Faster Game Solving via Initial Regret Reduction

Anonymous authors
Paper under double-blind review

Abstract

Counterfactual regret minimization (CFR) is the leading framework for solving two-player zero-sum imperfect-information games and underlies recent superhuman poker AI. A remaining source of inefficiency in CFR and CFR+ is uniform initialization: because every action is explored equally at the start, clearly suboptimal actions depress the average counterfactual value and inflate the initial positive regrets of the remaining actions, forcing many iterations to separate the viable actions that matter in equilibrium. Discounted variants such as DCFR and PDCFR+ mitigate this effect only indirectly by decaying accumulated regrets over many iterations. We propose initial regret reduction (IRR), a one-time reduction that rescales the first positive accumulated regret by a factor of at most one, preserves its direction, and leaves every later update unchanged. The factor comes from computing the regret of that round as if it had been played with a mixture of the uniform strategy and the next strategy induced by the first positive regret. We prove that the single reduction never enlarges the first positive regret, and that IRR-RM+ and IRR-PRM+ preserve the $O(1/\sqrt{T})$ no-regret rate. Combined with PCFR+, IRR-PCFR+ attains the lowest average-iterate exploitability among the compared algorithms on benchmark imperfect-information games other than the heads-up no-limit Texas hold'em subgames, where DCFR remains superior.

1 Introduction

Imperfect-information games model strategic interaction with hidden information, and counterfactual regret minimization (CFR) (Zinkevich et al., 2007) is the standard framework for computing Nash equilibria in two-player zero-sum imperfect-information games. Modern variants speed convergence in different ways: CFR+ clips negative regrets (Tammelin, 2014), Linear CFR and DCFR discount the regrets of early iterations (Brown & Sandholm, 2019), and PCFR+ and PDCFR+ add prediction to the update (Farina et al., 2021; Xu et al., 2024).

All of these variants share a default initial configuration: at every information set, the accumulated regret starts at zero and the strategy starts uniform. The accumulated regret stays zero until the *first positive regret round*, the earliest round whose instantaneous regret has a positive component, typically the first round. In this round, dominated actions pull down the average utility of the uniform strategy, pushing up the positive regrets of the other actions, all by the same amount. When this amount is large relative to the gaps among them, the solver needs many iterations to separate the actions that matter in equilibrium. We call this effect the uniform-initialization bias.

Discounting mitigates the bias indirectly: DCFR and PDCFR+ decay accumulated regret at every iteration, so the bias fades over time. We instead propose initial regret reduction (IRR), which scales down the accumulated regret at the first positive regret round with a single factor, without changing the strategies of that round or the next. To compute the factor, IRR first computes the regret of that round as if it had been played with a mixture of the uniform strategy and the next strategy induced by the first positive regret; the factor is then the ratio of the ℓ_1 norm of the positive regret under this mixture to the original one, and never exceeds one. After this round, the update is the standard PCFR+ recurrence; we call the resulting solver IRR-PCFR+.

Our contributions are as follows.

- To mitigate the slowdown caused by the uniform-initialization bias (Section 3), we propose initial regret reduction (IRR), which scales down the accumulated regret at the first positive

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Update rules of the CFR variants discussed in this paper; γ is the averaging exponent, $\odot$ the entrywise product, and $\tilde{\boldsymbol{R}}_j^{t+1}$ the predicted accumulated regret. In DCFR, the entrywise discount is $\boldsymbol{d}_j^{t-1}[a] = (t{-}1)^\alpha/((t{-}1)^\alpha+1)$ if $\boldsymbol{R}_j^{t-1}[a] > 0$ and $(t{-}1)^\beta/((t{-}1)^\beta+1)$ otherwise. In DCFR+ and PDCFR+, $d_t = (t{-}1)^\alpha/((t{-}1)^\alpha+1)$.

Algorithm	Accumulated regret $\boldsymbol{R}_j^t$	New strategy $\boldsymbol{x}_j^{t+1}$	Cumulative strategy $\boldsymbol{X}^t$
CFR	$\boldsymbol{R}_j^{t-1} + \boldsymbol{r}_j^t$	$[\boldsymbol{R}_j^t]^+/\|[\boldsymbol{R}_j^t]^+\|_1$	$\boldsymbol{X}^{t-1} + \dot{\boldsymbol{x}}^t$ ($\gamma = 0$)
CFR+	$[\boldsymbol{R}_j^{t-1} + \boldsymbol{r}_j^t]^+$	$[\boldsymbol{R}_j^t]^+/\|[\boldsymbol{R}_j^t]^+\|_1$	$\boldsymbol{X}^{t-1}(t{-}1)^\gamma/t^\gamma + \dot{\boldsymbol{x}}^t$ ($\gamma = 1$)
Linear CFR	$\boldsymbol{R}_j^{t-1} + t\,\boldsymbol{r}_j^t$	$[\boldsymbol{R}_j^t]^+/\|[\boldsymbol{R}_j^t]^+\|_1$	$\boldsymbol{X}^{t-1}(t{-}1)^\gamma/t^\gamma + \dot{\boldsymbol{x}}^t$ ($\gamma = 1$)
DCFR	$\boldsymbol{R}_j^{t-1} \odot \boldsymbol{d}_j^{t-1} + \boldsymbol{r}_j^t$	$[\boldsymbol{R}_j^t]^+/\|[\boldsymbol{R}_j^t]^+\|_1$	$\boldsymbol{X}^{t-1}(t{-}1)^\gamma/t^\gamma + \dot{\boldsymbol{x}}^t$ ($\gamma = 2$)
DCFR+	$[\boldsymbol{R}_j^{t-1} d_t + \boldsymbol{r}_j^t]^+$	$[\boldsymbol{R}_j^t]^+/\|[\boldsymbol{R}_j^t]^+\|_1$	$\boldsymbol{X}^{t-1}(t{-}1)^\gamma/t^\gamma + \dot{\boldsymbol{x}}^t$ ($\gamma = 4$)
PCFR+	$[\boldsymbol{R}_j^{t-1} + \boldsymbol{r}_j^t]^+$	$[\tilde{\boldsymbol{R}}_j^{t+1}]^+/\|[\tilde{\boldsymbol{R}}_j^{t+1}]^+\|_1$	$\boldsymbol{X}^{t-1}(t{-}1)^\gamma/t^\gamma + \dot{\boldsymbol{x}}^t$ ($\gamma = 2$)
PDCFR+	$[\boldsymbol{R}_j^{t-1} d_t + \boldsymbol{r}_j^t]^+$	$[\tilde{\boldsymbol{R}}_j^{t+1}]^+/\|[\tilde{\boldsymbol{R}}_j^{t+1}]^+\|_1$	$\boldsymbol{X}^{t-1}(t{-}1)^\gamma/t^\gamma + \dot{\boldsymbol{x}}^t$ ($\gamma = 5$)
IRR-PCFR+	$[\boldsymbol{R}_j^{t-1} + \boldsymbol{r}_j^t]^+$, with $\boldsymbol{r}_j^t \leftarrow s\,\boldsymbol{r}_j^t$ once at $t = \tau$	$[\tilde{\boldsymbol{R}}_j^{t+1}]^+/\|[\tilde{\boldsymbol{R}}_j^{t+1}]^+\|_1$	$\boldsymbol{X}^{t-1}(t{-}1)^\gamma/t^\gamma + \dot{\boldsymbol{x}}^t$ ($\gamma = 9$)

regret round without changing the strategies of that round or the next (Section 4). Compared with discounting, IRR applies the reduction only once and acts more aggressively.

- We prove that the scale factor never exceeds one, so the reduction never enlarges the first positive regret, and that IRR-RM+ and IRR-PRM+ retain the RM+ and PRM+ no-regret bounds up to a constant independent of T, preserving the $O(1/\sqrt{T})$ no-regret rate. In the online mirror descent view, IRR corresponds to a single per-information-set step-size change (Section 5).
- Experimentally, IRR-PCFR+ attains the lowest average-iterate exploitability on the benchmark games other than the heads-up no-limit Texas hold'em (HUNL) subgames and remains competitive on the last iterate; on HUNL, DCFR remains superior (Section 6).

