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

Dustin Morrill[1] morrill@ualberta.ca **Ryan D'Orazio**[2] **Marc Lanctot**[3]

James R. Wright[1] **Michael Bowling**[1, 3] **Amy R. Greenwald**[4]

[1]University of Alberta; Alberta Machine Intelligence Institute, Canada
[2]Université de Montréal; Mila, Canada
[3]DeepMind
[4]Brown University, United States

A Notation and Symbols

$\mathcal{A}_*$	The union of player i's action sets.
$\mathbb{h}(I)$	An arbitrary history in information set I.
v_I	The counterfactual value function at information set I.
$\phi \in \Phi^{\cdot}_{\mathcal{X}}$	A transformation from finite set $\mathcal{X}$ to itself. SW in the superscript of $\Phi^{\cdot}_{\mathcal{X}}$ denotes the set of swap transformations, EX denotes the external transformations, and IN denotes the internal transformations.
$\phi^1 : a \mapsto a$	The identity transformation.
$\phi^{\to a}$	The external transformation to a.
$\phi^{a \to a'}$	The internal transformation from a to a'.
$\Phi^{\text{IN}}_{\mathcal{I}_i}$	The set of player i's behavioral deviations.
$I \in \mathcal{I}_i$	One of player i's information sets.
$\mathbb{I}(h)$	The information set containing history h.
U	The maximum magnitude of any payoff.
$\mathbb{p}(I')$	The unique parent (immediate predecessor) of information set I'.
$\mathcal{P}$	The player choice function.
$s \in S$	A pure strategy profile.
S_c	The set of a game's random events or the set of pure strategies that could be assigned to chance.
P	The reach probability function.
$\mu \in \Delta^{\lvert S \rvert}$	A distribution over strategy profiles, often representing an empirical distribution of play and interpreted as a recommendation distribution.
ρ	Regret.
ρ^{CF}	Counterfactual regret.
Δ^d	The d-dimensional probability simplex.

$\pi \in \Pi$	A behavioral/mixed strategy profile.
$w \in W(\phi)$	A time selection function associated with transformation ϕ.
$W_I^\Phi(\phi_I)$	The set of time selection functions associated with action transformation ϕ_I corresponding to the deviation player memory probabilities generated by the set of behavioral deviations Φ in information set I.
u_i	Bounded utility function for player i.
$a \in \mathcal{A}(h) = \mathcal{A}(\mathbb{I}(h))$	An action from the set of legal actions at history h in information set $\mathbb{I}(h)$.
$a_h^{\to I'}$ or $a_{\mathbb{I}(h)}^{\to I'}$	The unique action that would need to be taken in history h or information set $I(h)$ to reach successor information set $I' \succ I(h)$.
d_*	The depth of player i's deepest information set.
d_I	The depth of information set I.
$g \in G_i$	A deviation player memory state string.
$h \in \mathcal{H}$	An action history.
$M(\phi)$	The size of the time selection function set associated with transformation ϕ.
M^*	The size of the largest time selection function set.
$n_\mathcal{A}$	The maximum number of actions available at any history.
$z \in \mathcal{Z} \subseteq \mathcal{H}$	A terminal history.

B Regret Matching for Time Selection

In an online decision problem (also called a *prediction with expert advice* problem), *regret matching* is a learning algorithm that accumulates a vector of regrets, $\rho^{1:t-1}$—one for each deviation or "expert", $\phi \in \Phi \subseteq \Phi_{S_i}^{\text{SW}}$—and chooses its mixed strategy, π^t, on each round as the fixed point of a linear operator. We generalize this algorithm and three extensions—regret matching$^+$, regret approximation, and predictions—to the time selection setting.

B.1 Background

Regret Matching

The regret matching operator is constructed from a vector of non-negative *link outputs*, $y^t \in \mathbb{R}_+^{|\Phi|}$, generated by applying a link function, $f : \mathbb{R}^{|S_i|} \to \mathbb{R}_+^{|S_i|}$, to the cumulative regrets, *i.e.*, $y^t = f(\rho^{1:t-1})$. The operator is defined as

$$L^t : \pi_i \mapsto \frac{1}{z^t} \sum_{\phi \in \Phi} \phi(\pi_i) y_\phi^t, \tag{1}$$

where $z^t = \sum_{\phi \in \Phi} y_\phi^t$ is the sum of the link outputs, and π_i^t is chosen arbitrarily if $z^t = 0$.

Regret bounds are generally derived for regret matching algorithms by choosing $f = \alpha g$ for some $\alpha > 0$, where g is part of a Gordon triple (Gordon 2005), (G, g, γ). A Gordon triple is a triple consisting of a potential function, $G : \mathbb{R}^n \to \mathbb{R}$, a scaled link function $g : \mathbb{R}^n \to \mathbb{R}_+^n$, and a size function, $\gamma : \mathbb{R}^n \to \mathbb{R}_+$, where they satisfy the generalized smoothness condition $G(x + x') \leq G(x) + x' \cdot g(x) + \gamma(x')$ for any $x, x' \in \mathbb{R}^n$. By applying the potential function to the cumulative regret, we can unroll the recursive bound to get a simple bound on the cumulative regret itself.

While its bounds are not quite optimal, Hart and Mas-Colell (2000)'s original regret matching algorithm, defined with the *rectified linear unit* (*ReLU*) link function, $\cdot^+ = \max\{\cdot, 0\}$, is often exceptionally effective in practice (see, *e.g.*, Waugh and Bagnell (2015); Burch (2017)). We focus our analysis on this link function, but our arguments readily apply to other link functions. Only the final regret bounds will change. We follow the typical convention for analyzing Hart and Mas-Colell (2000)'s regret matching with $\gamma(x) = \frac{1}{2}\|x\|_2^2$, $G(x) = \gamma(x^+)$, and $g = f$.

Regret Matching$^+$

Instead of the cumulative regrets, regret matching$^+$ updates a vector of pseudo regrets (sometimes called "q-regrets"), $q^{1:t} = (q^{1:t-1} + \rho^t)^+ \geq \rho^{1:t}$ (Tammelin 2014; Tammelin et al. 2015). If we assume a *positive invariant* potential function where $G((x+x')^+) \leq G(x+x')$, then the same regret bounds follow from the same arguments used in the analysis of regret matching D'Orazio (2020). Note that this condition is satisfied with equality for the quadratic potential $G(x) = \frac{1}{2}\|x^+\|_2^2$.

Regret Approximation

Approximate regret matching is regret matching with approximated cumulative regrets, $\widetilde{\rho^{1:t-1}} \approx \rho^{1:t-1}$ (Waugh et al. 2015; D'Orazio et al. 2020) or q-regrets, $\widetilde{q^{1:t-1}} \approx q^{1:t-1}$ (Morrill 2016; D'Orazio 2020). The regret of approximate regret matching depends on its approximation accuracy and motivates the use of function approximation when it is impractical to store and update the regret for each deviation individually. While it requires an extra assumption, we derive simpler approximate regret matching bounds than those derived by D'Orazio et al. (2020); D'Orazio (2020) through an analysis of regret matching with predictions.

Optimism via Predictions

Optimistic regret matching augments its link inputs by adding a prediction of the instantaneous regret on the next round, *i.e.*, $m^t \sim \rho^t$. If the predictions are accurate then the algorithm's cumulative regret will be very small. This is a direct application of optimistic Lagrangian Hedging (D'Orazio and Huang 2021) to Φ-regret. The general approach of adding predictions to improve the performance of regret minimizers originates with Rakhlin and Sridharan (2013); Syrgkanis et al. (2015).

D'Orazio and Huang (2021)'s analysis requires that G and g satisfy $G(x') \geq G(x) + \langle g(x),\, x' - x\rangle$, which is achieved, for example, if G is convex and g is a subgradient of G. Note that this is achieved for Hart and Mas-Colell (2000)'s regret matching because Greenwald, Li, and Marks (2006) shows that the ReLU function is the gradient of the convex quadratic potential $G(x) = \frac{1}{2}\|x^+\|_2^2$.

B.2 Time Selection

To adapt regret matching to the time selection framework, we treat each deviation–time selection function pair as a separate expert and sum over the link outputs corresponding to a given deviation to construct the regret matching operator. Our goal is then to ensure that each element of the cumulative regret matrix, $\rho^{1:T}$, grows sublinearly, where each index in the second dimension corresponds to a time selection function. Each deviation $\phi \in \Phi$ is assigned a finite set of time selection functions, $w \in W(\phi)$, so the regret matrix entries corresponding to (ϕ, w)-pairings where $w \notin W(\phi)$, are always zero.

To facilitate a unified analysis, we assume a general optimistic regret matching algorithm that, after $t-1$ rounds, uses link outputs $y_\phi^t = \sum_{w \in W(\phi)} w^t (x_{\phi,w}^t + m_{\phi,w}^t)^+$, where either $x^t = \rho^{1:t-1}$ or $x^t = q^{1:t-1}$ with $x^1 = \mathbf{0}$, and m^t is a matrix of arbitrary predictions or approximation errors. Notice that this means that $x^t + m^t$ can be generated from a function approximator instead of storing either term in a table. Denoting the weighted sum of the link outputs as $z^t = \sum_{\phi \in \Phi} y_\phi^t$, the regret matching operator has the same form as initially defined, *i.e.*,

$$L^t : \pi_i \mapsto \frac{1}{z^t} \sum_{\phi \in \Phi} \phi(\pi_i) y_\phi^t. \tag{2}$$

With this, we can bound the regret of optimistic regret matching thusly:

Theorem 2. *Establish deviation set $\Phi \subseteq \Phi_{S_i}^{\text{SW}}$ and finite time selection sets $W(\phi) = \{w \in [0,1]^T\}_{j=1}^{M(\phi)}$ for each deviation $\phi \in \Phi$. On each round $1 \leq t \leq T$, $(\Phi, \cdot^+)$-regret matching with respect to matrix x^t (equal to either $\rho^{1:t-1}$ or $q^{1:t-1}$) and predictions m^t chooses its strategy, $\pi_i^t \in \Pi_i$, to be the fixed point of $L^t : \pi_i \mapsto {}^1\!/_{z^t} \sum_{\phi \in \Phi} \phi(\pi_i) y_\phi^t$ or an arbitrary strategy when $z^t = 0$, where link outputs are generated from $y_\phi^t = \sum_{w \in W(\phi)} w^t (x_{\phi,w}^t + m_{\phi,w}^t)^+$ and $z^t = \sum_{\phi \in \Phi} y_\phi^t$. This*

algorithm ensures that

$$\rho^{1:T}(\phi, w) \leq \sqrt{\sum_{t=1}^{T} \sum_{\substack{\phi' \in \Phi, \\ \bar{w} \in W(\phi')}} \left(\bar{w}^t \rho(\phi'; \pi^t) - m^t_{\phi', \bar{w}} \right)^2}$$

for every deviation ϕ and time selection function w.

Proof. Let us overload $W = \bigcup_{\phi \in \Phi} W(\phi)$ and let $a_{\cdot,w} = [a_{\phi,w}]_{\phi \in \Phi}$ for any matrix $a \in \mathbb{R}^{|\Phi| \times |W|}$. Then, for any time selection function, $w \in W$, the quadratic potential function, $G(x) = \frac{1}{2}\|x^+\|_2^2$ is convex, positive invariant (with equality), has the ReLU function as its gradient (Greenwald, Li, and Marks 2006), and is smooth with respect to $\gamma(x) = \frac{1}{2}\|x\|_2^2$. Altogether, these properties imply that

$$G\Big(\big(x^t_{\cdot,w} + w^t \rho^t\big)^+ \Big) = G\Big(\big(x^t_{\cdot,w} + m^t_{\cdot,w} + w^t \rho^t - m^t_{\cdot,w}\big)^+ \Big) \tag{3}$$

$$= G\big(x^t_{\cdot,w} + m^t_{\cdot,w} + w^t \rho^t - m^t_{\cdot,w}\big) \tag{4}$$

$$\leq G\big(x^t_{\cdot,w} + m^t_{\cdot,w}\big) + \langle w^t \rho^t - m^t_{\cdot,w},\ \big(x^t_{\cdot,w} + m^t_{\cdot,w}\big)^+ \rangle + \gamma\big(w^t \rho^t - m^t_{\cdot,w}\big), \tag{5}$$

where $\rho^t = [\rho(\phi; \pi^t)]_{\phi \in \Phi}$ is the vector of instantaneous regrets on round t.

