)$ (an informed action deviation) or not (a blind action deviation (Morrill et al., 2021)), so any such deviation is reproduced by assigning either an internal or external transformation to the sole memory state at $I^!$, respectively, and identity transformations elsewhere.

Counterfactual deviations. A *counterfactual deviation* (Morrill et al., 2021) plays to reach a given "target" information set, I , and transforms the immediate strategy there so any such deviation is reproduced by assigning (i) external transformations to all of the information sets leading up to I , (ii) an external or internal transformation to the sole memory state at I (for the blind and informed variant, respectively), and (iii) identity transformations elsewhere.

Phrasing these deviation types as behavioral deviations allows us to identify complexity differences between these deviation types by counting the number of realizable memory states they admit. Across all action or counterfactual deviations, there is always exactly one memory state at each information set to which a non-identity transformation is assigned. Thus, a hindsight rational algorithm need only ensure its strategy cannot be improved by applying a single action transformation at each information set. Under the causal deviations, in contrast, the number of memory states realizable at information set I is at least the number of I's predecessors since there is at least one causal deviation that triggers at each of them and plays to I. This makes causal deviations more costly to compete with and gives them strategic power, though notably not enough to subsume either the action or counterfactual deviations (Morrill et al., 2021). Are there sets of behavioral deviations that subsume the causal, action, and counterfactual deviations without being much more costly than the causal deviations?

4. Partial Sequence Deviations

Notice that the causal, action, and counterfactual deviations are composed of contiguous blocks of the same type of action transformation. We can therefore understand these deviations as having distinct phases. The *correlation phase*

[1] Where the deviation player makes only single-action swaps, which, again, is a simplification made without a loss in generality (Dudík & Gordon, 2009; Greenwald et al., 2003).

is an initial sequence of identity transformations, where "correlation" references the fact that the identity transformation preserves any correlation that player i's behavior has with those of the other players. There are causal and action deviations with a correlation phase, but no counterfactual deviation exhibits such behavior. All of these deviation types permit a *de-correlation phase* that modifies the input strategy with external transformations, breaking correlation. Finally, the *re-correlation phase* is where identity transformations follow a de-correlation phase, but it is only present in action and counterfactual deviations. The informed variant of each deviation type separates these phases with a single internal transformation, which both modifies the strategy and preserves correlation. The action deviation type is the only one that permits all three phases, but the de-correlation phase is limited to a single action transformation.

Why not permit all three phases at arbitrary lengths to subsume the causal, action, and counterfactual deviations? We now introduce four types of *partial sequence deviations* based on exactly this idea, where each phase spans a "partial sequence" through the game.

The *blind partial sequence* (*BPS*) deviation has all three phases and lacks any internal transformations. Notice that there are at most $d_* n_{\mathcal{A}} |\mathcal{I}_i|$ BPS deviations and yet the behavior produced by any blind causal deviation given a pure strategy profile can be reproduced exactly by a BPS deviation. Thus, the benefit of a blind causal deviation cannot be more than the sum of the positive part of the benefit from each individual BPS deviation needed to reproduce the blind causal deviation's behavior under different strategy profiles. Since there are $\mathcal{O}(n_{\mathcal{A}}^{|\mathcal{I}_i|}|\mathcal{I}_i|)$ blind causal deviations, the BPS deviations capture the same strategic power with an exponential reduction in complexity. Even better, the set of BPS deviations includes the sets of blind action and blind counterfactual deviations. The empirical distribution of play of learners that are hindsight rational for BPS deviations thus converge towards what we could call a *BPS correlated equilibrium* in the intersection of the sets of extensive-form coarse-correlated equilibrium (EFCCE) (Farina et al., 2020a), agent-form coarse-correlated equilibrium (AFCCE) (Morrill et al., 2021), and counterfactual coarse-correlated equilibrium (CFCCE) (Morrill et al., 2021).

In general, re-correlation is strategically useful (Morrill et al., 2021) and adding it to a deviation type *decreases* its complexity! While this observation may be new in its generality, Zinkevich et al. (2007) implicitly uses this property of deviations in EFGs to design *counterfactual regret minimization* (*CFR*) so that it efficiently minimizes regret for the external deviations. This is because CFR minimizes regret for blind counterfactual deviations (Morrill et al., 2021), which are exactly external deviations augmented with re-correlation in the same way that BPS deviations are the blind causal deviations augmented with re-correlation.