2 Preliminaries

We first introduce notation for extensive-form games and the regret minimizers that CFR runs at each information set, then define the first positive regret round on which IRR acts. Table 1 summarizes the CFR variants we build on. We consider two-player zero-sum imperfect-information games in extensive form with perfect recall (Zinkevich et al., 2007). For a player, let $\mathcal{J}$ be the set of information sets, and let $\mathcal{A}_j$ be the set of legal actions at $j \in \mathcal{J}$, of size n_j. A *local strategy* at j is a distribution $\boldsymbol{x}_j \in \Delta^{n_j}$ over $\mathcal{A}_j$. The collection of local strategies induces a sequence-form strategy $\dot{\boldsymbol{x}} \in \mathcal{X}$. Writing the opponent's sequence-form strategy as $\dot{\boldsymbol{y}} \in \mathcal{Y}$ and the sparse payoff matrix $\boldsymbol{A}$, which encodes the minimizing player's losses, the game is the bilinear saddle point

$$\min_{\dot{\boldsymbol{x}}\in\mathcal{X}} \max_{\dot{\boldsymbol{y}}\in\mathcal{Y}} \dot{\boldsymbol{x}}^\top \boldsymbol{A}\dot{\boldsymbol{y}}, \tag{1}$$

whose solutions are the Nash equilibria. The *exploitability* of a profile $(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}})$ is $e(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}}) = \big(\delta_1 + \delta_2\big)/2$, where $\delta_1 = \dot{\boldsymbol{x}}^\top \boldsymbol{A}\dot{\boldsymbol{y}} - \min_{\dot{\boldsymbol{x}}'\in\mathcal{X}} \dot{\boldsymbol{x}}'^\top \boldsymbol{A}\dot{\boldsymbol{y}}$ and $\delta_2 = \max_{\dot{\boldsymbol{y}}'\in\mathcal{Y}} \dot{\boldsymbol{x}}^\top \boldsymbol{A}\dot{\boldsymbol{y}}' - \dot{\boldsymbol{x}}^\top \boldsymbol{A}\dot{\boldsymbol{y}}$ are the two players' incentives to deviate (Zinkevich et al., 2007; Farina et al., 2019a).

CFR runs a regret minimizer at each information set. At iteration t, the *counterfactual loss* $\boldsymbol{\ell}_j^t \in \mathbb{R}^{n_j}$ assigns to each action $a \in \mathcal{A}_j$ the loss of playing a at j and continuing under both players' current strategies, weighted by the opponent's reach probability of j and not by the player's own. The *instantaneous counterfactual regret* is $\boldsymbol{r}_j^t = \langle \boldsymbol{x}_j^t, \boldsymbol{\ell}_j^t\rangle \mathbf{1} - \boldsymbol{\ell}_j^t$: the excess of the loss of the current local strategy over that of each action. Regret Matching+ (RM+) (Tammelin, 2014), regret matching with negative accumulated regrets set to zero, accumulates $\boldsymbol{R}_j^t = [\boldsymbol{R}_j^{t-1} + \boldsymbol{r}_j^t]^+$, where $[\cdot]^+ = \max\{\cdot, 0\}$, and plays $\boldsymbol{x}_j^{t+1} = \boldsymbol{R}_j^t/\|\boldsymbol{R}_j^t\|_1$, which we take to be uniform when $\boldsymbol{R}_j^t = \mathbf{0}$. Predictive regret matching+ (PRM+) (Farina et al., 2021), the regret-matching core of PCFR+, stores the previous instantaneous regret as the prediction, $\boldsymbol{m}_j^{t+1} = \boldsymbol{r}_j^t$, forms the predicted accumulated regret $\tilde{\boldsymbol{R}}_j^{t+1} = [\boldsymbol{R}_j^t + \boldsymbol{m}_j^{t+1}]^+$, and reads out $\boldsymbol{x}_j^{t+1} = \tilde{\boldsymbol{R}}_j^{t+1}/\|\tilde{\boldsymbol{R}}_j^{t+1}\|_1$, again taking the strategy to be uniform when $\tilde{\boldsymbol{R}}_j^{t+1} = \mathbf{0}$. PCFR+ applies PRM+ at each information set inside CFR.

All accumulators start at zero. We write τ_j for the *first positive regret round* at j: the first iteration whose instantaneous counterfactual regret $\boldsymbol{r}_j^{\tau_j}$ has a positive component, so that $\boldsymbol{R}_j^{\tau_j} \neq \mathbf{0}$ while

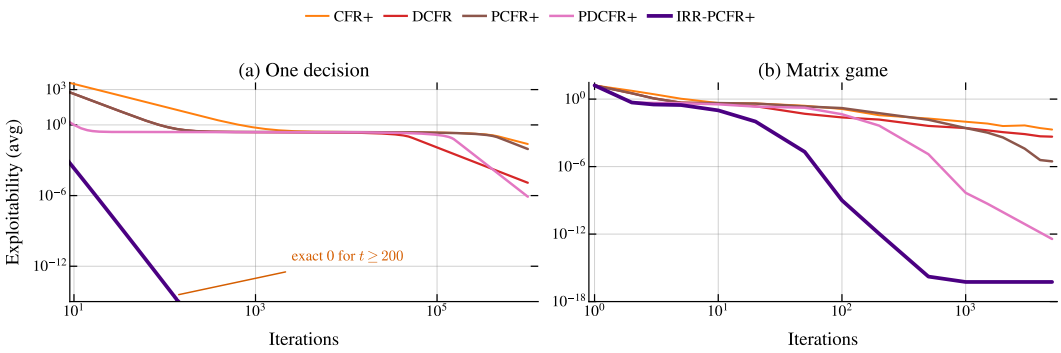

Figure 1: The two motivating examples, average-iterate exploitability. **(a)** the one-decision game of Section 3.1. **(b)** the matrix game of Section 3.2. Both panels use log–log axes.

$\boldsymbol{R}_j^{\tau_j - 1} = \mathbf{0}$ (τ when j is clear from context). The round τ_j is also called the *reduction round*: it is the only round IRR modifies (Section 4).

The *cumulative strategy* $\boldsymbol{X}^t$ and the *average strategy* $\bar{\boldsymbol{x}}^t = \boldsymbol{X}^t / \|\boldsymbol{X}^t\|_1$ aggregate the sequence-form iterates through the recursion in Table 1; for weights $w_t = t^\gamma$, this recursion is equivalent to adding the iterate $\dot{\boldsymbol{x}}^t$ with weight t^γ. Discounted variants decay $\boldsymbol{R}_j^{t-1}$ before each update: DCFR does so entrywise, whereas DCFR+ and PDCFR+ multiply by the scalar d_t (Table 1); Section 6.1 lists the parameter choices used in the experiments.

3 The uniform-initialization bias

Two examples show the uniform-initialization bias at the first positive regret round: a single decision with no opponent (Section 3.1), then a matrix game with a real opponent (Section 3.2).

3.1 Online learning with a highly suboptimal action

We start from a simple online learning setup: one decision among three actions, no opponent (Figure 1). The losses are $\boldsymbol{\ell}^1 = (0, -1, 10^6)$ for actions 1, 2, 3 (payoffs $(0, 1, -10^6)$). This is the game that motivates regret discounting in DCFR (Brown & Sandholm, 2019). The first strategy is uniform, so each action's instantaneous regret is its own loss subtracted from the average, $\boldsymbol{r}^1[a] = \frac{1}{n}\sum_{b=1}^{n} \boldsymbol{\ell}^1[b] - \boldsymbol{\ell}^1[a]$: the loss of action 3 contributes to the average and inflates the positive regrets of both other actions, putting the first accumulated regret at $\boldsymbol{R}^1 = (333{,}333,\ 333{,}334,\ 0)$. Action 2, the optimum, is thus measured against the *average* loss rather than against action 1, its only competitor: both regrets start at 3.3×10^5 even though the true gap between actions 1 and 2 is only one. Nor do later rounds fix the scale. The readout of RM+ and its variants is proportional to the accumulated regret, so the next strategy splits roughly evenly between actions 1 and 2, and the update moves them to roughly $(333{,}332.5,\ 333{,}334.5,\ 0)$, as Brown & Sandholm (2019) observe. After that, the difference between the two viable actions changes by only one per round, so CFR+ needs $471{,}406$ iterations to put probability 1 on action 2; PCFR+ is barely faster ($471{,}404$); discounting brings the count down (DCFR $52{,}126$, PDCFR+ $156{,}461$); IRR-PCFR+ needs 5. The count drops because IRR rescales exactly this first regret once (Section 4): with $\lambda = 0.99999$ the scale is $s \approx 10^{-5}$, so the accumulated regret after the first round becomes approximately $(3.33332925,\ 3.33333925,\ 0)$ (the same two nonzero coordinates, five orders of magnitude smaller), and the regret of action 1 reaches zero within a few updates.