By convexity, $G(a) - G(b) \leq \langle \nabla G(a),\ a - b \rangle$, for any vectors a and b, so we substitute $a = x^t_{\cdot,w} + m^t_{\cdot,w}$ and $b = x^t_{\cdot,w}$ to bound $G(x^t_{\cdot,w} + m^t_{\cdot,w}) - \langle m^t_{\cdot,w},\ (x^t_{\cdot,w} + m^t_{\cdot,w})^+ \rangle \leq G(x^t_{\cdot,w})$. Therefore,

$$G\Big(\big(x^t_{\cdot,w} + w^t \rho^t\big)^+ \Big) \leq G(x^t_{\cdot,w}) + w^t \langle \rho^t,\ \big(x^t_{\cdot,w} + m^t_{\cdot,w}\big)^+ \rangle + \gamma\big(w^t \rho^t - m^t_{\cdot,w}\big) \tag{6}$$

$$= G\Big(\big(x^t_{\cdot,w}\big)^+ \Big) + w^t \langle \rho^t,\ \big(x^t_{\cdot,w} + m^t_{\cdot,w}\big)^+ \rangle + \gamma\big(w^t \rho^t - m^t_{\cdot,w}\big). \tag{7}$$

Summing the potentials across time selection functions,

$$\sum_{w \in W} G\Big(\big(x^t_{\cdot,w} + w^t \rho^t\big)^+ \Big) \leq \sum_{w \in W} G\Big(\big(x^t_{\cdot,w}\big)^+ \Big) + w^t \langle \rho^t,\ \big(x^t_{\cdot,w} + m^t_{\cdot,w}\big)^+ \rangle + \gamma\big(w^t \rho^t - m^t_{\cdot,w}\big). \tag{8}$$

With some algebra, we can rewrite the sum of inner products:

$$\sum_{w \in W} w^t \langle \rho^t,\ (x^t_{\cdot,w} + m^t_{\cdot,w})^+ \rangle = \sum_{w \in W} \sum_{\phi \in \Phi} w^t \rho(\phi; \pi^t) \big(x^t_{\phi,w} + m^t_{\phi,w}\big)^+ \tag{9}$$

$$= \sum_{\phi \in \Phi} \rho(\phi; \pi^t) \sum_{w \in W(\phi)} w^t \big(x^t_{\phi,w} + m^t_{\phi,w}\big)^+ \tag{10}$$

$$= \sum_{\phi \in \Phi} \rho(\phi; \pi^t) y^t_\phi \tag{11}$$

$$= \langle \rho^t,\ y^t \rangle. \tag{12}$$

Since the strategy π^t_i is the fixed point of L^t generated from link outputs y^t, the Blackwell condition $\langle \rho^t,\ y^t \rangle \leq 0$ is satisfied with equality. For proof, see, for example, Greenwald, Li, and Marks (2006). The sum of potential functions after T rounds are then bounded as

$$\sum_{w \in W} G\Big(\big(x^T_{\cdot,w} + w^T \rho^T\big)^+ \Big) \leq \sum_{w \in W} G\Big(\big(x^T_{\cdot,w}\big)^+ \Big) + \gamma(w^T \rho^T - m^T_{\cdot,w}). \tag{13}$$

Expanding the definition of γ,

$$\sum_{w \in W} G\Big(\big(x^T_{\cdot,w} + w^T \rho^T\big)^+ \Big) \leq \sum_{w \in W} G\Big(\big(x^T_{\cdot,w}\big)^+ \Big) + \frac{1}{2} \sum_{w \in W} \sum_{\phi \in \Phi} \big(w^T \rho(\phi; \pi^T) - m^T_{\phi,w}\big)^2 \tag{14}$$

$$= \sum_{w \in W} G\Big(\big(x^T_{\cdot,w}\big)^+ \Big) + \frac{1}{2} \sum_{\substack{\phi \in \Phi, \\ w \in W(\phi)}} \big(w^T \rho(\phi; \pi^T) - m^T_{\phi,w}\big)^2. \tag{15}$$

Unrolling the recursion accross time,

$$= \frac{1}{2} \sum_{t=1}^{T} \sum_{\substack{\phi \in \Phi, \\ w \in W(\phi)}} \left(w^t \rho(\phi; \pi^t) - m^t_{\phi,w} \right)^2. \tag{16}$$

We lower bound

$$\sum_{w \in W} G\left(\left(x^{T+1}_{\cdot,w} \right)^+ \right) = \frac{1}{2} \sum_{w \in W} \sum_{\phi \in \Phi} \left(\left(x^{T+1}_{\phi,w} \right)^+ \right)^2 \tag{17}$$

$$\geq \frac{1}{2} \max_{\substack{\phi \in \Phi, \\ w \in W(\phi)}} \left(\left(x^{T+1}_{\phi,w} \right)^+ \right)^2 \tag{18}$$

so that

$$\frac{1}{2} \max_{\substack{\phi \in \Phi, \\ w \in W(\phi)}} \left(\left(x^{T+1}_{\phi,w} \right)^+ \right)^2 \leq \frac{1}{2} \sum_{t=1}^{T} \sum_{\substack{\phi \in \Phi, \\ w \in W(\phi)}} \left(w^t \rho(\phi; \pi^t) - m^t_{\phi,w} \right)^2. \tag{19}$$

Multiplying both sides by two, taking the square root, and applying $\rho^{1:T}(\phi, w) \leq \left(x^{T+1}_{\phi,w} \right)^+$, we arrive at the final bound,

$$\max_{\substack{\phi \in \Phi, \\ w \in W(\phi)}} \rho^{1:T}(\phi, w) \leq \sqrt{\sum_{t=1}^{T} \sum_{\substack{\phi' \in \Phi, \\ \bar{w} \in W(\phi')}} \left(\bar{w}^t \rho(\phi'; \pi^t_i) - m^t_{\phi',\bar{w}} \right)^2}. \tag{20}$$

Since the bound is true of the worst-case $\phi \in \Phi$ and $w \in W$, it is true of each pair, thereby proving the claim. □

If all of the predictions m^t are zero, then we arrive at a simple bound for exact regret matching. We only prove the bound for ordinary regret matching for simplicity but the result and arguments are identical for exact regret matching$^+$.

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, Li, and Marks 2006).*

Proof. Since $m^t = \mathbf{0}$ on every round t, we know from Theorem 2 that

$$\rho^{1:T}(\phi, w) \leq \sqrt{\sum_{t=1}^{T} \sum_{\substack{\phi' \in \Phi, \\ \bar{w} \in W(\phi')}} \left(\bar{w}^t \rho(\phi'; \pi^t_i) \right)^2} \tag{21}$$

$$= \sqrt{\sum_{t=1}^{T} \sum_{\phi' \in \Phi} \left(\rho(\phi'; \pi^t_i) \right)^2 \sum_{\bar{w} \in W(\phi')} \left(\bar{w}^t \right)^2}. \tag{22}$$

Since $0 \leq \bar{w}^t \leq 1$,

$$\leq \sqrt{M^* \sum_{t=1}^{T} \sum_{\phi' \in \Phi} \left(\rho(\phi'; \pi^t_i) \right)^2}. \tag{23}$$

Since $\sum_{\phi' \in \Phi} (\rho(\phi'; \pi_i^t))^2 \leq (2U)^2 \omega(\Phi)$ (see Greenwald, Li, and Marks (2006)),

$$\leq \sqrt{M^* (2U)^2 \omega(\Phi) T} \tag{24}$$

$$= 2U \sqrt{M^* \omega(\Phi) T}. \tag{25}$$

$$\tag{26}$$

This completes the argument. □

If $x^t + m^t$ is generated from a function attempting to approximate $x^t + [w^t \rho(\phi; \pi^t)]_{\phi \in \Phi, w \in W}$, then we can rewrite Theorem 2 in terms of its approximation error.

Corollary 2. *Establish deviation set $\Phi \subseteq \Phi_{S_i}^{\text{SW}}$ and finite time selection sets $W(\phi) = \{w \in [0,1]^T\}_{j=1}^{M(\phi)}$ for each deviation $\phi \in \Phi$. On each round $1 \leq t \leq T$, approximate $(\Phi, \cdot^+)$-regret matching with respect to matrix x^t (equal to either $\rho^{1:t-1}$ or $q^{1:t-1}$) chooses its strategy, $\pi_i^t \in \Pi_i$, to be the fixed point of $L^t : \pi_i \mapsto {}^1\!/\!_{z^t} \sum_{\phi \in \Phi} \phi(\pi_i) y_\phi^t$ or an arbitrary strategy when $z^t = 0$, where link outputs are generated from approximation matrix $\widetilde{y}^t \in \mathbb{R}^{|\Phi| \times |W|}$ as $y_\phi^t = \sum_{w \in W(\phi)} w^t (\widetilde{y}_{\phi,w}^t)^+$ and $z^t = \sum_{\phi \in \Phi} y_\phi^t$. This algorithm ensures that*

$$\rho^{1:T}(\phi, w) \leq \sqrt{\sum_{t=1}^{T} \sum_{\substack{\phi' \in \Phi, \\ \bar{w} \in W(\phi')}} \left(x_{\phi', \bar{w}}^t + \bar{w}^t \rho(\phi'; \pi^t) - \widetilde{y}_{\phi', \bar{w}}^t \right)^2}$$

for every deviation ϕ and time selection function w.

Proof. Since the predictions m^t are arbitrary, we can set $\widetilde{y}^t = x^t + m^t$, which implies that $m^t = \widetilde{y}^t - x^t$. Substituting this into the bound of Theorem 2, we arrive at the desired result. □

C EFR

EFR's regret decomposition is a straightforward generalization of CFR's by Zinkevich et al. (2007) but it requires us to build up some mathematical machinery.

C.1 Preliminaries

The parent action function. The action taken to reach a given information set from its parent is returned by $\mathfrak{a} : I' \mapsto a_{\mathbb{P}(I')}^{\to I'}$ ("blackboard a").

Terminal successor histories. Let the histories that terminate without further input from player i after taking action a in I be

$$\mathcal{Z}_i(I, a) = \left\{ z \in \mathcal{Z} \;\middle|\; \begin{array}{l} \exists h \in I,\, z \sqsupseteq ha, \\ \nexists h' \in \mathcal{H}_i, ha \sqsubseteq h' \sqsubset z \end{array} \right\}. \tag{27}$$

Child information sets. Let the child information sets of information set I after taking action a be

$$\mathcal{I}_i(I, a) = \left\{ I' \in \mathcal{I}_i \;\middle|\; \begin{array}{l} \forall h' \in I',\, \exists h \in I,\, h' \sqsupseteq ha, \\ \nexists h'' \in \mathcal{H}_i, ha \sqsubseteq h'' \sqsubset h' \end{array} \right\}. \tag{28}$$

Deviation player observations. Let

$$o : \mathcal{A}(I) \ni a; \Phi_{\mathcal{A}(I)}^{\text{SW}} \ni \phi = \begin{cases} * & \text{if } \phi \in \Phi_{\mathcal{A}(I)}^{\text{EX}} \\ a & \text{o.w.} \end{cases}$$

return the observation that the deviation player makes when they apply action transformation ϕ and action a is recommended. Now, we can characterize how memory probabilities evolve in general as the deviation player chooses actions. For any $I' \in \mathcal{I}_i(I, a')$ child of I following action a' and observation $b \in \{*\} \cup \mathcal{A}(I)$, the probability of memory gb under behavioral deviation ϕ is

$$w_\phi(I', gb) = w_\phi(I, g) \sum_{a \in \mathcal{A}(I)} \mathbb{1}\{\phi_{I,g}(a) = a' \wedge o(a; \phi_{I,g}) = b\} \pi_i(a \,|\, I). \tag{29}$$

C.2 Counterfactual Value

The counterfactual value of an action. The counterfactual value function is

$$v_I : a; \pi \mapsto \sum_{\substack{h \in I, \\ z \in \mathcal{Z}}} P(h; \pi_{-i}) P(ha, z; \pi_i, \pi_{-i}) u_i(z).$$