There are three versions of informed partial sequence deviations due to the asymmetry between informed causal and informed counterfactual deviations. A *causal partial sequence* (*CSPS*) deviation uses an internal transformation at the end of the correlation phase while a *counterfactual partial sequence* (*CFPS*) deviation uses an internal transformation at the start of the re-correlation phase. A *twice informed partial sequence* (*TIPS*) deviation uses internal transformations at both positions, making it the strongest of our partial sequence deviation types. CSPS subsumes the informed causal deviations but represents an exponentially smaller set because CSPS allows re-correlation. And TIPS achieves our initial goal as it subsumes the informed causal, informed action, and informed counterfactual deviations at the cost of an $n_{\mathcal{A}}$ factor compared to CSPS or CFPS. Each type of informed partial sequence deviation corresponds to a new mediated equilibrium concept in the intersection of previously studied equilibrium concepts.

Table C.1 in Appendix C gives a formal definition of each deviation type derived from behavioral deviations and Figure 1 gives a visualization of each type along with their relationships. The number of deviations contained within each deviation type is listed in Table 1.

5. Extensive-Form Regret Minimization

We now develop *extensive-form regret minimization* (*EFR*), a general and extensible algorithm that is hindsight rational for any given set of behavioral deviations. Its computational requirements and regret bound scale closely with the number of realizable memory states.

5.1. CFR and Previous Derivatives

CFR is based on evaluating actions with their *counterfactual value* (Zinkevich et al., 2007). Given a strategy profile, π, the counterfactual value for taking a in information set I is the expected utility for player i, assuming they play to reach I before playing π_i thereafter and that the other players play according to π_{-i} throughout, *i.e.*,

$$v_I(a;\pi) = \sum_{\substack{h \in I, \\ z \in \mathcal{Z}}} P(h;\pi_{-i}) \underbrace{P(ha, z;\pi) u_i(z)}_{\text{Future value given } ha.}.$$

The learner's performance is then measured at each information set in isolation according to *immediate counterfactual regret*, which is the extra counterfactual value achieved by choosing a given action instead of following π_i at I, *i.e.*, $\rho_I^{\text{CF}}(a;\pi) = v_I(a;\pi) - \mathbb{E}_{a' \sim \pi_i(I)} v_I(a';\pi)$.

Iterating on Zinkevich et al. (2007)'s approach to learning in EFGs, Celli et al. (2020) define *laminar subtree trigger regret* (*immediate trigger regret* in our terminology),

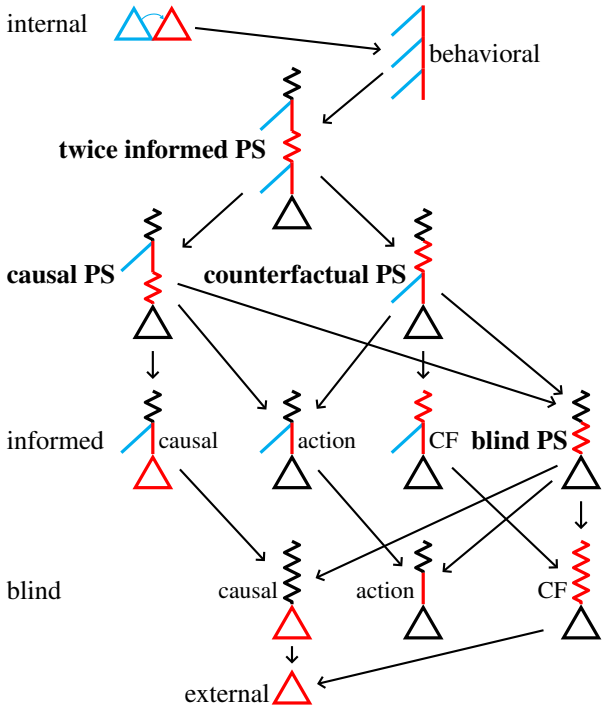

Figure 1. A summary of the deviation landscape in EFGs. Each pictogram is an abstract representation of a prototypical deviation. Games play out from top to bottom. Straight lines represent action transformations, zigzags are transformation sequences, and triangles are transformations of entire decision trees. Identity transformations are colored black; internal transformations have a cyan component representing the trigger action or strategy and a red component representing the target action or strategy; and external transformations only have a red component. Arrows denote ordering from a stronger to a weaker deviation type (and therefore a subset to superset equilibrium relationship).

as the regret under counterfactual values weighted by the probability that player i plays to a given predecessor and plays a particular action there. Their *ICFR* modification of *pure CFR* (Gibson, 2014)[2] is hindsight rational for informed causal deviations[3]. Morrill et al. (2021) also observe that simply weighting the counterfactual regret at I by the probability that player i plays to I modifies CFR so that it is hindsight rational for blind action deviations. We derive EFR by generalizing the idea of changing CFR's learning behavior by weighting counterfactual regrets.