3.2 A matrix game with dominated actions

The bias persists with a real opponent (Figure 1). We use the 3×3 zero-sum matrix game NFG-3 that motivates PDCFR+ (Xu et al., 2024): player 1's utilities are $U = ((1, 0, 5),\ (0, 2, 0),\ (0, 0, 100))$. For player 2, the minimizer, action 3 gives player 1 payoffs $(5, 0, 100)$ while action 1 gives $(1, 0, 0)$,

so action 3 is weakly dominated by action 1. After removing player 2's action 3, player 1's action 3 yields $(0, 0)$ while action 1 yields $(1, 0)$, so player 1's action 3 is weakly dominated by action 1 as well. In the reduced game, only actions 1 and 2 matter in equilibrium, i.e., the reduced game is the two-action game NFG-2. The slowdown is created before those eliminations. At the uniform start, player 2 still assigns probability $1/3$ to action 3. Against this uniform opponent, player 1's three actions have expected utilities 2, $2/3$, and $100/3$, while the played uniform strategy has utility 12, so player 1's first instantaneous regret is $(-10,\ -34/3,\ 64/3)$. Hence the first positive regret is carried entirely by action 3, and the next strategy places all weight on that action. Player 2 then faces losses $(0, 0, 100)$ across its three actions, whose uniform average is $100/3$, so the positive part of player 2's first instantaneous regret is $(100/3,\ 100/3,\ 0)$. Thus the two actions that matter in equilibrium for player 2 start from a large common base, even though after elimination the reduced game distinguishes them only through the small payoffs 1 and 2. We measure convergence by the first iteration at which the average-iterate exploitability falls below 10^{-4}: CFR+ needs 1051 iterations on the corresponding two-action game NFG-2, but $66{,}313$ on NFG-3 after the weakly dominated actions are added. On NFG-3 this count drops to $14{,}660$ for DCFR, $2{,}253$ for PCFR+, 336 for PDCFR+, and 41 for IRR-PCFR+. PDCFR+ reduces the slowdown by discounting the common base over many rounds; IRR attacks the same uniform-initialization bias more directly by shrinking the first positive regret once.

4 Initial regret reduction

This section defines IRR precisely, motivates the interpolation baseline and its single parameter, and states the algorithm.

Construction. At the reduction round τ, the accumulated regret $\boldsymbol{R}_j^{\tau-1}$ is zero. The round has already been played under the uniform strategy $\boldsymbol{x}_j^\tau$, and the next strategy $\boldsymbol{x}_j^{\tau+1}$ is the normalization of the positive part of the regret $\boldsymbol{r}_j^\tau$. IRR re-evaluates the same loss vector against the interpolated baseline

$$\boldsymbol{\sigma} = \lambda\, \boldsymbol{x}_j^{\tau+1} + (1-\lambda)\, \boldsymbol{x}_j^\tau, \tag{2}$$

a mixture of the played strategy and the next strategy. The scale factor is

$$s = \frac{\| \, [\langle \boldsymbol{\sigma}, \boldsymbol{\ell}_j^\tau \rangle \mathbf{1} - \boldsymbol{\ell}_j^\tau]^+ \, \|_1}{\| \, [\langle \boldsymbol{x}_j^\tau, \boldsymbol{\ell}_j^\tau \rangle \mathbf{1} - \boldsymbol{\ell}_j^\tau]^+ \, \|_1}, \tag{3}$$

whose denominator is the original ℓ_1 norm of $[\boldsymbol{r}_j^\tau]^+$ and whose numerator is the corresponding ℓ_1 norm under $\boldsymbol{\sigma}$. IRR replaces $\boldsymbol{r}_j^\tau$ by $s\, \boldsymbol{r}_j^\tau$ and then returns to the standard PCFR+ update (Algorithm 1).

Why this baseline? At round τ, the played strategy $\boldsymbol{x}_j^\tau$ is already fixed, and the next strategy $\boldsymbol{x}_j^{\tau+1}$ is read out from $[\boldsymbol{r}_j^\tau]^+$. Thus the quantity IRR modifies is $\boldsymbol{r}_j^\tau$ itself. The goal is to shrink the ℓ_1 norm of $[\boldsymbol{r}_j^\tau]^+$ to a smaller scale. Because τ is the first positive regret round, the played strategy $\boldsymbol{x}_j^\tau$ is uniform, while $\boldsymbol{x}_j^{\tau+1}$ is the next strategy induced by the first positive regret. IRR therefore computes the regret of that round as if it had been played with a mixture of the uniform strategy and that next strategy, and uses the resulting ℓ_1 norm of the positive regret as the target for the rescaling. IRR rescales rather than shifts: scaling preserves the support and direction of $[\boldsymbol{r}_j^\tau]^+$, whereas a shift can change the next strategy after clipping. The interpolation weight $\lambda \in (0, 1)$ controls how far the re-evaluation moves toward $\boldsymbol{x}_j^{\tau+1}$. We do not set $\lambda = 1$: if a single action carries the first positive regret, then $\boldsymbol{x}_j^{\tau+1}$ is the pure strategy on that action. Taking $\lambda = 1$ therefore makes $\boldsymbol{\sigma} = \boldsymbol{x}_j^{\tau+1}$, that is, a strategy with one action played with probability 1. Re-evaluating the same round against that pure strategy yields zero positive regret, so s would become 0 and the next readout would return to the uniform strategy. Keeping $\lambda < 1$ avoids this collapse while still moving the target mass toward the next-strategy baseline. If the re-evaluated mass vanishes ($S' = 0$), IRR instead sets $s = 1$, which guarantees $s > 0$.

Algorithm 1 IRR-PCFR+: the reduction branch at the first positive regret round; it runs at most once.

input: accumulated regret $\boldsymbol{R}_j^{t-1}$, instantaneous regret $\boldsymbol{r}_j^t$, counterfactual loss $\boldsymbol{\ell}_j^t$, current strategy $\boldsymbol{x}_j^t$, interpolation weight $\lambda \in (0,1)$, reduction flag used_j (initially FALSE)
$O \leftarrow \sum_a [\boldsymbol{r}_j^t[a]]^+$
if $\boldsymbol{R}_j^{t-1} = \mathbf{0}$ **and** $\text{used}_j = \text{FALSE}$ **and** $O > 0$ **then**
  $\boldsymbol{p}_j[a] \leftarrow [\boldsymbol{r}_j^t[a]]^+ / O$
  $\boldsymbol{\sigma} \leftarrow \lambda\, \boldsymbol{p}_j + (1-\lambda)\, \boldsymbol{x}_j^t$
  $S' \leftarrow \sum_a \left[\langle \boldsymbol{\sigma}, \boldsymbol{\ell}_j^t \rangle - \boldsymbol{\ell}_j^t[a]\right]^+$
  $s \leftarrow S'/O$ (1 if $S' = 0$)
  $\boldsymbol{r}_j^t \leftarrow s \cdot \boldsymbol{r}_j^t$; $\text{used}_j \leftarrow \text{TRUE}$
  ▷ here $\boldsymbol{p}_j = \boldsymbol{x}_j^{t+1}$ under the standard PCFR+ readout; the first round where this case applies is $t = \tau$
end if
$\boldsymbol{R}_j^t \leftarrow [\boldsymbol{R}_j^{t-1} + \boldsymbol{r}_j^t]^+$; update $\boldsymbol{x}_j^{t+1}$ and the average as PCFR+

5 Theoretical guarantees

The goal of this section is to show that IRR-PCFR+ keeps the convergence guarantee of PCFR+. We first prove that the reduction is safe: the scale factor never exceeds one, so the reduction never enlarges the first positive regret (Theorem 1). We then bound the local regret: at any information set, IRR-RM+ and IRR-PRM+ retain the RM+ and PRM+ no-regret bounds up to a constant independent of T (Theorems 2 and 3). Composing these local bounds with the CFR framework of Appendix B (Theorems 7–9) then gives the rate (Corollary 1). Section 5.2 closes with a complementary view: through the chain from Blackwell approachability to online linear optimization (OLO) and then to online mirror descent (OMD), the reduction becomes a one-time step-size change at that information set (Theorem 4). All proofs appear in Appendix C.

5.1 Scale and no-regret

Theorem 1. *The scale factor of Algorithm 1 satisfies $s \le 1$.*

Theorem 2 (No-regret guarantee for IRR-RM+)**.** *Let RM+ face an arbitrary (possibly adaptive) sequence of loss vectors $\boldsymbol{\ell}^1, \dots, \boldsymbol{\ell}^T$ on action set $\mathcal{A}$, whose induced instantaneous regrets satisfy $\|\boldsymbol{r}^t\|_\infty \le G$ for all t. Let $\kappa \in \{0,1\}$ indicate whether the reduction occurred; if $\kappa = 1$, let τ be the reduction round and $s \in (0,1]$ the scale used there. The resulting algorithm remains no-regret: its T-round regret $R_{\text{IRR}}^T = \max_a \sum_{t=1}^T \boldsymbol{r}^t[a]$ obeys*

$$R_{\text{IRR}}^T \le G\sqrt{|\mathcal{A}|T} + G\,\kappa, \tag{4}$$

so the excess over the standard RM+ bound is a constant independent of T.