We overload $v_I(\pi_i'(I); \pi) = \mathbb{E}_{a' \sim \pi_i'(I)} v_I(a'; \pi)$. If the same strategy is used in I as well as the following information sets, we overload $v_I(\pi) = v_I(\pi_i(I); \pi)$. By splitting up the histories that lead out of I into those that terminate without further input from i and those that lead to to child information sets, we can decompose counterfactual values recursively:

$$\begin{aligned}
v_I(a; \pi) &= \sum_{\substack{h \in I, \\ z \in \mathcal{Z}}} P(h; \pi_{-i}) P(ha, z; \pi) u_i(z) && (30) \\
&= \sum_{\substack{h \in I, \\ z \in \mathcal{Z}}} P(ha, z; \pi_i) \underbrace{P(z; \pi_{-i}) u_i(z)}_{\text{Terminal counterfactual values.}} && (31) \\
&= \underbrace{\sum_{z \in \mathcal{Z}_i(I,a)} P(z; \pi_{-i}) u_i(z)}_{\text{Expected value from terminal histories.}} + \underbrace{\sum_{\substack{h' \in I' \in \mathcal{I}_i(I,a) \\ z \in \mathcal{Z}}} P(h', z; \pi_i) P(z; \pi_{-i}) u_i(z).}_{\text{Expected value from non-terminal histories.}} && (32)
\end{aligned}$$

If we define $r(I, a; \pi_{-i}) = \sum_{z \in \mathcal{Z}_i(I,a)} P(z; \pi_{-i}) u_i(z)$, then

$$\begin{aligned}
&= r(I, a; \pi_{-i}) + \sum_{I' \in \mathcal{I}_i(I,a)} \underbrace{\sum_{a' \in \mathcal{A}(I')} \pi_i(a' \mid I') \sum_{\substack{h' \in I' \\ z \in \mathcal{Z}}} P(h'a', z; \pi_i) P(z; \pi_{-i}) u_i(z)}_{v_{I'}(\pi)}. && (33) \\
&= \underbrace{r(I, a; \pi_{-i})}_{\text{Expected immediate value.}} + \underbrace{\sum_{I' \in \mathcal{I}_i(I,a)} v_{I'}(\pi)}_{\text{Expected future value.}}. && (34)
\end{aligned}$$

The counterfactual value of a behavioral deviation. To account for the deviation player's memory when evaluating behavioral deviations, we must define a new variation of counterfactual value.

Definition 1. *The counterfactual value of behavioral deviation $\phi \in \Phi^{\text{IN}}_{\mathcal{I}_i}$ from information set I and memory state $g \in G_i$, given strategy profile $\pi \in \Pi$, is*

$$\hat{v}_{I,g}(\phi; \pi) = \sum_{a \in \mathcal{A}(I)} \pi_i(a \mid I) \left(r(I, \phi_{I,g}(a); \pi_{-i}) + \sum_{I' \in \mathcal{I}_i(I, \phi_{I,g}(a))} \hat{v}_{I', go(a; \phi_{I,g})}(\phi; \pi) \right).$$

At the start of the game, this counterfactual value matches the expected value of $\phi(\pi_i)$, *i.e.*,

$$\hat{v}_{I,g}(\phi; \pi) = \mathbb{E}_{s_i' \sim \phi(\pi_i)}[u_i(s_i', \pi_{-i})].$$

If there are no non-identity internal transformations in ϕ, then $\hat{v}$ reduces to the usual counterfactual value function under strategy profile $(\phi(\pi_i), \pi_{-i})$. If all of the transformations following I are identity transformations, then the counterfactual value of ϕ is just the counterfactual value of applying the transformation at I and memory state g, *i.e.*,

$$\begin{aligned}
\hat{v}_{I,g}(\phi_{\preceq I, \sqsubseteq g}; \pi) &= \sum_{a \in \mathcal{A}(I)} \pi_i(a \mid I) \left(r(I, \phi_{I,g}(a); \pi_{-i}) + \sum_{I' \in \mathcal{I}_i(I, \phi_{I,g}(a))} \hat{v}_{I', go(a; \phi_{I,g})}(\phi^1; \pi) \right) && (35) \\
&= \sum_{a \in \mathcal{A}(I)} \pi_i(a \mid I) \left(r(I, \phi_{I,g}(a); \pi_{-i}) + \sum_{I' \in \mathcal{I}_i(I, \phi_{I,g}(a))} v_{I'}(\pi) \right) && (36) \\
&= v_I(\phi_{I,g}(\pi_i(I)); \pi). && (37)
\end{aligned}$$

C.3 Regret

Next, we generalize the idea of immediate and full regret to behavioral deviations.

Immediate regret. The immediate regret of behavioral deviation ϕ from information set I and memory state g is the immediate counterfactual regret weighted by the probability of g, *i.e.*,

$$\rho_I(\phi_{\preceq I, \sqsubseteq g}; \pi) = w_\phi(I, g; \pi_i)(v_I(\phi_{I,g}(\pi_i(I)); \pi) - v_I(\pi)) \tag{38}$$

$$= w_\phi(I, g; \pi_i)\rho_I^{\text{CF}}(\phi_{I,g}; \pi). \tag{39}$$

Full regret. The full regret of behavioral deviation ϕ at information set I and memory state g is

$$\rho_{I,g}(\phi; \pi) = w_\phi(I, g; \pi_i)(\hat{v}_{I,g}(\phi; \pi) - v_I(\pi)). \tag{40}$$

Some intuition can be gained about generalized immediate and full regret by making a formal connection between memory probabilities and reach probabilities. If zero non-identity internal transformations are used in behavioral deviation ϕ, then there is a unique memory state g that ϕ realizes in information set I and its probability coincides with the reach probability of the deviation strategy, $\phi(\pi_i)$, *i.e.*, $w_\phi(I, g; \pi_i) = P(\mathbb{h}(I); \phi(\pi_i))$. After internal transformation $\phi^{a' \to a}$ that outputs $a \neq a'$ leading to a set of successors $\{I' \succ I\}$, there are two possible memory states, ga and ga'. The reach probability at any I' is then the sum of memory probabilities across these states, *i.e.*, $P(\mathbb{h}(I'); \phi(\pi_i)) = w_\phi(I', ga; \pi_i) + w_\phi(I', ga'; \pi_i)$. Thus, the full counterfactual regret at I' weighted by a memory probability is nearly the difference in expected payoff from I' between $\phi(\pi_i)$ and π_i assuming that both play to I' according to $\phi(\pi_i)$. Minimizing regret with respect to each memory state is a stronger property than minimizing regret with respect to each deviation strategy reach probability because memory states distinguish between the different ways an information set could be reached.

At the start of the game, there is only one memory state, $\varnothing$, and $w_\phi(I, \varnothing; \pi_i) = 1$, which means that $\rho_{I,g}(\phi; \pi)$ reduces to the difference in expected value achieved by $\phi(\pi_i)$ and π_i, *i.e.*,

$$\rho_{I,g}(\phi; \pi) = \mathbb{E}_{s_i' \sim \phi(\pi_i)}[u_i(s_i', \pi_{-i})] - u_i(\pi).$$

Thus, bounding full regret at the start of the game ensures hindsight rationality. We achieve this by showing a recursive decomposition between full and immediate regret so that an algorithm must only minimize immediate regret at each information set to be hindsight rational. Lemma 1 is the key observation required for this decomposition.

Lemma 1. *Given strategy profile π and behavioral deviation ϕ, consider $\phi_{I,g}(\pi_i)$, the strategy for player i that applies ϕ at information set I assuming memory state g and thereafter follows π_i. The regret for re-correlating after I and g—that is, the difference between the counterfactual value of $\phi(\pi_i)$ and $\phi_{I,g}(\pi_i)$ from I and g, weighted by the probability of g—is equal to the sum of full regrets at I's and g's children,* i.e.*,*

$$\hat{v}_{I,g}(\phi; \pi) - \hat{v}_{I,g}(\phi_{\preceq I, \sqsubseteq g}; \pi) = \sum_{\substack{a' \in \mathcal{A}(I), \\ I' \in \mathcal{I}_i(I, a'), \\ b \in \{*\} \cup \mathcal{A}(I)}} \rho_{I', gb}(\phi; \pi).$$

Proof.

$$\hat{v}_{I,g}(\phi; \pi) - \hat{v}_{I,g}(\phi_{\preceq I, \sqsubseteq g}; \pi) \tag{41}$$

$$= w_\phi(I, g; \pi_i) \sum_{a \in \mathcal{A}(I)} \pi_i(a \mid I) \left(r(I, \phi_{I,g}(a); \pi_{-i}) + \sum_{I' \in \mathcal{I}_i(I, \phi_{I,g}(a))} \hat{v}_{I', go(a; \phi_{I,g})}(\phi; \pi) \right) \tag{42}$$

$$- \pi_i(a \mid I) \left(r(I, \phi_{I,g}(a); \pi_{-i}) + \sum_{I' \in \mathcal{I}_i(I, \phi_{I,g}(a))} v_{I'}(\pi) \right)$$

$$= w_\phi(I, g; \pi_i) \sum_{a \in \mathcal{A}(I)} \pi_i(a \mid I) \sum_{I' \in \mathcal{I}_i(I, \phi_{I,g}(a))} \hat{v}_{I', go(a; \phi_{I,g})}(\phi; \pi) - v_{I'}(\pi). \tag{43}$$

Let $\Delta\hat{v}_{I',gb} = \hat{v}_{I',gb}(\phi;\pi) - v_{I'}(\pi)$, then,

$$= w_\phi(I,g;\pi_i) \sum_{a\in\mathcal{A}(I)} \pi_i(a\,|\,I) \sum_{I'\in\mathcal{I}_i(I,\phi_{I,g}(a))} \Delta\hat{v}_{I',go(a;\phi_{I,g})} \tag{44}$$

$$= \sum_{\substack{a'\in\mathcal{A}(I),\\ b\in\{*\}\cup\mathcal{A}(I)}} w_\phi(I,g;\pi_i) \sum_{a\in\mathcal{A}(I)} \pi_i(a\,|\,I) \sum_{I'\in\mathcal{I}_i(I,a')} \mathbb{1}\{\phi_{I,g}(a) = a' \wedge o(a;\phi_{I,g}) = b\}\Delta\hat{v}_{I',gb} \tag{45}$$

$$= \sum_{\substack{a'\in\mathcal{A}(I),\\ I'\in\mathcal{I}_i(I,a'),\\ b\in\{*\}\cup\mathcal{A}(I)}} \Delta\hat{v}_{I',gb} \underbrace{w_\phi(I,g;\pi_i) \sum_{a\in\mathcal{A}(I)} \pi_i(a\,|\,I)\mathbb{1}\{\phi_{I,g}(a) = a' \wedge o(a;\phi_{I,g}) = b\}}_{w_\phi(I',gb;\pi_i)} \tag{46}$$

$$= \sum_{\substack{a'\in\mathcal{A}(I),\\ I'\in\mathcal{I}_i(I,a'),\\ b\in\{*\}\cup\mathcal{A}(I)}} \rho_{I',gb}(\phi;\pi), \tag{47}$$

as required. □

The following two corollaries will help us to present a simple regret bound for EFR.

Corollary 3. *Lemma 1 has three cases:*

$$\hat{v}_{I,g}(\phi;\pi) - \hat{v}_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi)$$
$$= \begin{cases} \sum_{I'\in\bigcup_{a\in\mathcal{A}(I)}\mathcal{I}_i(I,a)} \rho_{I',ga}(\phi;\pi) & \text{if } \phi_{I,g} = \phi^1 \\ \sum_{I'\in\mathcal{I}_i(I,a\quad)} \rho_{I',g*}(\phi;\pi) & \text{if } \exists a\quad \in\mathcal{A}(I), \phi_{I,g} = \phi^{\to a} \\ \sum_{I'\in\mathcal{I}_i(I,a\quad)} \rho_{I',ga'}(\phi;\pi) & \\ + \sum_{I'\in\bigcup_{a\neq a'}\mathcal{I}_i(I,a)} \rho_{I',ga}(\phi;\pi) & \text{if } \exists a' \neq a\quad \in\mathcal{A}(I), \phi_{I,g} = \phi^{a'\to a}\quad . \end{cases}$$

Proof. Case 1: assume that $\phi_{I,g} = \phi^1$ (the identity transformation), then $\phi_{I,g}(a) = a$ and $o(a;\phi_{I,g}) = a$ for all $a \in \mathcal{A}(I)$. Thus,

$$\hat{v}_{I,g}(\phi;\pi) - \hat{v}_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi) = \sum_{I'\in\bigcup_{a\in\mathcal{A}(I)}\mathcal{I}_i(I,a)} \rho_{I',ga}(\phi;\pi), \tag{48}$$

as required.