5.2. Time Selection

The key insight that leads to EFR is that each of the deviation player's memory states corresponds to a different weighting function, which reduces the problem of minimizing immediate regret with respect to all weightings simultaneously to *time selection regret minimization* (Blum & Mansour, 2007). In a time selection problem, there is a finite set of $M(\phi)$ time selection functions, $W(\phi) = \{t \mapsto w_j^t \in [0,1]\}_{j=1}^{M(\phi)}$, for each deviation $\phi \in \Phi \subseteq \Phi_{S_i}^{\text{SW}}$ that maps the round t to a weight. The regret with respect to deviation ϕ and time selection function $w \in W(\phi)$ after T rounds is $\rho^{1:T}(\phi, w) \doteq \sum_{t=1}^{T} w^t \rho(\phi; \pi^t)$. The goal is to ensure that each of these regrets grow sublinearly, which can be accomplished by simply treating each (ϕ, w)-pair as a separate transformation (here called an *expert*) and applying a no-regret algorithm[4].

We introduce a (Φ, f)-regret matching (Hart & Mas-Colell,

Table 1. A rough accounting of (i) realizable memory states, (ii) action transformations, and (iii) the total number of deviations showing dominant terms. Columns (i) and (ii) are with respect to a single information set.

type	(i)	(ii)	(iii)
internal	N/A	N/A	$n_{\mathcal{A}}^{2\lvert\mathcal{I}_i\rvert}$
behavioral	$n_{\mathcal{A}}^{d_*}$	$n_{\mathcal{A}}^2$	$n_{\mathcal{A}}^{d_*+2}\lvert\mathcal{I}_i\rvert^{\dagger}$
TIPS	$d_* n_{\mathcal{A}}$	$n_{\mathcal{A}}^2$	$d_* n_{\mathcal{A}}^3 \lvert\mathcal{I}_i\rvert$
CSPS	$d_* n_{\mathcal{A}}$	$n_{\mathcal{A}}{}^{\ddagger}$	$d_* n_{\mathcal{A}}^2 \lvert\mathcal{I}_i\rvert$
CFPS	d_*	$n_{\mathcal{A}}^2$	$d_* n_{\mathcal{A}}^2 \lvert\mathcal{I}_i\rvert$
BPS	d_*	$n_{\mathcal{A}}$	$d_* n_{\mathcal{A}} \lvert\mathcal{I}_i\rvert$
informed causal	d_*	N/A	$n_{\mathcal{A}}^{\lvert\mathcal{I}_i\rvert+1}\lvert\mathcal{I}_i\rvert$
informed action	1	$n_{\mathcal{A}}^2$	$n_{\mathcal{A}}^2 \lvert\mathcal{I}_i\rvert$
informed CF	1	$n_{\mathcal{A}}^2$	$n_{\mathcal{A}}^2 \lvert\mathcal{I}_i\rvert$
blind causal	d_*	N/A	$n_{\mathcal{A}}^{\lvert\mathcal{I}_i\rvert}\lvert\mathcal{I}_i\rvert$
blind action	1	$n_{\mathcal{A}}$	$n_{\mathcal{A}} \lvert\mathcal{I}_i\rvert$
blind CF	1	$n_{\mathcal{A}}$	$n_{\mathcal{A}} \lvert\mathcal{I}_i\rvert$
external	N/A	N/A	$n_{\mathcal{A}}^{\lvert\mathcal{I}_i\rvert}$

† This is the number of behavioral deviations that only assign non-identity transformations to each predecessor information set leading up to a given target information set, which are representative in the same way that, *e.g.*, BPS deviations capture all the strategic power of blind causal deviations.

‡ One memory state at each information set is associated with the set of internal transformations which contains $\mathcal{O}(n_{\mathcal{A}}^2)$ transformations, but this is dominated by the number of external transformations associated with every other memory state in non-root information sets.

[2]Pure CFR purifies the learner's strategy on each round by sampling actions at each information set.

[3]Actually, ICFR is hindsight rational for CSPS deviations as well, but of course this was previously not understood.

[4] The *regret matching++* algorithm (Kash et al., 2020) could ostensibly be used to minimize regret with respect to all time selection functions simultaneously without using more than $|S_i|$ computation and memory, however, there is an error in the proof of the regret bound. In Appendix D, we give an example where regret matching++ suffers linear regret and we show that no algorithm can have a sublinear bound on the sum of positive instantaneous regrets.