Theorem 3 (No-regret guarantee for IRR-PRM+)**.** *Let PRM+ face an arbitrary (possibly adaptive) sequence of loss vectors $\boldsymbol{\ell}^1, \dots, \boldsymbol{\ell}^T$ on action set $\mathcal{A}$, whose induced instantaneous regrets satisfy $\|\boldsymbol{r}^t\|_\infty \le G$ for all t. Let $\kappa \in \{0,1\}$ indicate whether the reduction occurred; if $\kappa = 1$, let τ be the reduction round and $s \in (0,1]$ the scale used there. Run the standard prediction rule $\boldsymbol{m}^1 = \mathbf{0}$ and $\boldsymbol{m}^{t+1} = \boldsymbol{r}^t$, except that if $\kappa = 1$, then $\boldsymbol{m}^{\tau+1} = s\,\boldsymbol{r}^\tau$. The resulting algorithm remains no-regret: its T-round regret $R_{\text{IRR}}^T = \max_a \sum_{t=1}^T \boldsymbol{r}^t[a]$ obeys*

$$R_{\text{IRR}}^T \;\le\; \sqrt{3}\,G\sqrt{|\mathcal{A}|T} \;+\; G\,\kappa. \tag{5}$$

The bound has the same shape as Theorem 2: the reduction still costs only the constant $G\kappa$, and the predicted readout changes only the leading constant to $\sqrt{3}$. Setting $\kappa = 0$ (that is, taking no reduction) recovers a no-regret bound for PRM+ itself.

Corollary 1 (Rate in extensive-form games)**.** *Composing Theorem 2 for IRR-RM+ and Theorem 3 for IRR-PRM+ with the CFR framework of Appendix B gives: for any non-decreasing weights $w_t \ge 0$, not all zero, the w-weighted average strategy profile is a Nash equilibrium up*

Algorithm 2 IRR-PRM+.

state: $\boldsymbol{R}^0 \leftarrow \mathbf{0}$, $\boldsymbol{m}^1 \leftarrow \mathbf{0}$
function NextStrategy($\boldsymbol{m}^t$)
  $\tilde{\boldsymbol{R}}^t \leftarrow [\boldsymbol{R}^{t-1} + \boldsymbol{m}^t]^+$
  if $\tilde{\boldsymbol{R}}^t \neq \mathbf{0}$ **return** $\boldsymbol{x}^t \leftarrow \tilde{\boldsymbol{R}}^t / \|\tilde{\boldsymbol{R}}^t\|_1$
  else return $\boldsymbol{x}^t \leftarrow \mathbf{1}/|\mathcal{A}|$
function ObserveLoss($\boldsymbol{\ell}^t$)
  $\boldsymbol{r}^t \leftarrow \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t$
  $\boldsymbol{r}^t \leftarrow s\,\boldsymbol{r}^t$ ▷ once, at $t = \tau$, s as in Algorithm 1
  $\boldsymbol{R}^t \leftarrow [\boldsymbol{R}^{t-1} + \boldsymbol{r}^t]^+, \quad \boldsymbol{m}^{t+1} \leftarrow \boldsymbol{r}^t$

Algorithm 3 Closed-form predictive OMD update over $\mathbb{R}^{|\mathcal{A}|}_{\geq 0}$, with step sizes $\eta_\tau = s$ and $\eta_t = 1$ otherwise.

state: $\boldsymbol{z}^0 \leftarrow \mathbf{0}$, $\boldsymbol{v}^1 \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathcal{A}|}$
step sizes: $\eta_t = s$ if $t = \tau$, 1 otherwise
function NextDecision($\boldsymbol{v}^t$)
  return $\boldsymbol{\theta}^t \leftarrow [\boldsymbol{z}^{t-1} + \boldsymbol{v}^t]^+$
function ObservePayoff($\boldsymbol{u}^t$)
  $\boldsymbol{z}^t \leftarrow [\boldsymbol{z}^{t-1} + \eta_t \boldsymbol{u}^t]^+$
  $\boldsymbol{v}^{t+1} \leftarrow \eta_t \boldsymbol{u}^t$

to $\frac{w_T}{\sum_{t=1}^T w_t}\, O\big(G\,|\mathcal{I}|\,(\sqrt{nT}+1)\big)$*, where* $|\mathcal{I}| = |\mathcal{J}_1| + |\mathcal{J}_2|$ *is the total number of information sets across both players and* $n = \max_j n_j$*. For the averaging family* $w_t = t^\gamma$ *of Section 2 with* $\gamma \geq 0$*, the bound is* $O\big((\gamma{+}1)\,G\,|\mathcal{I}|\,(\sqrt{n}+1)/\sqrt{T}\big)$*.*

The uniform average is the case $\gamma = 0$. The averaging step rests on Theorem 9, the weighted-averaging theorem of DCFR (Brown & Sandholm, 2019), which generalized the linear averaging of CFR+ (Tammelin, 2014) to any non-decreasing weights. Appendix C combines the theorem with Theorems 2 and 3 and tracks the share the reduction withholds.

In finite extensive-form games the constant G is concrete: each counterfactual loss is the opponent reach of the information set times a conditional continuation value, so its range is bounded by the payoff range $\max_z u_i(z) - \min_z u_i(z)$ over terminal states z. The same bound therefore applies to every action's instantaneous regret, and Theorems 2 and 3 hold with $G = \max_z u_i(z) - \min_z u_i(z)$.

5.2 The OMD correspondence

Theorems 1 through 3 give the regret guarantees. We now re-express the same local solver through the Blackwell–OLO view of Farina et al. (2021), in which the OLO oracle is instantiated as an OMD update. The displayed algorithm below is that OMD update itself; the surrounding approachability construction then applies a forcing action to map its nonnegative readout back to the simplex.

Fix one information set and write $\mathcal{A}$ for its local action set. Regret matching on the simplex is a Blackwell approachability procedure (Hart & Mas-Colell, 2000). Abernethy et al. (2011) show that approachability of a closed convex cone C can be implemented through no-regret online linear optimization over the polar cone C°. In the simplex regret problem the target cone is $\mathbb{R}^{|\mathcal{A}|}_{\leq 0}$, so the oracle domain is the nonnegative orthant $\mathbb{R}^{|\mathcal{A}|}_{\geq 0}$.

Farina et al. (2021) instantiate that OLO oracle as predictive OMD with regularizer $\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$. Their algorithm (Farina et al., 2021, Algorithm 3) forms the approachability action by a forcing action: for the simplex regret problem one may take $g(\boldsymbol{\theta}) = \boldsymbol{\theta}/\|\boldsymbol{\theta}\|_1$ when $\boldsymbol{\theta} \neq \mathbf{0}$, while at $\boldsymbol{\theta} = \mathbf{0}$ any simplex point is valid. In our local solver we choose the uniform strategy in that zero case, matching Algorithm 2. When the payoff observed by the oracle is the local regret increment, $\boldsymbol{u}^t = \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t = \boldsymbol{r}^t_j$, this reproduces RM$^+$. In the predictive version, the update stores $\boldsymbol{v}^{t+1} = \eta_t \boldsymbol{u}^t$ for the next decision; at constant step size this is exactly the prediction $\boldsymbol{m}^{t+1} = \boldsymbol{r}^t$ of PRM$^+$, so the same construction reproduces PRM$^+$. The averaging of PCFR+ remains outside this local construction.

IRR changes only one round of that local OMD update. At the reduction round τ, the pre-update OMD state is zero, so we simply use the step size $\eta_\tau = s$ chosen by IRR for that round. Because Algorithm 3 also stores the observed vector scaled by the step size, $\boldsymbol{v}^{t+1} = \eta_t \boldsymbol{u}^t$, as the next prediction, the same one-time factor is carried into it, exactly as IRR-PRM+ stores the scaled regret once at $t = \tau + 1$. After the forcing action maps the OMD readout back to the simplex, this gives the same local strategy sequence. Running the same construction at every information set recovers the local core of IRR-PCFR+. The theorem below records this correspondence.