Case 2: assume that $\phi_{I,g} = \phi^{\to a}\quad$ (an external transformation), then $\phi_{I,g}(a) = a\quad$ and $o(a;\phi_{I,g}) = *$ for all $a \in \mathcal{A}(I)$. Thus,

$$\hat{v}_{I,g}(\phi;\pi) - \hat{v}_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi) = \sum_{I'\in\mathcal{I}_i(I,a\quad)} \rho_{I',g*}(\phi;\pi), \tag{49}$$

as required.

Case 3: assume that $\phi_{I,g} = \phi^{a'\to a}\quad, a' \neq a\quad$ (a non-identity internal transformation), then

$$\hat{v}_{I,g}(\phi;\pi) - \hat{v}_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi) = \sum_{I'\in\mathcal{I}_i(I,a\quad)} \rho_{I,ga'}(\phi;\pi) + \sum_{I'\in\bigcup_{a\neq a'}\mathcal{I}_i(I,a)} \rho_{I,ga}(\phi;\pi), \tag{50}$$

as required. □

Corollary 4. *If the full regret following information set I and memory state g is always bounded by $C \geq 0$, then the regret for re-correlating after I and g is bounded as*

$$\hat{v}_{I,g}(\phi;\pi) - \hat{v}_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi) \leq \left(1 + \mathbb{1}\left\{\phi_{I,g} \in \Phi^{\text{IN}}_{\mathcal{A}(I)} \setminus \{\phi^1\}\right\}\right) \bigcup_{a\in\mathcal{A}(I)} \mathcal{I}_i(I,a)\ C$$

Proof. Case 1: assume that $\phi_{I,g} = \phi^1$ (the identity transformation), then

$$\hat{v}_{I,g}(\phi;\pi) - \hat{v}_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi) = \sum_{I' \in \bigcup_{a\in\mathcal{A}(I)} \mathcal{I}_i(I,a)} \rho_{I',ga}(\phi;\pi) \tag{51}$$

$$\leq \bigcup_{a\in\mathcal{A}(I)} \mathcal{I}_i(I,a)\ C, \tag{52}$$

as required.

Case 2: assume that $\phi_{I,g} = \phi^{\to a}$ (an external transformation), then

$$\hat{v}_{I,g}(\phi;\pi) - \hat{v}_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi) = \sum_{I' \in \mathcal{I}_i(I,a\ )} \rho_{I',g*}(\phi;\pi) \tag{53}$$

$$\leq \bigcup_{a\in\mathcal{A}(I)} \mathcal{I}_i(I,a)\ C, \tag{54}$$

as required.

Case 3: assume that $\phi_{I,g} = \phi^{a' \to a}$, $a' \neq a$ (an internal transformation), then

$$\hat{v}_{I,g}(\phi;\pi) - \hat{v}_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi) = \sum_{I' \in \mathcal{I}_i(I,a\ )} \rho_{I,ga'}(\phi;\pi) + \sum_{I' \in \bigcup_{a\neq a'} \mathcal{I}_i(I,a)} \rho_{I,ga}(\phi;\pi) \tag{55}$$

$$\leq\ \mathcal{I}_i(I,a\ )\ C + \bigcup_{a\neq a'} \mathcal{I}_i(I,a)\ C \tag{56}$$

$$\leq 2 \bigcup_{a\in\mathcal{A}(I)} \mathcal{I}_i(I,a)\ C. \tag{57}$$

as required. □

We can now state our decomposition result:

Lemma 2. *The full regret of behavioral deviation ϕ from information set I and memory state g is bounded by the immediate regret at I plus the full regret at each of I's and g's children.*

Proof.

$$\rho_{I,g}(\phi;\pi) = v_{I,g}(\phi;\pi) - v_I(\pi). \tag{58}$$

Adding and subtracting $v_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi)$,

$$= \underbrace{v_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi) - v_I(\pi)}_{\text{Immediate regret.}} + \underbrace{v_{I,g}(\phi;\pi) - v_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi)}_{\text{Regret for re-correlating after } I \text{ and } g.} \tag{59}$$

Applying Lemma 1,

$$= \rho_{I,g}(\phi_{\preceq I,\sqsubseteq g};\pi) + \sum_{\substack{a'\in\mathcal{A}(I),\\ I'\in\mathcal{I}_i(I,a'),\\ b\in\{*\}\cup\mathcal{A}(I)}} \rho_{I',gb}(\phi;\pi), \tag{60}$$

as required. □

Corollary 5. *If the full regret of each child of information set I is bounded by $C \geq 0$, then*

$$\rho_{I,g}(\phi;\pi) \leq \rho_I(\phi_{\preceq I,\sqsubseteq g}) + \left(1 + \mathbb{1}\left\{\phi_{I,g} \in \Phi^{\text{IN}}_{\mathcal{A}(I)} \setminus \{\phi^1\}\right\}\right) \bigcup_{a\in\mathcal{A}(I)} \mathcal{I}_i(I,a)\ C.$$

Proof. Lemma 1 and Corollary 4. □

Using Corollary 5 and instantiating EFR with exact regret matching, we can derive a simple regret bound that depends on the number of immediate regrets associated with a given subset of behavioral deviations.

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 $D = \max_{I \in \mathcal{I}_i, \phi_I \in \Phi_I} |W_I^{\Phi}(\phi_I)| \omega(\Phi_I)$. Then, EFR's cumulative regret after T rounds with respect to Φ is upper bounded by $2^{n_{\text{IN}}+1} U |\mathcal{I}_i| \sqrt{DT}$.*

Proof. EFR keeps track of each immediate regret for each transformation associated with each realizable memory state g in each information set I. EFR's immediate strategies at each I on each round are chosen according to time selection regret matching on the cumulative immediate regrets and memory state probabilities there. Therefore, the cumulative immediate regret at each information set and memory state is bounded as $\sum_{t=1}^{T} \rho_I(\phi_{\preceq I, \sqsubseteq g}) \leq 2U\sqrt{DT}$ according to Corollary 1. Working from the leaves of the information set tree to the roots, we recursively bound the full regret according to Corollary 5. The full regret at each information set is then bounded as $\sum_{t=1}^{T} \rho_{I,g}(\phi; \pi^t) \leq 2^{n_{\text{IN}}+1} U |\mathcal{I}_i| \sqrt{DT}$. EFR's cumulative regret with respect to Φ is equal to its cumulative full regret at the start of the game, so the former is bounded by $2^{n_{\text{IN}}+1} U |\mathcal{I}_i| \sqrt{DT}$ as well, which concludes the argument. □

See Table C.2 for EFR parameters each deviation type.

The variable D in the EFR regret bound that depends on the particular behavioral deviation subset with which it is instantiated is often the number of immediate regrets generated by that subset divided by the number of information sets. D is slightly larger for CSPS because it uses the union of internal and external transformations for its action transformation set, Φ_I, at all information sets except those at the beginning of the game. Since our bound depends on the maximum number of memory states associated with any $\phi \in \Phi_I$ and $\omega(\Phi_I)$ counts the maximum number of non-trivial ways an action can be transformed according to the transformations in Φ_I, their product ends up being larger for CSPS than the number of valid combinations between memory states and action transformations. See Table C.2 for D values corresponding to each partial sequence deviation type.

D Regret Matching++

Kash, Sullins, and Hofmann (2020) presents the regret matching++ algorithm and claims that it is no-external-regret. This algorithm's proposed regret bound implies a sublinear bound on cumulative positive regret, which would further imply that it has the same bound with respect to *all possible* time selection functions. The surprising aspect of this result is that the algorithm does not require any information about any of the possible time selection functions and requires no more computation or storage than basic regret matching. The following result, Theorem 3, shows that there is actually no algorithm that can achieve a sublinear bound on cumulative positive regret. This result proves that regret matching++ cannot be no-external-regret as claimed. Appendix D.2 identifies the mistake in the regret matching++ bound proof.

D.1 Linear Lower Bound on the Sum of Positive Regrets

Theorem 3. *The worst-case maximum cumulative positive regret, $Q^T = \max_{a \in \mathcal{A}} \sum_{t=1}^{T} (r^t(a) - \langle \pi^t, r^t \rangle)^+$, under a sequence of reward functions chosen from the class of bounded reward functions, $(r^t \in \{r : r \in \mathbb{R}^{|\mathcal{A}|}, \|r\|_\infty \leq 1\})_{t=1}^{T}$, of any algorithm that chooses policies $\pi^t \in \Delta^{|\mathcal{A}|}$ over a finite set of actions, $\mathcal{A}$, in an online fashion over T rounds, is at least $T/4$.*

Proof. Without loss of generality, consider a two action environment, $\mathcal{A} = (a, a')$, and any learning algorithm that deterministically chooses a distribution, $\pi^t \in \Delta^2$, over them on each round t. The environment gets to see the learner's policy before presenting a reward function. If the learner weights one action more than the other, the environment gives a reward of zero for the action with the larger

Table C.1: Formal definition of the strategy generated by a deviation of each type given pure strategy $s_i \in S_i$ at each information set $I \in \mathcal{I}_i$.

behavioral	$\forall I', \exists a^!_{I'}, a'_{I'}$, $\begin{cases} a'_I & \text{if } \forall \bar{I} \preceq I, s_i(\bar{I}) = a^!_{\bar{I}} \\ s_i(I) & \text{o.w.} \end{cases}$	in. causal	$\exists I^!, a^!, s'$, $\begin{cases} s'(I) & \text{if } I \succeq I^!, s_i(I^!) = a^! \\ s_i(I) & \text{o.w.} \end{cases}$
TIPS	$\exists I^!, a^!, I\ , a\ , a\ ^!$, $\begin{cases} a & \text{if } I = I\ , s_i(I\ ) = a\ ^!, \\ & s_i(I^!) = a^! \\ a_I^{\rightarrow I} & \text{if } I \succeq I^!, s_i(I^!) = a^! \\ s_i(I) & \text{o.w.} \end{cases}$	in. action	$\exists I^!, a\ , a^!$, $\begin{cases} a & \text{if } I{=}I^!, s_i(I^!){=}a^! \\ s_i(I) & \text{o.w.} \end{cases}$
CSPS	$\exists I^!, a^!, I\ , a\ ,$ $\begin{cases} a & \text{if } I = I\ , s_i(I^!) = a^! \\ a_I^{\rightarrow I} & \text{if } I \succeq I^!, s_i(I^!) = a^! \\ s_i(I) & \text{o.w.} \end{cases}$	in. CF	$\exists I\ , a\ , a\ ^!$, $\begin{cases} a & \text{if } I{=}I\ , s_i(I\ ){=}a\ ^! \\ a_I^{\rightarrow I} & \text{if } I \preceq I \\ s_i(I) & \text{o.w.} \end{cases}$
CFPS	$\exists I^!, I\ , a\ , a\ ^!$, $\begin{cases} a & \text{if } I = I\ , s_i(I\ ) = a\ ^! \\ a_I^{\rightarrow I} & \text{if } I \succeq I^! \\ s_i(I) & \text{o.w.} \end{cases}$	blind causal	$\exists I^!, s'$, $\begin{cases} s'(I) & \text{if } I \succeq I^! \\ s_i(I) & \text{o.w.} \end{cases}$
BPS	$\exists I^!, I\ , a\ ,$ $\begin{cases} a & \text{if } I = I \\ a_I^{\rightarrow I} & \text{if } I \succeq I^! \\ s_i(I) & \text{o.w.} \end{cases}$	blind action	$\exists I^!, a\ ,$ $\begin{cases} a & \text{if } I{=}I^! \\ s_i(I) & \text{o.w.} \end{cases}$
		blind CF	$\exists I\ , a\ ,$ $\begin{cases} a & \text{if } I{=}I \\ a_I^{\rightarrow I} & \text{if } I \preceq I \\ s_i(I) & \text{o.w.} \end{cases}$

weight and one to the action with the smaller weight. Formally, if $\pi^t(a) \geq 0.5$, then $r^t(a) = 0$, $r^t(a') = 1$, and vice-versa otherwise.