2000; Greenwald et al., 2006) algorithm for the time selection setting with a regret bound that depends on the size of the largest time selection function set, $M^* = \max_{\phi \in \Phi} M(\phi)$.[5]

Corollary 1. *Given deviation set $\Phi \subseteq \Phi^{\text{SW}}_{S_i}$ and finite time selection sets $W(\phi) = \{w_j \in [0,1]^T\}_{j=1}^{M(\phi)}$ for each deviation $\phi \in \Phi$, $(\Phi, \cdot^+)$-regret matching chooses a strategy on each round $1 \leq t \leq T$ as the fixed point of $L^t : \pi_i \mapsto {}^1\!/_{z^t} \sum_{\phi \in \Phi} \phi(\pi_i) y^t_\phi$ or an arbitrary strategy when $z^t = 0$, where* link outputs *are generated from exact regrets $y^t_\phi = \sum_{w \in W(\phi)} w^t(\rho^{1:t-1}(\phi, w))^+$ and $z^t = \sum_{\phi \in \Phi} y^t_\phi$. This algorithm ensures that $\rho^{1:T}(\phi, w) \leq 2U\sqrt{M^* \omega(\Phi) T}$ for any deviation ϕ and time selection function w, where $\omega(\Phi) = \max_{a \in S_i} \sum_{\phi \in \Phi} \mathbb{1}\{\phi(s_i) \neq s_i\}$ is the maximal activation of Φ (Greenwald et al., 2006).*

This result is a consequence of two more general theorems presented in Appendix B, one that allows regret approximations à la D'Orazio et al. (2020) (motivating the use of function approximation) and another that allows predictions of future regret, *i.e.*, optimistic regret matching (D'Orazio & Huang, 2021). Appendix B also contains analogous results for the *regret matching*$^+$ (Tammelin, 2014; Tammelin et al., 2015) modification of regret matching.

5.3. Memory Probabilities

Just as we use the reach probability function to capture the frequency that a mixed strategy plays to reach a particular history, we define a *memory probability function*, w_ϕ, to capture the frequency that the deviation player, playing behavioral deviation ϕ, reaches information set I with memory state g given mixed recommendations, π_i. It is the product of the probabilities that π_i plays each action in g, *i.e.*, $w_\phi(I, \varnothing; \pi_i) = 1$, $w_\phi(I', ga; \pi_i) = w_\phi(I, g; \pi_i)\pi_i(a \mid I)$, and $w_\phi(I', g*; \pi_i) = w_\phi(I, g; \pi_i)$, for all $I' \succ I$. Under pure recommendations, the memory probability function expresses realizability. We overload

$$G_i(I, \phi) = \{g \in G_i \mid \exists s_i \in S_i, w_\phi(I, g; s_i) = 1\}$$

as the set of memory states that ϕ can realize in I.

5.4. EFR

We define the immediate regret of behavioral deviation ϕ at information set I and memory state g as the immediate counterfactual regret for not applying action transformation $\phi_{I,g}$ weighted by the probability of g, *i.e.*, $w_\phi(I, g; \pi_i)\rho^{\text{CF}}_I(\phi_{I,g}; \pi)$, where we generalize counterfactual regret to action transformations as $\rho^{\text{CF}}_I(\phi_{I,g}; \pi) = \mathbb{E}_{a \sim \phi_{I,g}(\pi_i(I))}[\rho^{\text{CF}}_I(a; \pi)]$. By treating $t \mapsto w_\phi(I, g; \pi^t_i)$ for each memory state g in I as a time selection function, we reduce the problem of minimizing immediate regret to time selection regret minimization.

EFR is given a set of behavioral deviations, $\Phi \subseteq \Phi^{\text{IN}}_{\mathcal{I}_i}$, and gathers all transformations at information set I across realizable memory states into $\Phi_I = \{\phi_{I,g}\}_{\phi \in \Phi,\, g \in G_i(I,\phi)}$. Each action transformation, $\phi_I \in \Phi_I$ is associated with the set of time selection functions

$$W^\Phi_I(\phi_I) = \{t \mapsto w_{\phi'}(I, g; \pi^t_i)\}_{\phi' \in \Phi,\, g \in G_i(I,\phi'),\, \phi'_{I,g} = \phi_I}$$

derived from memory probabilities. EFR then chooses its immediate strategy at I according to a time selection regret minimizer. Applying the same procedure at each information set, EFR minimizes immediate regret at all information sets and memory states simultaneously.

Hindsight rationality requires us to relate immediate regret to full regret. The full regret of behavioral deviation ϕ at information set I and memory state g, $\rho_{I,g}(\phi; \pi)$, is the expected value achieved by $\phi(\pi_i)$ from I and g minus that of π_i, weighted by the probability of g. The full regret at the start of the game on any given round is then exactly the total performance difference between ϕ and the learner. The full regret decomposes across successive information sets and memory states, *i.e.*,

$$\rho_{I,g}(\phi; \pi) = \underbrace{\rho_I(\phi_{\preceq I, \sqsubseteq g})}_{\text{Immediate regret.}} + \sum_{\substack{a' \in \mathcal{A}(I),\\ I' \in \mathcal{I}_i(I,a'),\\ b \in \{*\} \cup \mathcal{A}(I)}} \underbrace{\rho_{I',gb}(\phi; \pi)}_{\text{Full regret at successor.}},$$

where $\phi_{\preceq I, \sqsubseteq g}$ is the behavioral deviation that deploys ϕ at all $\bar{I} \preceq I$ and $\bar{g} \sqsubseteq g$ but the identity transformation otherwise. Therefore, minimizing immediate regret at every information set and memory state also minimizes full regret at every information set and memory state, including those at the start of the game. Finally, this implies that minimizing immediate regret with respect to any given set of behavioral deviations $\Phi \subseteq \Phi^{\text{IN}}_{\mathcal{I}_i}$ ensures hindsight rationality with respect to Φ. EFR's regret is bounded according to the following theorem:

Theorem 1. *Instantiate EFR for player i with exact regret matching and a set of behavioral deviations $\Phi \subseteq \Phi^{\text{IN}}_{\mathcal{I}_i}$. Let the maximum number of information sets along the same line of play where non-identity internal transformations are allowed before a non-identity transformation within any single deviation be n_{IN}. Let $D = \max_{I \in \mathcal{I}_i, \phi_I \in \Phi_I} |W^\Phi_I(\phi_I)| \omega(\Phi_I)$. Then, EFR's cumulative regret after T rounds with respect to Φ is upper bounded by $2^{n_{\text{IN}}+1} U |\mathcal{I}_i| \sqrt{DT}$.*

See Appendix C for technical details.

[5] While we only present the bound for the *rectified linear unit* (*ReLU*) link function, $\cdot^+ : x \mapsto \max\{0, x\}$, the arguments involved in proving Corollary 1 apply to any link function; only the final bound would change.

Algorithm 1 EFR update for player i with exact regret matching.

Input: Strategy profile, $\pi^t \in \Pi$, $t \geq 1$, and behavioral deviations, $\Phi \subseteq \Phi^{\text{IN}}_{\mathcal{I}_i}$.
initialize table $\rho^{1:0}_{\cdot,\cdot}(\cdot) = 0$.
\# Update cumulative immediate regrets:
for $I \in \mathcal{I}_i$, $\phi_I \in \Phi_I$, $w \in W^\Phi_I(\phi_I)$ **do**
  $\rho^{1:t}_{I,w}(\phi_I) \leftarrow \rho^{1:t-1}_{I,w}(\phi_I) + w^t \rho^{\text{CF}}_I(\phi_I; \pi^t)$
end for
\# Construct π_i^{t+1} with regret matching:
for $I \in \mathcal{I}_i$ from the start of the game to the end **do**
  for $\phi_I \in \Phi_I$ **do**
    \# π_i^{t+1} need only be defined at $\bar{I} \prec I$.
    $y^{t+1}_{\phi_I} \leftarrow \sum_{w \in W^\Phi_I(\phi_I)} w^{t+1} \left(\rho^{1:t}_{I,w}(\phi_I)\right)^+$
  end for
  $z^{t+1} \leftarrow \sum_{\phi_I \in \Phi_I} y^{t+1}_{\phi_I}$
  if $z^{t+1} > 0$ **then**
    $\pi_i^{t+1}(I) \leftarrow$ a fixed point of linear operator $L^t : \Delta^{|\mathcal{A}(I)|} \ni \sigma \mapsto \frac{1}{z^{t+1}} \sum_{\phi_I \in \Phi_I} y^{t+1}_{\phi_I} \phi_I(\sigma)$
  else
    $[\pi_i^{t+1}(a \mid I) \leftarrow \frac{1}{\mathcal{A}(I)}]_{a \in \mathcal{A}(I)}$ \# Arbitrary.
  end if
end for
output π_i^{t+1}

5.5. Discussion

The variable D in the EFR regret bound depends on the given behavioral deviations and is essentially the maximum number of realizable memory states times the number of action transformations across information sets. See Table C.2 in Appendix C for the D value for each deviation type.

Algorithm 1 provides an implementation of EFR with exact regret matching. Notice that as a matter of practical implementation, EFR only requires Φ_I and W^Φ_I for all information sets $I \in \mathcal{I}_i$, which are often easier to specify than Φ. Table C.2 in Appendix C shows the Φ_I and W^Φ_I parameters corresponding to each deviation type, as well as the D and n_{IN} values that determine each EFR instance's regret bound. Thanks to this feature, EFR always operates on representative deviations from Φ that are additionally augmented with re-correlation. This both potentially improves EFR's performance and ensures that learning is efficient even for some exponentially large deviation sets, like the external, blind causal, and informed causal deviations.

For example, it is equivalent to instantiate EFR with the blind causal deviations or the BPS deviations. Likewise for the informed causal deviations and the CSPS deviations, where EFR reduces to a variation of ICFR (Celli et al., 2020). To be precise, ICFR is pure EFR (analogous to pure CFR) instantiated with the CSPS deviations except that the external and internal action transformation learners at separate memory states within an information set are sampled and updated independently in ICFR. EFR therefore improves on this algorithm (beyond its generality) because EFR's action transformation learners share all experience, potentially leading to faster learning, and EFR enjoys a deterministic finite time regret bound.