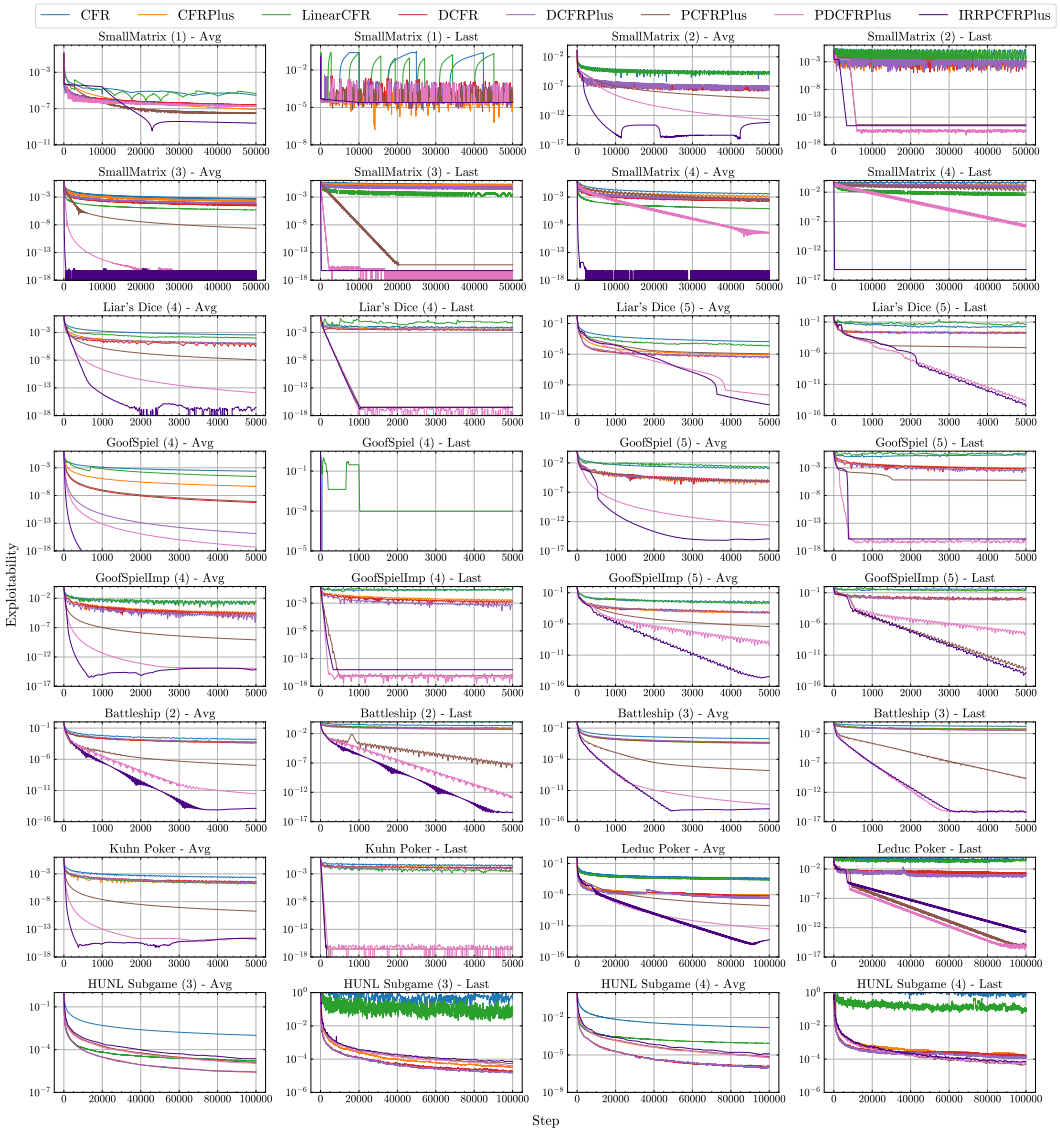

Figure 2: Exploitability of the average iterate and the last iterate on every benchmark game; linear iteration axes, logarithmic exploitability axes, budgets as in Section 6.1.

Theorem 4 (IRR-PRM+ as a one-time OMD step-size change)**.** *Fix an arbitrary (possibly adaptive) sequence of losses and an information set with local action set $\mathcal{A}$, and let a run of IRR-PRM+ (Algorithm 2) reduce at round τ with scale $s \in (0, 1]$. When the approachability construction of Farina et al. (2021, Algorithm 3) is set up with oracle domain $\mathbb{R}^{|\mathcal{A}|}_{\geq 0}$ and the OMD update of Algorithm 3 as the regret minimizer, with step sizes $\eta_\tau = s$ and $\eta_t = 1$ otherwise, observed payoffs $\boldsymbol{u}^t = \boldsymbol{r}^t_j$, and the uniform strategy at the zero readout, it produces the same strategy sequence as IRR-PRM+.*

6 Experiments

6.1 Setup

The benchmark comprises 16 two-player zero-sum imperfect-information games: four matrix-derived games, the ten-game standard sequential benchmark of Xu et al. (2024), and HUNL Sub-

games (3) and (4), with Subgame (4) the larger-pot case (game details in Appendix A). The SmallMatrix family contains SmallMatrix (1) and (2), the 2×2 hard instance $A_\delta = \left(\left(\frac{1}{2}+\delta,\ \frac{1}{2}\right),\ (0,\ 1)\right)$ of Cai et al. (2024) with $\delta = 10^{-4}$ and 10^{-2}, respectively; SmallMatrix (3) is the NFG-3 example of Xu et al. (2024), and SmallMatrix (4) is a 10×10 extension that keeps two actions that matter in equilibrium and adds eight dominated ones. The sequential benchmark contains Kuhn poker, Leduc poker, Liar's dice (4) and (5), Goofspiel (4) and (5), the imperfect-information Goofspiel (4) and (5), and Battleship (2) and (3). We run $T = 5000$ iterations in most games, including Battleship (3), with $50{,}000$ for the SmallMatrix variants and $100{,}000$ for Leduc and HUNL Subgames (3) and (4).

All algorithms run on the same implementation (the PDCFR+ package) with alternating updates (the two players' strategies are updated in turn within an iteration). We report the exploitability of both the average iterate, which is what the no-regret theory certifies, and the last iterate, which shows whether a method is actually stabilizing: optimistic and forgetful methods can have qualitatively different last-iterate behavior, both in normal-form games (Daskalakis & Panageas, 2019; Cai et al., 2024; 2025) and in extensive-form games (Lee et al., 2021). We compare eight algorithms: CFR, CFR+ (Tammelin, 2014), Linear CFR (LCFR) and DCFR (Brown & Sandholm, 2019) ($\alpha = 1.5$, $\beta = 0$, $\gamma = 2$; LCFR is the special case DCFR$(1, 1, 1)$ in that paper), DCFR+ and PDCFR+ (Xu et al., 2024) with the configuration used by their authors ($\alpha = 1.5$, $\gamma = 4$ for DCFR+; $\alpha = 2.3$, $\gamma = 5$ for PDCFR+), PCFR+ (Farina et al., 2021) ($\gamma = 2$), and IRR-PCFR+ ($\lambda = 0.99999$, $\gamma = 9$, reduction applied once). λ and γ were selected by a sweep on this benchmark and then frozen for all reported runs, following the protocol of Xu et al. (2024). The exponent γ affects only the average-iterate weighting, not the regret updates or the last-iterate readout.

6.2 Main results

In average-iterate exploitability, IRR-PCFR+ attains the lowest value by the end of the run in every non-HUNL benchmark game (Figure 2). The SmallMatrix games show the uniform-initialization bias most clearly, especially SmallMatrix (3) and (4), where dominated actions make the effect explicit. Beyond these diagnostic cases, the same advantage extends across the broader sequential benchmark, including Liar's dice, Goofspiel, Battleship, Kuhn poker, and Leduc poker. On the HUNL subgames, DCFR remains strongest.

The last-iterate results are more heterogeneous, but still favorable to IRR-PCFR+ (Figure 2). Several SmallMatrix games drop rapidly under IRR, and the method is best or tied-best on many other games as well. This is notable because, relative to PCFR+, IRR changes only one regret vector at one round of each information set, yet in most games that single intervention produces substantially faster convergence.

7 Related work

We position IRR against the CFR variants it builds on, the approachability view, and stability and last-iterate studies.

CFR variants. CFR introduced the counterfactual-regret decomposition for extensive-form zero-sum games (Zinkevich et al., 2007). CFR+ replaces regret matching by regret matching+ and clips negative regrets, which greatly improves practical speed (Tammelin, 2014). Linear CFR (LCFR) and DCFR then discount early iterations in the regrets and/or in the average iterate (Brown & Sandholm, 2019). PCFR+ and PDCFR+ add prediction to the update, following the broader line that links predictable online learning, Blackwell approachability, and online mirror descent with forcing action (Rakhlin & Sridharan, 2013; Abernethy et al., 2011; Farina et al., 2021; Xu et al., 2024). Warm starting initializes the accumulators from a precomputed strategy and requires a single traversal of the game tree (Brown & Sandholm, 2016). Scale-invariant regret matching normalizes updates to remove dependence on payoff scale (Zhang et al., 2025). IRR addresses the uniform-initialization bias from a different direction: it needs no prior strategy, and instead of repeated per-iteration discounting it applies a one-time rescaling to the first positive regret, whose initial scale the uniform start can distort.

Approachability and mirror descent. Regret matching can be seen as a Blackwell approachability procedure (Hart & Mas-Colell, 2000). Abernethy et al. (2011) connect this view to no-regret

online linear optimization, and Farina et al. (2021) instantiate the resulting oracle as online mirror descent together with a forcing action for the predictive regret-matching core of PCFR+. This is the perspective in which our one-time rescaling becomes a one-time per-information-set step-size change in the OMD update, while the forcing action stays standard.