Let $a_{\text{low}} = a'$ if $\pi^t(a) \geq 0.5$ and $a_{\text{low}} = a$ otherwise. The positive regrets on any round t are $(1 - \pi^t(a_{\text{low}}))^+ \geq 0.5$ and $(0 - (1 - \pi^t(a_{\text{low}})))^+ = 0$. So the learner is forced to suffer at least 0.5 positive regret on each round for one of the actions. Since there are only two actions, then over T rounds one of the actions must have accumulated a regret of 0.5 on at least $T/2$ rounds. The cumulative positive regret for this action must then be $T/4$. Therefore, the maximum cumulative positive regret of any deterministic algorithm in this environment must be at least $T/4$.

To extend this result to include algorithms that stochastically choose π^t, we simply need to consider the expected cumulative positive regret and notice that the rectified linear unit function $(\cdot^+)$ is convex. By Jensen's inequality and the fact that the max of an expectation is no larger than the expectation of the max, the expected cumulative positive regret is lower bounded by the cumulative positive regret under the learner's expected distributions, $\mathbb{E}[\pi^t]$, *i.e.*, $\mathbb{E}[Q^T] \geq \max_{a \in \mathcal{A}} \sum_{t=1}^{T} (r^t(a) - \langle \mathbb{E}[\pi^t], r^t \rangle)^+ \geq T/4$. Since $\mathbb{E}[\pi^t]$ is a single distribution, we have reduced the stochastic case to the deterministic case, thereby showing they have the same regret lower bound. □

Table C.2: EFR parameters and regret bound constants for different deviation types.

type	Φ_I for all $I \in \mathcal{I}_i$	W_I^Φ as function of ϕ_I for all $I \in \mathcal{I}_i$	$\max_{I \in \mathcal{I}_i, \phi_I \in \Phi_I} W_I^\Phi(\phi_I)$	D	n_{IN}
behavioral	$\Phi^{\text{IN}}_{\mathcal{A}(I)} \setminus \{\phi^1\}$	$\{t \mapsto 1\} \cup \left\{t \mapsto \prod_{\bar{I} \preceq \bar{I}'} \pi_i^t(a_{\bar{I}} \mid \bar{I})\right\}_{\bar{I}' \prec I,\, \forall \bar{I} \preceq \bar{I}', a_{\bar{I}} \in \mathcal{A}(\bar{I})}$	$n_{\mathcal{A}}^{d_*}$	$n_{\mathcal{A}}^{d_*}(n_{\mathcal{A}}^2 - n_{\mathcal{A}})$	d_*
TIPS	$\Phi^{\text{IN}}_{\mathcal{A}(I)} \setminus \{\phi^1\}$	$\{t \mapsto 1\} \cup \left\{t \mapsto P(\mathbb{h}(I'); \pi_i^t)\pi_i^t(a' \mid I')\right\}_{I' \prec I, a' \in \mathcal{A}(I')}$	$d_* n_{\mathcal{A}} + 1$	$(d_* n_{\mathcal{A}} + 1)(n_{\mathcal{A}}^2 - n_{\mathcal{A}})$	1
CSPS / in. causal	$\Phi^{\text{EX}}_{\mathcal{A}(I)} \cup \Phi^{\text{IN}}_{\mathcal{A}(I)} \setminus \{\phi^1\}$	$\begin{cases} \{t \mapsto 1\} \cup \\ \left\{t \mapsto P(\mathbb{h}(I'); \pi_i^t)\pi_i^t(a' \mid I')\right\}_{I' \prec I, a' \in \mathcal{A}(I')} & \text{if } \phi_I \in \Phi^{\text{EX}}_{\mathcal{A}(I)} \\ \left\{t \mapsto P(\mathbb{h}(I); \pi_i^t)\right\} & \text{o.w.} \end{cases}$	$d_* n_{\mathcal{A}}$	$d_* n_{\mathcal{A}}(n_{\mathcal{A}}^2 - 2)$	1
CFPS	$\Phi^{\text{IN}}_{\mathcal{A}(I)} \setminus \{\phi^1\}$	$\{t \mapsto 1\} \cup \left\{t \mapsto P(\mathbb{h}(I'); \pi_i^t)\right\}_{I' \preceq I}$	$d_* + 1$	$(d_* + 1)(n_{\mathcal{A}}^2 - n_{\mathcal{A}})$	0
BPS / blind causal	$\Phi^{\text{EX}}_{\mathcal{A}(I)}$	$\{t \mapsto 1\} \cup \left\{t \mapsto P(\mathbb{h}(I'); \pi_i^t)\right\}_{I' \preceq I}$	$d_* + 1$	$(d_* + 1)(n_{\mathcal{A}} - 1)$	0
in. action	$\Phi^{\text{IN}}_{\mathcal{A}(I)} \setminus \{\phi^1\}$	$\left\{t \mapsto P(\mathbb{h}(I); \pi_i^t)\right\}$	1	$n_{\mathcal{A}}^2 - n_{\mathcal{A}}$	0
in. CF	$\Phi^{\text{IN}}_{\mathcal{A}(I)} \setminus \{\phi^1\}$	$\{t \mapsto 1\}$	1	$n_{\mathcal{A}}^2 - n_{\mathcal{A}}$	0
blind action	$\Phi^{\text{EX}}_{\mathcal{A}(I)}$	$\left\{t \mapsto P(\mathbb{h}(I); \pi_i^t)\right\}$	1	$n_{\mathcal{A}} - 1$	0
blind CF / external	$\Phi^{\text{EX}}_{\mathcal{A}(I)}$	$\{t \mapsto 1\}$	1	$n_{\mathcal{A}} - 1$	0

D.2 Regret Matching++ Regret Bound Bug

Define the cumulative positive regret of action $a \in \mathcal{A}$ as $Q_a^T = \sum_{t=1}^T (\rho_a^t)^+ = \sum_{t=1}^T (r_a^t - \langle \pi^t, r^t \rangle)^+$, where r^t is the reward function on round t. Kash, Sullins, and Hofmann (2020) bounds $(\max_a Q_a^T)^2 \leq \sum_a (Q_a^T)^2 = \sum_a \left(Q_a^{T-1} + (\rho_a^T)^+\right)^2$. They then state that $\left(Q_a^{T-1} + (\rho_a^T)^+\right)^2 \leq \left(Q_a^{T-1} + \rho_a^T\right)^2 + \Delta^2$, where Δ is the diameter of the reward range. This is false in general:

$$\begin{aligned} \left(Q_a^{T-1} + (\rho_a^T)^+\right)^2 &= \left(Q_a^{T-1}\right)^2 + \left(\rho_a^T\right)^2 + 2Q_a^{T-1}(\rho_a^T)^+ && (61) \\ &\leq \left(Q_a^{T-1}\right)^2 + \left(\rho_a^T\right)^2 + 2Q_a^{T-1}\rho_a^T + 2Q_a^{T-1}\Delta && (62) \\ &= \left(Q_a^{T-1} + \rho_a^T\right)^2 + 2Q_a^{T-1}\Delta, && (63) \end{aligned}$$

where $2Q_a^{T-1}\Delta > \Delta^2$ if $Q_a^{T-1} > \Delta/2$. There are scenarios where Equation (63) is tight so it is unclear how this bound could be improved. Attempting the rest of the proof, we get

$$\sum_a \left(Q_a^{T-1} + (\rho_a^T)^+\right)^2 \leq |\mathcal{A}|\Delta^2 + \sum_a \left(Q_a^{T-1}\right)^2 + 2\Delta \sum_a Q_a^{T-1}.$$

Unrolling the recursion exactly is messy, but the extra $2\Delta \sum_a Q_a^{T-1}$ term ensures that the bound will be no smaller than $\sum_a Q_a^{T-1} + (\rho_a^T)^+ \leq 2^{\frac{T}{2}} |\mathcal{A}|^{\frac{T-1}{2}} \Delta^T$.

E Experiments

E.1 Games

Leduc Hold'em Poker

Leduc hold'em poker (Southey et al. 2005) is a two-player poker game with a deck of six cards (two suits and three ranks). At the start of the game, both players ante one chip and receive one private card. There are two betting rounds and there is a maximum of two raises on each round. Bet sizes are limited to two chips in the first round and four in the second. If one player folds, the other wins. At the start of the second round, a public card is revealed. A showdown occurs at the end of the second round if no player folds. The strongest hand in a showdown is a pair (using the public card), and if no player pairs, players compare the ranks of their private cards. The player with the stronger hand takes all chips in the pot or players split the pot if their hands have the same strength. Payoffs are reported in milli-big blinds (mbb) (where the ante is considered a big blind) for consistency with the way performance is reported in other poker games.

Imperfect Information Goofspiel

Imperfect information goofspiel (Ross 1971; Lanctot 2013) is a bidding game for N players. Each player is given a hand of n ranks that they play to bid on n point cards. On each round, one point card is revealed and each player simultaneously bids on the point card. The point cards might be sorted in ascending order ($\uparrow$), descending order ($\downarrow$), or they might be shuffled (R). If there is one bid that is greater than all the others, the player who made that bid wins the point card. If there is a draw, the bid card is instead discarded. The player with the most points wins so payoffs are reported in win percentage. We use five goofspiel variants:

- two-player, 5-ranks, ascending (goofspiel$(5, \uparrow, N = 2)$, denoted as $g_{2,5,\uparrow}$ in the main paper),
- two-player, 5-ranks, descending (goofspiel$(5, \downarrow, N = 2)$),
- two-player, 4-ranks, random (goofspiel$(4, R, N = 2)$),
- three-player, 4-ranks, ascending (goofspiel$(4, \uparrow, N = 3)$, denoted as $g_{3,4,\uparrow}$ in the main paper), and
- three-player, 4-ranks, descending (goofspiel$(4, \downarrow, N = 3)$).

Sheriff

Sheriff is a two-player, non-zero-sum negotiation game resembling the Sheriff of Nottingham board game and it was introduced by Farina et al. (2019). At the beginning of the game, the "smuggler" player chooses zero or more illegal items (maximum of three) to add to their cargo. The rest of the game proceeds over four rounds.

At the beginning of each round, the smuggler signals how much they would be willing to pay the "sheriff" player to bribe them into not inspecting the smuggler's cargo, between zero and three. The sheriff responds by signalling whether or not they would inspect the cargo. On the last round, the bribe amount chosen by the smuggler and the sheriff's decision about whether or not to inspect the cargo are binding.

If the cargo is not inspected, then the smuggler receives a payoff equal to the number of illegal items included within, minus their bribe amount, and the sheriff receives the bribe amount. Otherwise, the sheriff inspects the cargo. If the sheriff finds an illegal item, then the sheriff forces the smuggler to pay them two times the number of illegal items. Otherwise, the sheriff compensates the smuggler by paying them three.

Tiny Bridge

A miniature version of bridge created by Edward Lockhart, inspired by a research project at University of Alberta by Michael Bowling, Kate Davison, and Nathan Sturtevant. We use the smaller two-player rather than the full four-player version. See the implementation from Lanctot et al. (2019) for more details.

Tinay Hanabi

A miniature two-player version of Hanabi described by Foerster et al. (2019). The game is fully cooperative and the optimal score is ten. Both players take only one action so all EFR instances collapse except when they differ in their choice of Φ_I.

E.2 Alternative Φ_I Choices

When implementing EFR for deviations that set the action transformations at each information set to the internal transformations, we have the option of implementing these variants by using the union of the internal and external transformations without substantially changing the variant's theoretical properties. We test how this impacts practical performance within EFR variants for informed counterfactual deviations, CFPS deviations, and TIPS deviations. These variants have an "EX+ IN" subscript.