Crucially, EFR's generality does not come at a computational cost. EFR reduces to the CFR algorithms previously described to handle counterfactual and action deviations (Zinkevich et al., 2007; Morrill et al., 2021). Furthermore, EFR inherits CFR's flexibility as it can be used with Monte Carlo sampling (Lanctot et al., 2009; Burch et al., 2012; Gibson et al., 2012; Johanson et al., 2012), function approximation (Waugh et al., 2015; Morrill, 2016; D'Orazio et al., 2020; Brown et al., 2019; Steinberger et al., 2020; D'Orazio, 2020), variance reduction (Schmid et al., 2019; Davis et al., 2020), and predictions (Rakhlin & Sridharan, 2013; Farina et al., 2019; D'Orazio & Huang, 2021; Farina et al., 2020b).

6. Experiments

Our theoretical results show that EFR variants utilizing more powerful deviation types are pushed to accumulate higher payoffs during learning in worst-case environments. Do these deviation types make a practical difference outside of the worst case?

We investigate the performance of EFR with different deviation types in nine benchmark game instances from *OpenSpiel* (Lanctot et al., 2019). We evaluate each EFR variant by the expected payoffs accumulated over the course of playing each game in each seat over 1000 rounds under two different regimes for selecting the other players. In the "fixed regime", other players play their parts of the fixed sequence of strategy profiles generated with self-play before the start of the experiment using one of the EFR variants under evaluation. In the "simultaneous regime", the other players are EFR instances themselves. In games with more than two players, all other players share the same EFR variant and we only record the score for the solo EFR instance. The fixed regime provides a test of how well each EFR variant adapts when the other players are gradually changing in an oblivious way where comparison is simple, while the simultaneous regime is a possibly more realistic test of dynamic adaptation where it is more difficult to draw definitive conclusions about relative effectiveness.

Since we evaluate expected payoff, use expected EFR updates, and use exact regret matching, all results are deterministic and hyperparameter-free. To compute the regret matching fixed point when internal transformations are used, we solve a linear system with the Jacobi singular value al-

Table 2. The payoff of each EFR instance averaged across both 1000 rounds and each instance pairing (eight pairs in total) in two-player and three-player goofspiel (measured in win frequency between zero and one), and Sheriff (measured in points between -6 and $+6$). The top group of algorithms use weak deviation types ($\text{ACT}_{\text{IN}} \rightarrow$ informed action deviations, $\text{CF} \rightarrow$ blind counterfactual, and $\text{CF}_{\text{IN}} \rightarrow$ informed counterfactual) and the middle group use partial sequence deviation types. The BHV instance uses the full set of behavioral deviations.

	fixed			simultaneous		
	$g_{2,5,\uparrow}$	$g_{3,4,\uparrow}$	Sheriff	$g_{2,5,\uparrow}$	$g_{3,4,\uparrow}$ [†]	Sheriff
ACT_{IN}	0.51	0.48	0.28	0.45	0.86	0.00
CF	0.56	0.51	0.48	0.50	0.88	0.34
CF_{IN}	0.57	0.51	0.60	0.50	0.92	0.37
BPS	0.58	0.51	0.58	0.50	0.85	0.34
CF	0.58	0.52	0.70	0.51	0.84	0.37
CSPS	0.59	0.52	0.61	0.51	0.91	0.37
TIPS	0.60	0.53	0.82	0.51	0.87	0.38
BHV	0.63	0.53	0.91	0.51	0.92	0.38

[†] In three-player goofspiel, players who tend to play the same actions perform worse. Since the game is symmetric across player seats, two players who use the same (deterministic) algorithm will always employ the same strategies and often play the same actions, giving the third player a substantial advantage. The win percentage for all variants in the simultaneous regime tends to be high because we only record the score for each variant when they are instantiated in a single seat. The relative comparison is still informative.

gorithm implemented by the `jacobiSvd` method from the Eigen C++ library (Guennebaud et al., 2010). Experimental data and code for generating both the data and final results are available on GitHub.[6] Experiments took roughly 20 hours to complete on a 2.10GHz Intel® Xeon® CPU E5-2683 v4 processor with 10 GB of RAM.

Appendix E hosts the full set of results but a representative summary from two variants of imperfect information goofspiel (Ross, 1971; Lanctot, 2013) (a two-player and a three-player version denoted as $g_{2,5,\uparrow}$ and $g_{3,4,\uparrow}$, respectively, both zero-sum) and Sheriff (two-player, non-zero-sum) is presented in Table 2. See Appendix E.1 for descriptions of all games.