Stability and last-iterate behavior. Farina et al. (2023) give counterexamples in which RM+ and its predictive variant are unstable in games, together with fixes that restore stability. For optimistic learning in zero-sum games, Daskalakis & Panageas (2019) establish last-iterate convergence, Cai et al. (2024) show that fast last-iterate rates require forgetful algorithms, and Cai et al. (2025) relate average- iterate and last-iterate guarantees. In extensive-form games, Lee et al. (2021) show that several optimistic treeplex methods enjoy last-iterate convergence even though vanilla CFR relies on averaging.

8 Conclusion

We study a simple failure mode of CFR-type methods: the first positive regret at an information set can start at the wrong scale because it is formed under the uniform strategy. We call this effect the uniform-initialization bias. IRR reduces that regret once and then returns to the standard PCFR+ update. The theory shows that this one-time reduction never enlarges the first positive regret, and that IRR-RM+ and IRR-PRM+ preserve the standard no-regret rate. Experimentally, IRR-PCFR+ attains the lowest average-iterate exploitability across the benchmark games except the HUNL subgames, with especially clear gains on the matrix-derived games that expose the uniform-initialization bias most cleanly. This single round of rescaling already accelerates convergence markedly on most of the benchmark, and in the online mirror descent view of Section 5.2 it is exactly a one-time step-size change (Theorem 4). On HUNL, DCFR remains superior.

AI use statement

We used large language models to help check proofs and polish the writing. They were not used to run experiments, to write the paper as a whole, or to derive the core theory or results. All final technical content and claims were produced and verified by the authors.

Reproducibility statement

There is no random seed: every run is fully deterministic. Re-running the same configuration yields exactly identical outputs and numbers. All numbers are recomputed from stored per-run logs. The algorithms are implemented in the PDCFR+ package (Xu et al., 2024): Algorithm 1 states the reduction, Table 1 and Section 6.1 fix the update rules and the configuration, and Appendix A gives the game details. Game definitions are standard OpenSpiel/PokerRL benchmarks. Code and logs will be released.

References

Jacob Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *JMLR Workshop and Conference Proceedings*, pp. 27–46, 2011.

Noam Brown and Tuomas Sandholm. Strategy-based warm starting for regret minimization in games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1829–1836, 2019.

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Fast last-iterate convergence of learning in games requires forgetful algorithms. In *Advances in Neural Information Processing Systems*, pp. 23406–23434, 2024.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. From average-iterate to last-iterate convergence in games: A reduction and its applications. In *Advances in Neural Information Processing Systems*, pp. 52398–52428, 2025.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *Innovations in Theoretical Computer Science Conference (ITCS)*, volume 124 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 27:1–27:18, 2019.

Gabriele Farina, Christian Kroer, Noam Brown, and Tuomas Sandholm. Stable-predictive optimistic counterfactual regret minimization. In *International Conference on Machine Learning*, pp. 1853–1862, 2019a.

Gabriele Farina, Chun Kai Ling, Fei Fang, and Tuomas Sandholm. Correlation in extensive-form games: Saddle-point formulation and benchmarks. In *Advances in Neural Information Processing Systems*, pp. 9229–9239, 2019b.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive Blackwell approachability: Connecting regret matching and mirror descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5363–5371, 2021.

Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, and Haipeng Luo. Regret matching+: (in)stability and fast convergence in games. In *Advances in Neural Information Processing Systems*, 2023.

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

Harold W. Kuhn. A simplified two-person poker. In *Contributions to the Theory of Games*, volume 1, pp. 97–103. Princeton University Press, 1950.

Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. *arXiv preprint arXiv:2106.14326*, 2021.

Viliam Lisý, Marc Lanctot, and Michael H. Bowling. Online monte carlo counterfactual regret minimization for search in imperfect information games. In *International Conference on Autonomous Agents and Multiagent Systems*, pp. 27–36, 2015.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

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

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: Opponent modelling in poker. In *Conference on Uncertainty in Artificial Intelligence*, pp. 550–558, 2005.

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

Hang Xu, Kai Li, Bingyun Liu, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Minimizing weighted counterfactual regret with optimistic online mirror descent. In *International Joint Conference on Artificial Intelligence*, pp. 5272–5280, 2024.

Brian Hu Zhang, Ioannis Anagnostides, and Tuomas Sandholm. Scale-invariant regret matching and online learning with optimal convergence: Bridging theory and practice in zero-sum games. *arXiv preprint arXiv:2510.04407*, 2025.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, volume 20, 2007. Extended version with appendix; Lemma 7 and Corollary 6 cited from the extended version.

Appendix

The appendix follows the paper's argument:

- Benchmark games and size metadata appear in Appendix A.
- The CFR convergence framework appears in Appendix B, following Zinkevich et al. (2007) and Brown & Sandholm (2019).
- Proofs of the claims of Section 5, including the approachability-to-OMD chain used in Section 5.2 and the OMD correspondence of Theorem 4, appear in Appendix C.

A Benchmark games

This appendix describes the benchmark games and the matrix examples used to interpret Figure 2. All games are two-player zero-sum. Kuhn and Leduc poker, Liar's dice, Goofspiel, and Battleship are the OpenSpiel benchmarks of Xu et al. (2024); the HUNL subgames use the PokerRL implementation distributed with that code base.

Kuhn poker (Kuhn, 1950). The deck has three cards (J, Q, K). Each player antes one chip and receives one private card, and a single betting round follows in which a player may bet one chip; the fold, call, and check actions have their usual meaning. If neither player folds, the higher card wins the pot. Utilities are chips won or lost.

Leduc poker (Southey et al., 2005). The deck has six cards, two suits of three ranks. Each player antes one chip and receives one private card, and the game unfolds over two betting rounds with maximum bet sizes of two and four chips. After the first round one public card is revealed. A private card that pairs the public card wins outright; otherwise the higher private card wins.

Liar's dice (Lisý et al., 2015). Each player rolls one x-sided die ($x = 4, 5$) under a cup and observes the outcome privately. Players alternate. A bid of the form "at least p dice show face q" (the highest face counts as any face) must raise the previous bid; a player may instead call liar, which ends the game and reveals the dice: if the last bid is false its bidder loses, otherwise the caller loses. The winner's utility is $+1$, the loser's -1.

Goofspiel (Ross, 1971). Each player holds the cards $1, \dots, x$ ($x = 4, 5$), and a point deck with the same values is revealed one card per round. In each round both players simultaneously bid one card for the point card; the higher bid takes it, and equal bids discard it. After x rounds the player with more points wins; the winner's utility is $+1$, the loser's -1. The point deck is fixed in decreasing order, as in the PDCFR+ benchmark. Goofspiel-imp is the imperfect-information variant in which a player learns only whether each bid was won or lost, not the opponent's card.

Battleship (Farina et al., 2019b). Each player secretly places one ship of size 1×2 and value 2 on their own grid (2×2 or 3×2). Players alternate shooting at cells of the opponent's grid, no cell more than once; the game ends when a ship is sunk or both players have completed three shots. The utility is the value of the opponent's sunk ship minus that of one's own.

HUNL subgames. The HUNL panels in Figure 2 report Subgames (3) and (4) of the four heads-up no-limit Texas hold'em endgames generated by Libratus and studied by Brown & Sandholm (2019). Both begin at the start of the final betting round with the board dealt and fixed initial beliefs over the two players' private hands. Subgame (4) starts from the larger-pot river state, with $3{,}750$ chips in the pot; Subgame (3) starts from a smaller-pot river state. Betting is discretized as in that paper: the first bet of a round uses bet sizes of 0.5 and 1 times the pot and all-in, and subsequent bets use 1 times the pot and all-in.

Matrix games. The top four panels of Figure 2 are matrix games. SmallMatrix (1) and (2) are 2×2 zero-sum games adapted from examples in Cai et al. (2024); player 1's payoff matrices are $\big((0.5001, 0.5), (0, 1)\big)$ and $\big((0.51, 0.5), (0, 1)\big)$, respectively. SmallMatrix (3) is the 3×3 NFG-3

Table 2: Size statistics for the benchmark games, following the PDCFR+ benchmark table: total histories, total information sets across both players, terminal histories, tree depth, and maximum infoset cardinality.