E.3 Results

We present four sets of figures to summarize the performance of each EFR variant in the fixed and simultaneous regimes described in Section 7.

The first three sets of figures illustrate how each variant performs on average in each round individually. Figures E.1 and E.3 show the running average expected payoff of each variant over rounds, averaged over play with all EFR variants (including itself). These figures summarize the progress that each variant makes over rounds to adapt to and correlate with its companion variant, on average. Figures E.2 and E.4 show the instantaneous expected payoff of each variant over rounds, averaged over play with all EFR variants. Figures E.5 and E.6 show the same data as in Figures E.1 and E.3 except according to runtime rather than rounds. Tiny Hanabi is omitted because it is too small to make meaningful runtime comparisons between EFR variants.

Figure E.8 show the average expected payoff of each variant paired with each other variant (including itself) after 1000 rounds. These figures summarize how well each variant works with each other variant.

References

Burch, N. 2017. *Time and space: Why imperfect information games are hard*. Ph.D. thesis, University of Alberta.

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

D'Orazio, R.; and Huang, R. 2021. Optimistic and Adaptive Lagrangian Hedging. In *Reinforcement Learning in Games Workshop at the Thirty-Fifth AAAI Conference on Artificial Intelligence*.

D'Orazio, R.; Morrill, D.; Wright, J. R.; and Bowling, M. 2020. 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.

Farina, G.; Ling, C. K.; Fang, F.; and Sandholm, T. 2019. Correlation in Extensive-Form Games: Saddle-Point Formulation and Benchmarks. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Foerster, J.; Song, F.; Hughes, E.; Burch, N.; Dunning, I.; Whiteson, S.; Botvinick, M.; and Bowling, M. 2019. Bayesian action decoder for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, 1942–1951. PMLR.

Gordon, G. J. 2005. No-regret algorithms for structured prediction problems. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.

Greenwald, A.; Li, Z.; and Marks, C. 2006. Bounds for Regret-Matching Algorithms. In *ISAIM*.

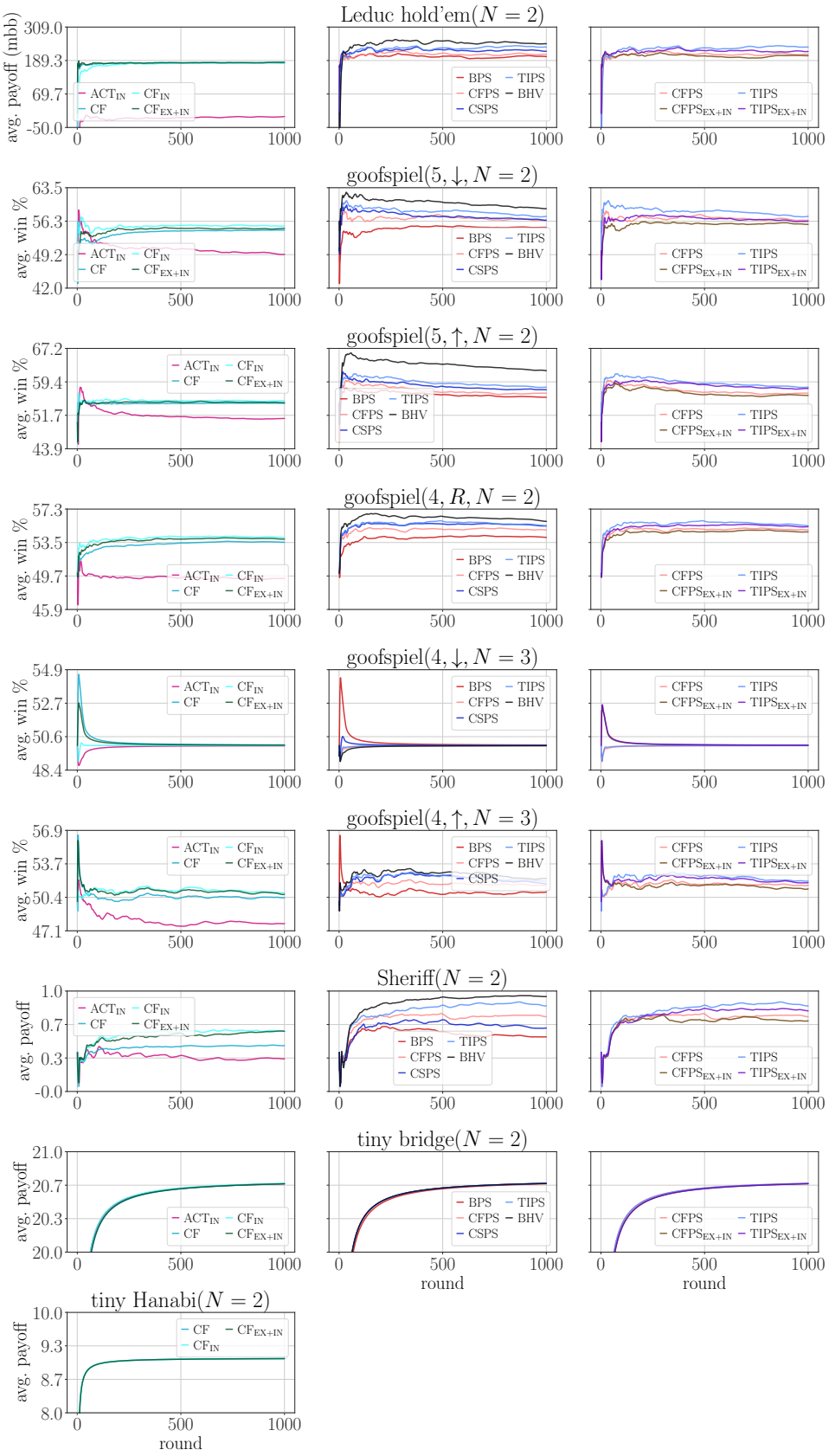

Figure E.1: The expected payoff accumulated by each EFR variant over rounds averaged over play with all EFR variants in each game in the fixed regime.

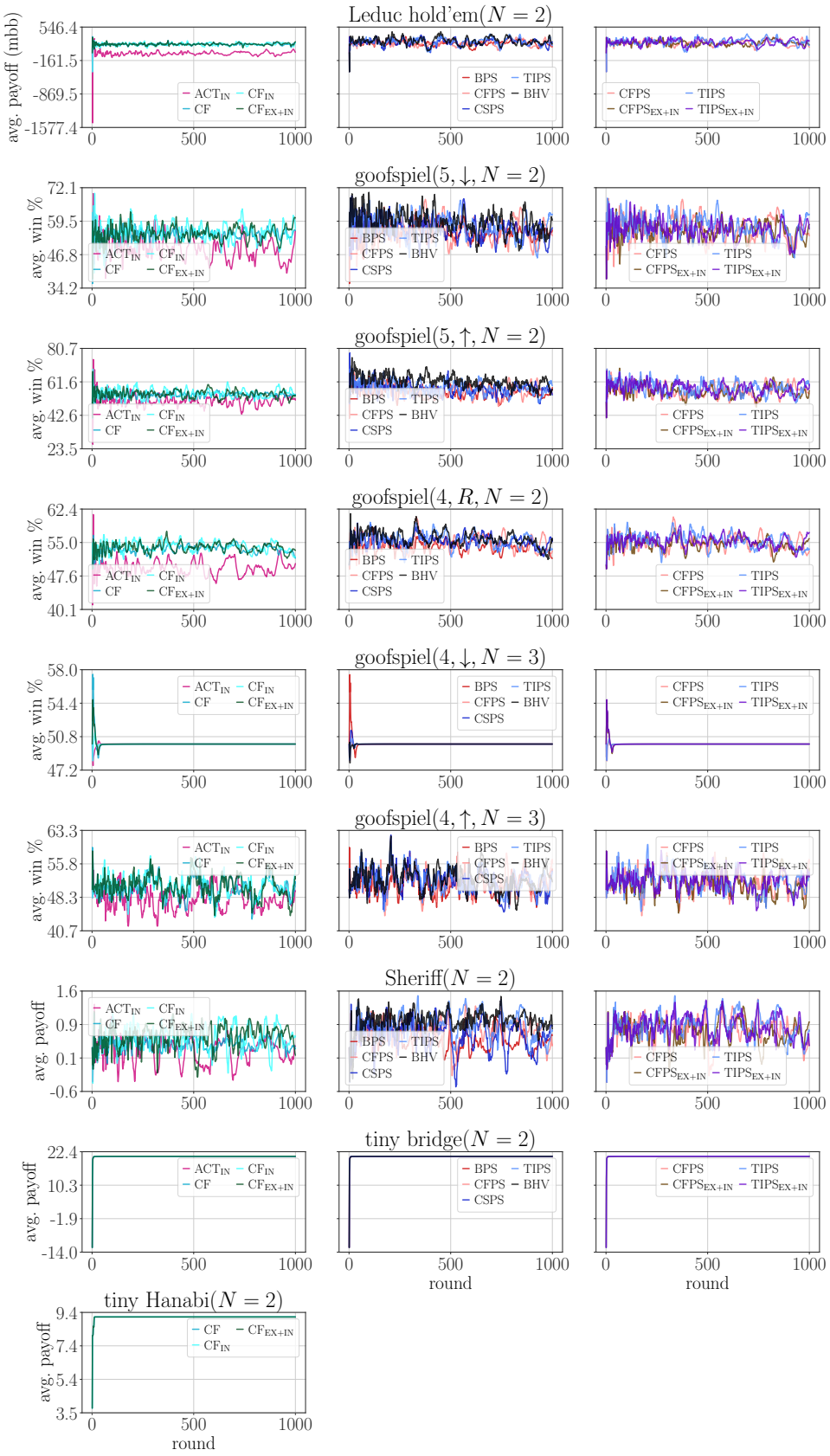

Figure E.2: The instantaneous payoff achieved by each EFR variant on each round averaged over play with all EFR variants in each game in the fixed regime.

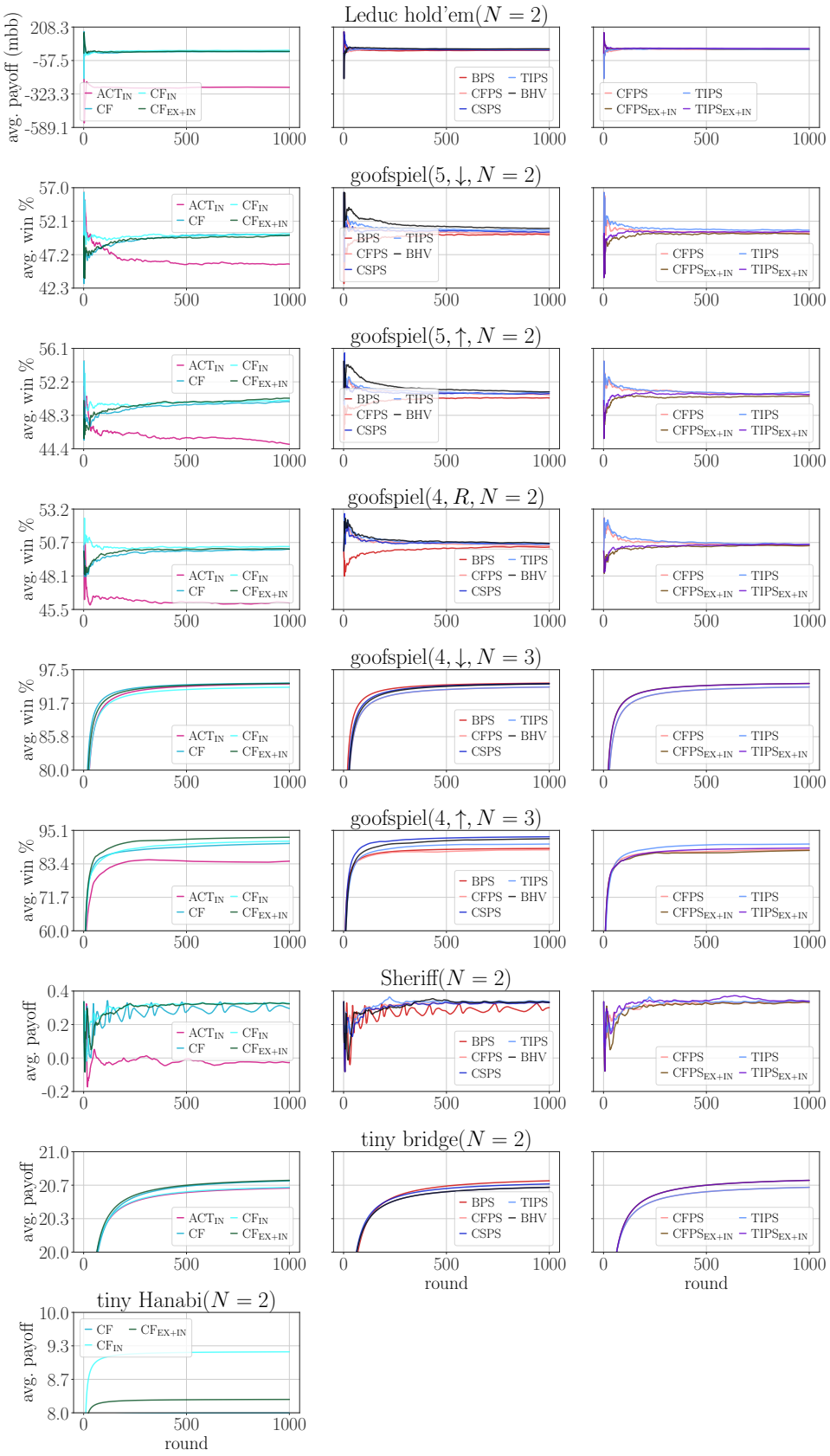

Figure E.3: The expected payoff accumulated by each EFR variant over rounds averaged over play with all EFR variants in each game in the simultaneous regime.