Stronger deviations consistently lead to better performance in both the fixed and the simultaneous regime. The behavioral deviations (BHV) and the informed action deviations (ACT_{IN}) often lead to the best and worst performance, respectively, and this is true of each scenario in Table 2. In many cases however, TIPS or CSPS yield similar performance to BHV. A notable outlier from the scenarios in Table 2 is three-player goofspiel with a descending point deck. Here, blind counterfactual (CF) and BPS deviations lead to better performance in the first few rounds before all variants quickly converge to play that achieves essentially the same payoff (see Figures E.1-E.4 in Appendix E).

[6] https://github.com/dmorrill10/hr_edl_experiments

7. Conclusions

We introduced EFR, an algorithm that is hindsight rational for any given set of behavioral deviations. While the full set of behavioral deviations leads to generally intractable computational requirements, we identified four partial sequence deviation types that are both tractable and powerful in games with moderate lengths.

An important tradeoff within EFR is that using stronger deviation types generally leads to slower strategy updates, demonstrated by Figures E.5-E.6 in Appendix E where learning curves are plotted according to runtime. Often in a tournament setting, the number of rounds and computational budget may be fixed so that running faster cannot lead to more reward for the learner, but there may be reward to gain by running faster in other scenarios. Quantifying the potential benefit of using a stronger deviation type in particular games could aid in navigating this tradeoff.

Alternatively, perhaps the learner can navigate this tradeoff on their own. Algorithms like the fixed-share forecaster (Herbster & Warmuth, 1998) or context tree weighting (Willems et al., 1993) efficiently minimize regret across large structured sets of experts, effectively avoiding a similar tradeoff. This approach could also address a second tradeoff, which is that stronger deviation types lead to EFR regret bounds with larger constant factors even if the best deviation is part of a "simpler" class, *e.g.*, the regret bound that TIPS EFR has with respect to counterfactual deviations is larger than that of CFR even though a TIPS EFR instance might often accumulate more reward in order to compete with the larger TIPS deviations. Perhaps an EFR variant can be designed that would compete with large sets of behavioral deviations, but its regret bound would scale with the "complexity" (in a sense that has yet to be rigorously defined) of the best deviation rather than the size of the whole deviation set. Ideally, its computational cost would be independent of the deviation set size or would at least scale with the complexity of the best deviation.

Acknowledgements

Computation provided by WestGrid and Compute Canada. Dustin Morrill, Michael Bowling, and James Wright are supported by the Alberta Machine Intelligence Institute (Amii) and NSERC. Michael Bowling and James Wright hold Canada CIFAR AI Chairs at Amii. Amy Greenwald is supported in part by NSF Award CMMI-1761546. Thanks to Ian Gemp, Gabriele Farina, and anonymous reviewers for constructive comments and suggestions.

References

Aumann, R. J. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1): 67–96, 1974.

Blum, A. and Mansour, Y. From external to internal regret. *Journal of Machine Learning Research*, 8(Jun): 1307–1324, 2007.

Brown, N., Lerer, A., Gross, S., and Sandholm, T. Deep counterfactual regret minimization. In *Proceedings of the 36th International Conference on Machine Learning (ICML-19)*, pp. 793–802, 2019.

Burch, N., Lanctot, M., Szafron, D., and Gibson, R. Efficient monte carlo counterfactual regret minimization in games with many player actions. In *Advances in Neural Information Processing Systems*, pp. 1880–1888, 2012.

Celli, A., Marchesi, A., Farina, G., and Gatti, N. No-regret learning dynamics for extensive-form correlated equilibrium. *Advances in Neural Information Processing Systems*, 33, 2020.

Davis, T., Schmid, M., and Bowling, M. Low-variance and zero-variance baselines for extensive-form games. In *International Conference on Machine Learning*, pp. 2392–2401. PMLR, 2020.

D'Orazio, R. *Regret Minimization with Function Approximation in Extensive-Form Games*. Master's thesis, University of Alberta, 2020.

D'Orazio, R. and Huang, R. Optimistic and adaptive lagrangian hedging. In *Reinforcement Learning in Games Workshop at the Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

D'Orazio, R., Morrill, D., Wright, J. R., and Bowling, M. Alternative function approximation parameterizations for solving games: An analysis of f-regression counterfactual regret minimization. In *Proceedings of The Nineteenth International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 5 2020.

Dudík, M. and Gordon, G. J. A sampling-based approach to computing equilibria in succinct extensive-form games. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI-2009)*, 2009.

Farina, G., Kroer, C., Brown, N., and Sandholm, T. Stable-predictive optimistic counterfactual regret minimization. In *International Conference on Machine Learning*, pp. 1853–1862, 2019.

Farina, G., Bianchi, T., and Sandholm, T. Coarse correlation in extensive-form games. In *Thirty-Fourth AAAI Conference on Artificial Intelligence, February 7-12, 2020, New York, New York, USA*, 2020a.