Game	Histories	Infosets	Terminal histories	Depth	Max infoset cardinality
SmallMatrix (1)	7	2	4	2	2
SmallMatrix (2)	7	2	4	2	2
SmallMatrix (3)	13	2	9	2	3
SmallMatrix (4)	111	2	100	2	10
Kuhn poker	58	12	30	5	2
Leduc poker	9,457	936	5,520	11	5
Liar's dice (4)	8,181	1,024	4,080	11	4
Liar's dice (5)	51,181	5,120	25,575	13	5
Goofspiel (4)	1,077	270	576	6	8
Goofspiel (5)	26,931	3,252	14,400	8	48
Goofspiel-imp (4)	1,077	162	576	6	14
Goofspiel-imp (5)	26,931	2,124	14,400	8	46
Battleship (2)	10,069	3,286	5,568	8	4
Battleship (3)	732,607	81,027	552,132	8	7
HUNL Subgame (3)	398,112,843	69,184	261,126,360	9	1,980
HUNL Subgame (4)	244,005,483	43,240	158,388,120	7	1,980

example of Xu et al. (2024), with player 1 payoff matrix $\big((1,0,5),(0,2,0),(0,0,100)\big)$. For player 2, action 3 gives player 1 payoffs $(5,0,100)$ while action 1 gives $(1,0,0)$, so action 3 is weakly dominated by action 1; after eliminating it, player 1's action 3 yields $(0,0)$ while action 1 yields $(1,0)$, so the reduced game is the 2×2 game on the first two actions. SmallMatrix (4) is a 10×10 extension of the same construction: two actions remain viable and the other eight are dominated but assigned large and heterogeneous payoffs. We do not reproduce that full matrix here; for the experiments, the key point is that it keeps the same two viable actions while making the payoffs of the dominated actions much larger.

Table 2 reports size statistics for the benchmark games, following the PDCFR+ benchmark table. For the OpenSpiel games, the numbers come from OpenSpiel profiling; for the HUNL subgames, they are obtained from the local PokerRL/Libratus benchmark definition. For the matrix games, we apply OpenSpiel's turn-based conversion before profiling, so each game contributes one decision state per player and the history counts are measured on that converted tree. Depth is the maximum number of actions in a history.

B CFR DERIVATION

This appendix assembles the convergence framework of CFR and proves each piece. Theorem 5 through Theorem 8 follow Zinkevich et al. (2007), with theorem numbers cited from the extended version; Theorem 9 follows Brown & Sandholm (2019). Theorem 5 decomposes the full counterfactual regret at an information set into an immediate term plus terms at the successor information sets, which reduces whole-game regret minimization to regret matching at each information set separately. Theorem 6 unfolds this recursion along the subtree, and Theorem 7 sums the unfolded bound over all information sets. Theorem 8 converts low overall regret of both players into proximity of the weighted average profile to a Nash equilibrium, and Theorem 9 carries the guarantee to any non-decreasing round weighting, the form used by the γ-weighted average of Section 2. The input the framework needs is a bound on each accumulator; the results of Section 5 provide it, and inserting them into the chain yields Corollary 1, with the composition recorded in Appendix C.

Notation. For player i and information set I, let $\pi^{\sigma}_{-i}(I)$ be the reach probability of I under σ excluding player i's own contribution (the *opponent's reach*). Let Σ_i be the set of sequence-form strategies for player i. For a deviation $\sigma^0\in\Sigma_i$, write $\sigma|I\to a$ for the strategy that plays action a at I and follows σ elsewhere, and $\sigma|D(I)\to\sigma^0$ for the strategy that follows σ^0 on the subtree rooted at I and σ on all other parts of the tree. The *counterfactual utility* $u_i(\sigma,I)$ is the expected payoff of player i conditional on reaching I and playing according to σ thereafter. For each action $a\in\mathcal{A}(I)$,

write the corresponding action-wise counterfactual regret term as

$$R_{T,i}(I,a) \;=\; \frac{1}{T}\sum_{t=1}^{T}\pi_{-i}^{\sigma_t}(I)\big(u_i(\sigma_t \mid I\to a, I) - u_i(\sigma_t, I)\big), \tag{6}$$

which changes only the action chosen at I and leaves the continuation below I unchanged. The *immediate counterfactual regret* is

$$R_{T,i,\text{imm}}(I) \;=\; \max_{a\in\mathcal{A}(I)} R_{T,i}(I,a) \;=\; \max_{a\in\mathcal{A}(I)} \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{r}_I^t[a], \tag{7}$$

where $\boldsymbol{r}_I^t$ is the instantaneous counterfactual regret of Section 2, its information set written as a subscript. The terminology follows the two sources: *immediate* is the single-information-set quantity of Zinkevich et al. (2007), the regret in the decisions at I itself as opposed to the full counterfactual regret below, while *instantaneous* is the single-round quantity of Brown & Sandholm (2019). The *full counterfactual regret* is

$$R_{T,i,\text{full}}(I) \;=\; \frac{1}{T}\max_{\sigma^0\in\Sigma_i}\sum_{t=1}^{T}\pi_{-i}^{\sigma_t}(I)\big(u_i(\sigma_t \mid D(I)\to\sigma^0, I) - u_i(\sigma_t, I)\big), \tag{8}$$

which allows an arbitrary continuation deviation in the whole subtree below I. Thus $R_{T,i}(I,a)$ is the fixed-action local term, $R_{T,i,\text{imm}}(I)$ is the best one-action deviation, and $R_{T,i,\text{full}}(I)$ is the best deviation over the subtree below I. For each $a \in \mathcal{A}(I)$, let $\text{Succ}_i(I,a)$ be the set of information sets of player i immediately reachable after action a at I, and let $\text{Succ}_i(I) = \bigcup_a \text{Succ}_i(I,a)$; the transition probability from (I,a) to a successor I_0 under σ_t is written $\text{succ}_i^{\sigma_t}(I_0 \mid I, a)$. Under perfect recall these successor families are disjoint across actions, and the subtrees rooted at them partition what lies below I. Let $\mathcal{D}_i(I)$ be the information sets of player i reachable from I, including I itself. The regret that the decomposition ultimately controls is the *overall regret*: for weights $w_t \geq 0$, not all zero, with $W_T = \sum_{t=1}^{T} w_t$,

$$R_{i,w} \;=\; \max_{\sigma^0\in\Sigma_i}\frac{1}{W_T}\sum_{t=1}^{T} w_t\big(u_i(\sigma^0, \sigma_{-i}^t) - u_i(\sigma^t)\big), \tag{9}$$

the weighted overall regret against the best fixed deviation, that is, the best single strategy σ^0 used against every round. The uniform average $w_t \equiv 1$ is the special case. Likewise, the weighted versions of the local regret quantities replace each uniform average $(1/T)\sum_{t=1}^{T}(\cdot)$ by $(1/W_T)\sum_{t=1}^{T} w_t(\cdot)$.

Theorem 5 (Exact decomposition of full counterfactual regret; per-action form of Corollary 6 of Zinkevich et al. (2007))**.** *For every player i and information set I,*

$$\begin{aligned} R_{T,i,\text{full}}(I) = \max_{a\in\mathcal{A}(I)} \Bigg\{ &\frac{1}{T}\sum_{t=1}^{T}\pi_{-i}^{\sigma_t}(I)\big(u_i(\sigma_t \mid I\to a, I) - u_i(\sigma_t, I)\big) \\ &+ \sum_{I_0\in\text{Succ}_i(I,a)} R_{T,i,\text{full}}(I_0)\Bigg\}. \end{aligned} \tag{10}$$

Taking the positive part of both sides yields the form used in the proofs below, with the first term written explicitly as $R_{T,i}(I,a)$; maximizing it over a gives $R_{T,i,\text{imm}}(I)$.

Proof. Start from the definition of $R_{T,i,\text{full}}(I)$:

$$R_{T,i,\text{full}}(I) = \frac{1}{T}\max_{\sigma^0\in\Sigma_i}\sum_{t=1}^{T}\pi_{-i}^{\sigma_t}(I)\big(u_i(\sigma_t \mid D(I)\to\sigma^0, I) - u_i(\sigma_t, I)\big). \tag{11}$$

Step 1: separate by the first action.

Any σ^0 specifies some action a at I. Decompose the maximum over strategies into a maximum over the first action followed by a maximum over the subtree given that action:

$$\max_{\sigma^0 \in \Sigma_i} = \max_{a \in \mathcal{A}(I)} \max_{\substack{\sigma^0 \in \Sigma_i \\ \sigma^0(I)=a}} . \tag{12}$$

Step 2: decompose the utility difference.