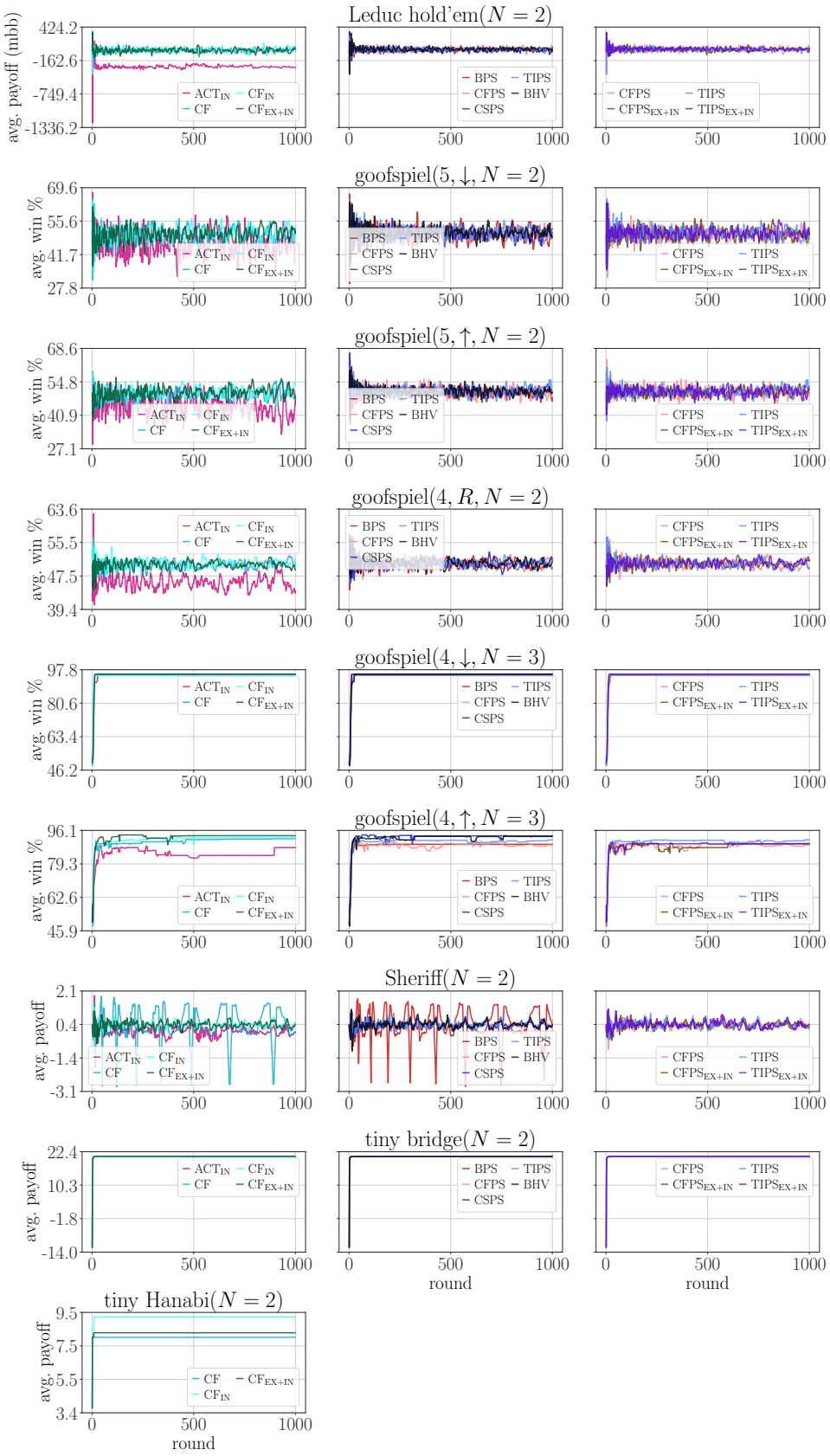

Figure E.4: The instantaneous payoff achieved by each EFR variant on each round averaged over play with all EFR variants in each game in the simultaneous regime.

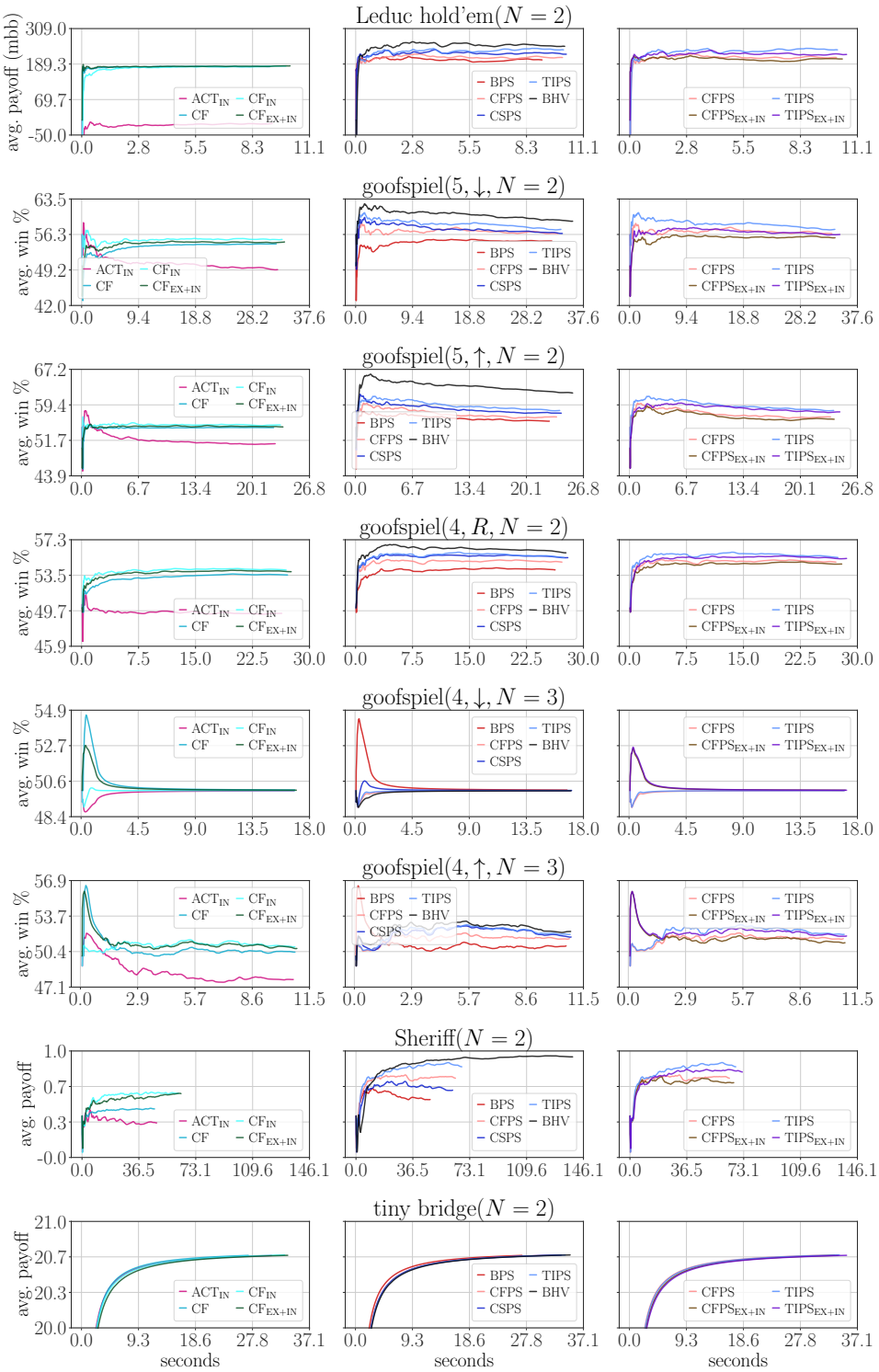

Figure E.5: The expected payoff accumulated by each EFR variant over runtime averaged over play with all EFR variants in each game in the fixed regime.

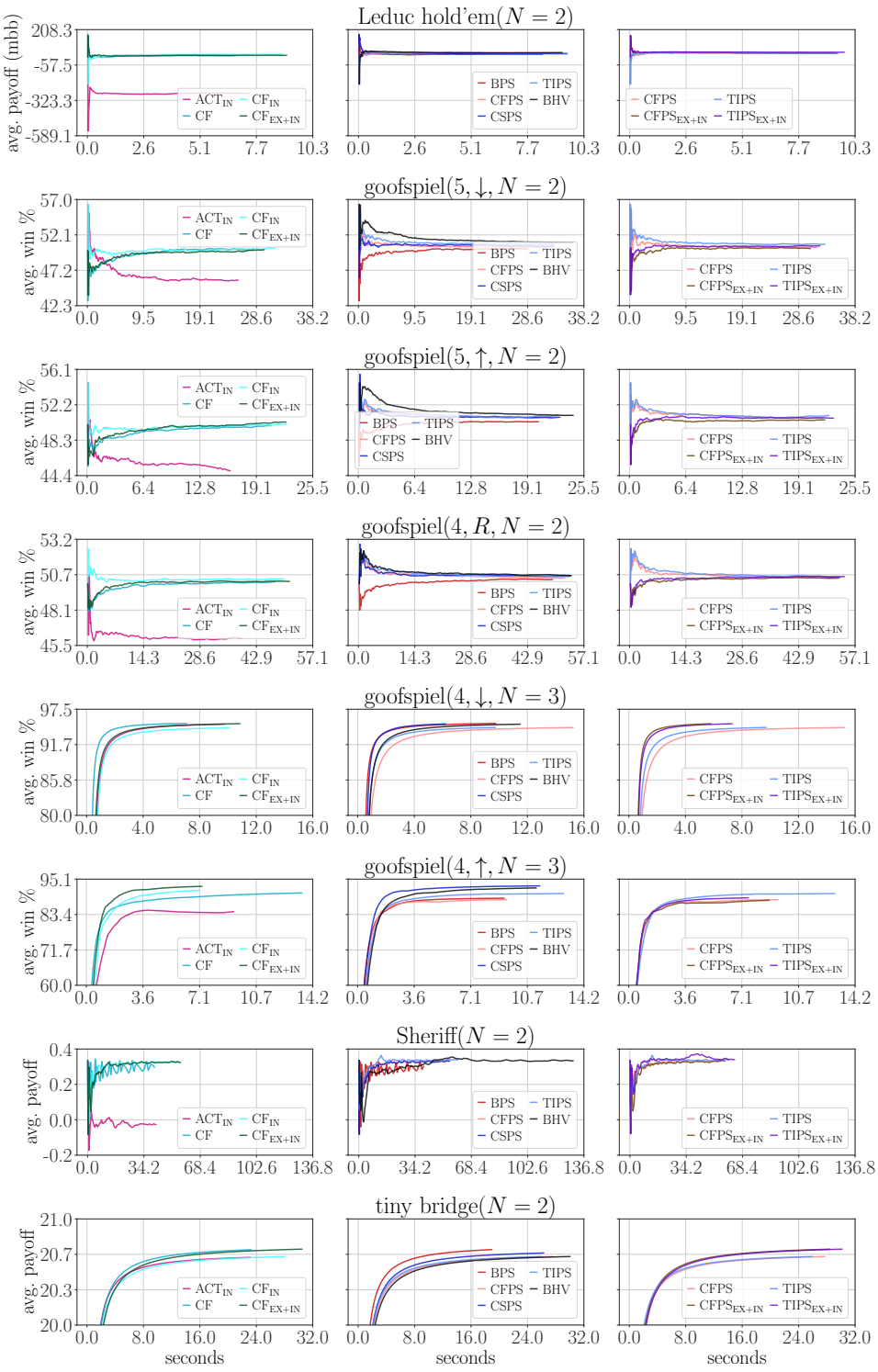

Figure E.6: The expected payoff accumulated by each EFR variant over runtime averaged over play with all EFR variants in each game in the simultaneous regime.