Farina, G., Kroer, C., and Sandholm, T. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. *arXiv preprint arXiv:2007.14358*, 2020b.

Forges, F. and von Stengel, B. Computionally Efficient Coordination in Games Trees. THEMA Working Papers 2002-05, THEMA (THéorie Economique, Modélisation et Applications), Université de Cergy-Pontoise, 2002.

Foster, D. P. and Vohra, R. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1-2):7–35, 1999.

Gibson, R. *Regret Minimization in Games and the Development of Champion Multiplayer Computer Poker-Playing Agents*. PhD thesis, University of Alberta, 2014.

Gibson, R., Lanctot, M., Burch, N., Szafron, D., and Bowling, M. Generalized sampling and variance in counterfactual regret minimization. In *Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12)*, pp. 1355–1361, 2012.

Gordon, G. J., Greenwald, A., and Marks, C. No-regret learning in convex games. In *Proceedings of the 25th international conference on Machine learning*, pp. 360–367, 2008.

Greenwald, A., Jafari, A., and Marks, C. A general class of no-regret learning algorithms and game-theoretic equilibria. In *Proceedings of the 2003 Computational Learning Theory Conference*, pp. 1–11, 8 2003.

Greenwald, A., Li, Z., and Marks, C. Bounds for regret-matching algorithms. In *ISAIM*, 2006.

Guennebaud, G., Jacob, B., et al. Eigen. *URl: http://eigen. tuxfamily. org*, 2010.

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

Herbster, M. and Warmuth, M. K. Tracking the best expert. *Machine learning*, 32(2):151–178, 1998.

Johanson, M., Bard, N., Lanctot, M., Gibson, R., and Bowling, M. Efficient nash equilibrium approximation through monte carlo counterfactual regret minimization. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2012.

Kash, I. A., Sullins, M., and Hofmann, K. Combining no-regret and q-learning. In *Proceedings of The Nineteenth International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, May 2020.

Kuhn, H. W. Extensive games and the problem of information. *Contributions to the Theory of Games*, 2:193–216, 1953.

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

Lanctot, M., Waugh, K., Zinkevich, M., and Bowling, M. Monte Carlo sampling for regret minimization in extensive games. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22*, pp. 1078–1086, 2009.

Lanctot, M., Lockhart, E., Lespiau, J.-B., Zambaldi, V., Upadhyay, S., Pérolat, J., Srinivasan, S., Timbers, F., Tuyls, K., Omidshafiei, S., Hennes, D., Morrill, D., Muller, P., Ewalds, T., Faulkner, R., Kramár, J., Vylder, B. D., Saeta, B., Bradbury, J., Ding, D., Borgeaud, S., Lai, M., Schrittwieser, J., Anthony, T., Hughes, E., Danihelka, I., and Ryan-Davis, J. OpenSpiel: A framework for reinforcement learning in games. *CoRR*, abs/1908.09453, 2019. URL http://arxiv.org/abs/1908.09453.

Morrill, D. *Using Regret Estimation to Solve Games Compactly*. Master's thesis, University of Alberta, 2016.

Morrill, D., D'Orazio, R., Sarfati, R., Lanctot, M., Wright, J. R., Greenwald, A., and Bowling, M. Hindsight and sequential rationality of correlated play. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, February 2-9, 2021, virtual*, 2021.

Rakhlin, S. and Sridharan, K. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pp. 3066–3074, 2013.

Ross, S. M. Goofspiel — the game of pure strategy. *Journal of Applied Probability*, 8(3):621–625, 1971.

Schmid, M., Burch, N., Lanctot, M., Moravcik, M., Kadlec, R., and Bowling, M. Variance reduction in monte carlo counterfactual regret minimization (vr-mccfr) for extensive form games using baselines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2157–2164, 2019.

Steinberger, E., Lerer, A., and Brown, N. Dream: Deep regret minimization with advantage baselines and model-free learning. *arXiv preprint arXiv:2006.10410*, 2020.

Tammelin, O. Solving large imperfect information games using cfr+. *arXiv preprint arXiv:1407.5042*, 2014.

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

von Stengel, B. and Forges, F. Extensive-form correlated equilibrium: Definition and computational complexity. *Mathematics of Operations Research*, 33(4):1002–1022, 2008.

Waugh, K., Morrill, D., Bagnell, J. A., and Bowling, M. Solving games with functional regret estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Willems, F. M., Shtarkov, Y. M., and Tjalkens, T. J. Context tree weighting: a sequential universal source coding procedure for fsmx sources. In *1993 IEEE International Symposium on Information Theory*, pp. 59. Institute of Electrical and Electronics Engineers, 1993.

Zinkevich, M., Johanson, M., Bowling, M. H., and Piccione, C. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, pp. 1729–1736, December 2007.