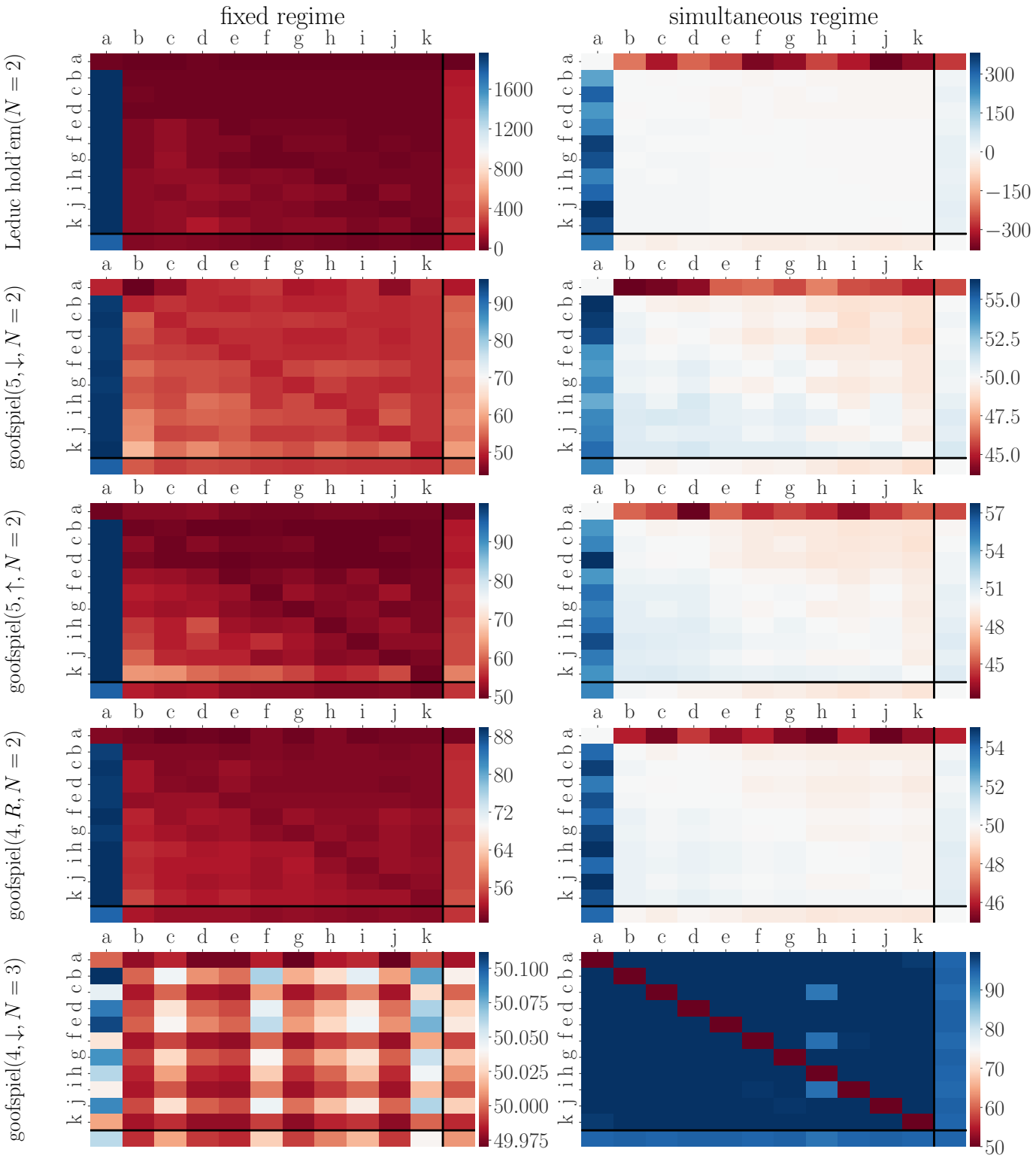

Figure E.7: (1 / 2) The average expected payoff accumulated by each EFR variant (listed by row) from playing with each other EFR variant (listed by column) in each game after 1000 rounds where a $\rightarrow$ ACT_{IN}, b $\rightarrow$ CF, c $\rightarrow$ CF_{IN}, d $\rightarrow$ $\text{CF}_{\text{EX+IN}}$, e $\rightarrow$ BPS, f $\rightarrow$ CFPS, g $\rightarrow$ $\text{CFPS}_{\text{EX+IN}}$, h $\rightarrow$ CSPS, i $\rightarrow$ TIPS, j $\rightarrow$ $\text{TIPS}_{\text{EX+IN}}$, k $\rightarrow$ BHV. The bottom rows and farthest right columns represent the column and row averages, respectively.

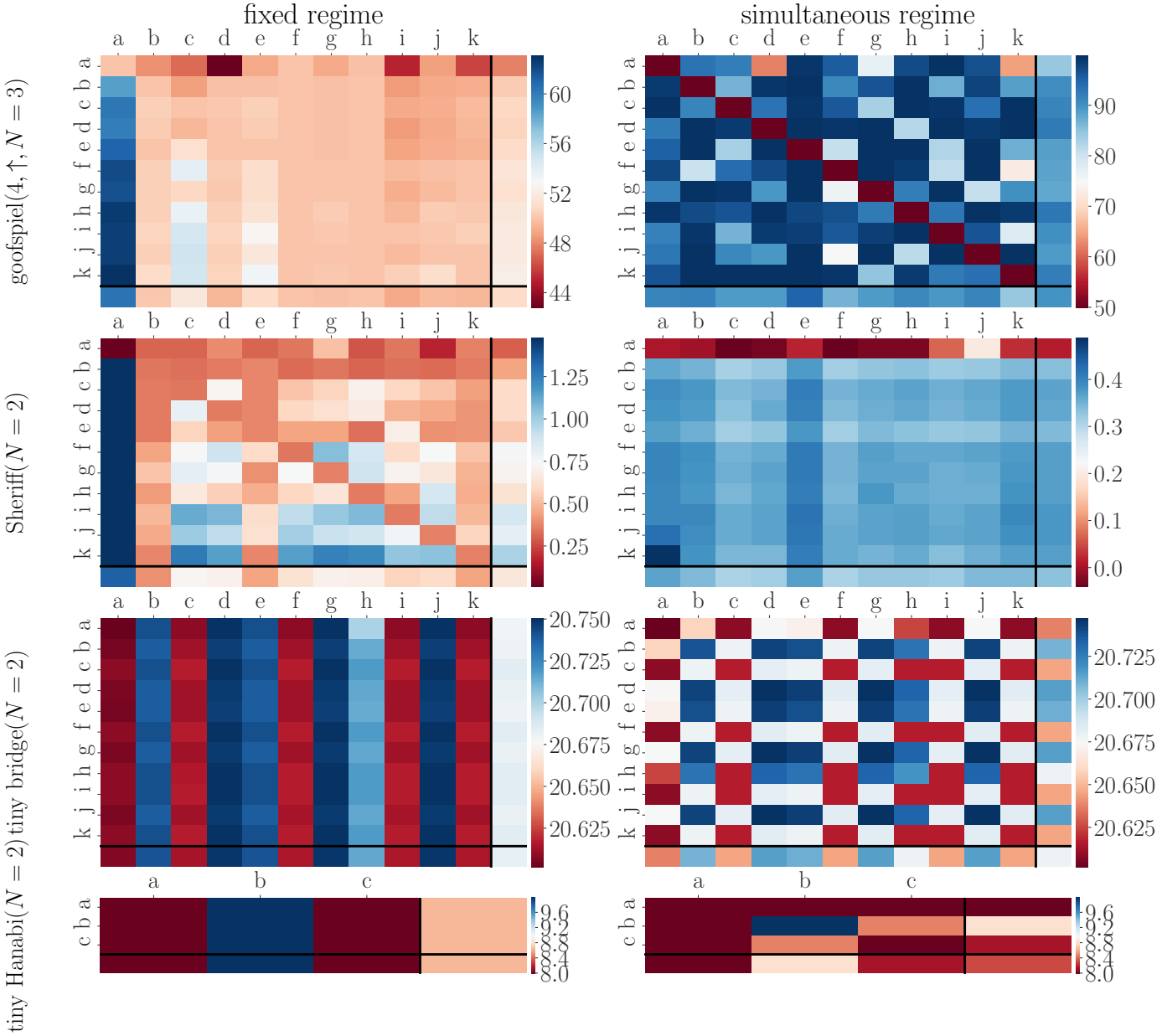

Figure E.8: (2 / 2) The average expected payoff accumulated by each EFR variant (listed by row) from playing with each other EFR variant (listed by column) in each game after 1000 rounds where a $\rightarrow$ ACT_{IN}, b $\rightarrow$ CF, c $\rightarrow$ CF_{IN}, d $\rightarrow$ $\text{CF}_{\text{EX+IN}}$, e $\rightarrow$ BPS, f $\rightarrow$ CFPS, g $\rightarrow$ $\text{CFPS}_{\text{EX+IN}}$, h $\rightarrow$ CSPS, i $\rightarrow$ TIPS, j $\rightarrow$ $\text{TIPS}_{\text{EX+IN}}$, k $\rightarrow$ BHV. The bottom rows and farthest right columns represent the column and row averages, respectively.

Hart, S.; and Mas-Colell, A. 2000. A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica* 68(5): 1127–1150.

Kash, I. A.; Sullins, M.; and Hofmann, K. 2020. 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.

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

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. 2019. OpenSpiel: A Framework for Reinforcement Learning in Games. *CoRR* abs/1908.09453. URL http://arxiv.org/abs/1908.09453.

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

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

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

Southey, F.; Bowling, M. H.; Larson, B.; Piccione, C.; Burch, N.; Billings, D.; and Rayner, D. C. 2005. Bayes' Bluff: Opponent Modelling in Poker. In *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Edinburgh, Scotland, July 26-29, 2005*, 550–558. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1216&proceeding_id=21.

Syrgkanis, V.; Agarwal, A.; Luo, H.; and Schapire, R. E. 2015. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, 2989–2997.

Tammelin, O. 2014. Solving Large Imperfect Information Games Using CFR+. *arXiv preprint arXiv:1407.5042* .

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*.

Waugh, K.; and Bagnell, J. A. 2015. A Unified View of Large-Scale Zero-Sum Equilibrium Computation. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Waugh, K.; Morrill, D.; Bagnell, J. A.; and Bowling, M. 2015. Solving Games with Functional Regret Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zinkevich, M.; Johanson, M.; Bowling, M.; and Piccione, C. 2007. Regret Minimization in Games with Incomplete Information. Technical Report TR07-14, University of Alberta.