For a fixed action a at I with $\sigma^0(I) = a$, the two strategies make the same first move at I; any difference appears only if play reaches a successor information set at which player i acts again. Therefore

$$\begin{aligned} & u_i(\sigma_t \mid D(I) \to \sigma^0, I) - u_i(\sigma_t \mid I \to a, I) \\ = & \sum_{I_0 \in \mathrm{Succ}_i(I,a)} \mathrm{succ}_i^{\sigma_t}(I_0 \mid I, a) \big(u_i(\sigma_t \mid D(I) \to \sigma^0, I_0) - u_i(\sigma_t \mid I \to a, I_0)\big) \\ = & \sum_{I_0 \in \mathrm{Succ}_i(I,a)} \mathrm{succ}_i^{\sigma_t}(I_0 \mid I, a) \big(u_i(\sigma_t \mid D(I) \to \sigma^0, I_0) - u_i(\sigma_t, I_0)\big), \end{aligned} \tag{13}$$

where the second equality uses that from each successor I_0 onward $\sigma_t \mid I \to a$ coincides with σ_t. Adding $u_i(\sigma_t \mid I \to a, I) - u_i(\sigma_t, I)$ to both sides gives

$$\begin{aligned} & u_i(\sigma_t \mid D(I) \to \sigma^0, I) - u_i(\sigma_t, I) \\ = & \big(u_i(\sigma_t \mid I \to a, I) - u_i(\sigma_t, I)\big) \\ & + \sum_{I_0 \in \mathrm{Succ}_i(I,a)} \mathrm{succ}_i^{\sigma_t}(I_0 \mid I, a) \big(u_i(\sigma_t \mid D(I) \to \sigma^0, I_0) - u_i(\sigma_t, I_0)\big). \end{aligned} \tag{14}$$

The first term is the immediate effect of playing a at I; the sum over successors captures the downstream effect on the subtree below (I, a).

Step 3: multiply by the opponent's reach and sum.

Multiply equation 14 by $\pi_{-i}^{\sigma_t}(I)$. By construction,

$$\pi_{-i}^{\sigma_t}(I)\, \mathrm{succ}_i^{\sigma_t}(I_0 \mid I, a) = \pi_{-i}^{\sigma_t}(I_0), \tag{15}$$

because under perfect recall every path that reaches a node of I_0 passes through I with player i taking a: nodes in the same information set share the player's action-observation history, which determines the information sets visited and the actions taken there. The player's action at I contributes nothing to $\pi_{-i}^{\sigma_t}$, and no further player-i action separates I from I_0, so $\pi_{-i}^{\sigma_t}(I_0)$ equals the opponent and chance reach at I times the transition probability along the committed edge (I, a). Substituting the result into the branch of equation 11 corresponding to action a gives

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^{T} \pi_{-i}^{\sigma_t}(I) \big(u_i(\sigma_t \mid I \to a, I) - u_i(\sigma_t, I)\big) \\ & \quad + \frac{1}{T} \max_{\substack{\sigma^0 \in \Sigma_i \\ \sigma^0(I)=a}} \sum_{I_0 \in \mathrm{Succ}_i(I,a)} \sum_{t=1}^{T} \pi_{-i}^{\sigma_t}(I_0) \big(u_i(\sigma_t \mid D(I) \to \sigma^0, I_0) - u_i(\sigma_t, I_0)\big). \end{aligned} \tag{16}$$

Step 4: factorize the subtree maximization. Perfect recall guarantees that distinct successors $I_0 \in \mathrm{Succ}_i(I, a)$ have disjoint descendant sets $\mathcal{D}_i(I_0)$. Thus, once the first action a is fixed, every global deviation σ^0 with $\sigma^0(I) = a$ restricts to one continuation strategy on each subtree $D(I_0)$, and conversely any choice of continuation strategies on these subtrees can be pasted into one global deviation below I. Hence the maximum separates:

$$\begin{aligned} & \max_{\substack{\sigma^0 \in \Sigma_i \\ \sigma^0(I)=a}} \sum_{I_0 \in \mathrm{Succ}_i(I,a)} \sum_{t=1}^{T} \pi_{-i}^{\sigma_t}(I_0) \big(u_i(\sigma_t \mid D(I) \to \sigma^0, I_0) - u_i(\sigma_t, I_0)\big) \\ = & \sum_{I_0 \in \mathrm{Succ}_i(I,a)} \max_{\tilde{\sigma}^0_{I_0}} \sum_{t=1}^{T} \pi_{-i}^{\sigma_t}(I_0) \big(u_i(\sigma_t \mid D(I_0) \to \tilde{\sigma}^0_{I_0}, I_0) - u_i(\sigma_t, I_0)\big), \end{aligned} \tag{17}$$

Each inner maximum is $T \cdot R_{T,i,\text{full}}(I_0)$ by definition, so the branch for action a reduces to

$$\frac{1}{T}\sum_{t=1}^{T}\pi_{-i}^{\sigma_t}(I)\big(u_i(\sigma_t \mid I \to a, I) - u_i(\sigma_t, I)\big) + \sum_{I_0 \in \text{Succ}_i(I,a)} R_{T,i,\text{full}}(I_0). \tag{18}$$

Step 5: maximize over the first action. Taking the outer maximum over $a \in \mathcal{A}(I)$ recovers the claimed recurrence equation 10.

Connection to Corollary 6. Applying the positive-part operator $x \mapsto x^+ = \max\{x, 0\}$ to both sides of equation 10 and using $x^+ + y^+ \geq (x+y)^+$ gives

$$\begin{aligned} R^+_{T,i,\text{full}}(I) \leq \max_{a \in \mathcal{A}(I)} \Bigg\{ \Big[\frac{1}{T}\sum_{t=1}^{T}\pi_{-i}^{\sigma_t}(I)\big(u_i(\sigma_t \mid I \to a, I) - u_i(\sigma_t, I)\big)\Big]^+ \\ + \sum_{I_0 \in \text{Succ}_i(I,a)} R^+_{T,i,\text{full}}(I_0) \Bigg\}, \end{aligned} \tag{19}$$

which is a per-action strengthening of Corollary 6 in Zinkevich et al. (2007). This bound is the foundation of CFR: it reduces whole-game regret minimization to independent regret-matching at each information set; the next three results make that reduction explicit. □

Theorem 6 (Unfolding the recursion; Lemma 7 of Zinkevich et al. (2007), extended to weighted sums)**.** *For every player i and information set I,*

$$R^+_{T,i,\text{full}}(I) \;\leq\; \sum_{I' \in \mathcal{D}_i(I)} R^+_{T,i,\text{imm}}(I'), \tag{20}$$

and, for the weighted versions obtained by replacing each uniform average $(1/T)\sum_{t=1}^{T}(\cdot)$ with $(1/W_T)\sum_{t=1}^{T} w_t(\cdot)$,

$$R^{+,w}_{i,\text{full}}(I) \;\leq\; \sum_{I' \in \mathcal{D}_i(I)} R^{+,w}_{i,\text{imm}}(I'). \tag{21}$$

Proof. Induction on $|\mathcal{D}_i(I)|$. If I has no successors, $\text{Succ}_i(I) = \emptyset$ and the positive part of equation 10 gives $R^+_{T,i,\text{full}}(I) \leq \max_a\{R_{T,i}(I,a)\}^+ = R^+_{T,i,\text{imm}}(I)$, the base case. Otherwise, taking the positive part of equation 10 and using $x^+ + y^+ \geq (x+y)^+$,

$$R^+_{T,i,\text{full}}(I) \leq R^+_{T,i,\text{imm}}(I) + \sum_{a \in \mathcal{A}(I)} \sum_{I_0 \in \text{Succ}_i(I,a)} R^+_{T,i,\text{full}}(I_0) = R^+_{T,i,\text{imm}}(I) + \sum_{I_0 \in \text{Succ}_i(I)} R^+_{T,i,\text{full}}(I_0), \tag{22}$$

where the equality collects the (disjoint) successor families of the actions. Each I_0 satisfies $|\mathcal{D}_i(I_0)| < |\mathcal{D}_i(I)|$ (no successor contains I in its descendants), so the induction hypothesis applies to each term, and $\mathcal{D}_i(I) = \{I\} \cup \bigcup_{I_0 \in \text{Succ}_i(I)} \mathcal{D}_i(I_0)$ with the union disjoint under perfect recall gives the claim. The weighted statement is obtained in the same way after replacing each uniform average $(1/T)\sum_{t=1}^{T}(\cdot)$ by $(1/W_T)\sum_{t=1}^{T} w_t(\cdot)$ throughout. This is immediate because w_t depends only on the round t: on each round the same scalar multiplies the immediate term and every successor-subtree term, so the recursion and the induction are unchanged. □

Theorem 7 (Local no-regret implies low overall regret; Theorem 3 of Zinkevich et al. (2007), extended to arbitrary nonnegative weights)**.** *For every player i and any weights $w_t \geq 0$, not all zero,*

$$R_{i,w} \;\leq\; \sum_{I \in \mathcal{J}_i} R^{+,w}_{i,\text{imm}}(I), \tag{23}$$

where $\mathcal{J}_i$ is the set of player i's information sets.

Theorem 7 is obtained by applying Theorem 6 at a formal root. At that root, the full subtree regret is exactly the whole-game overall regret $R_{i,w}$ from equation 9, namely the weighted overall regret against one fixed deviation. No monotonicity is needed here: the root reduction works for any nonnegative weights. Monotonicity is needed only in Theorem 9, at the partial-summation step.

Proof. Let I_{root} denote a formal root information set of player i with one action leading to the original tree. This changes no strategy and no payoff. Because I_{root} has only one action, its immediate regret is always zero, while choosing a deviation below I_{root} is exactly choosing a fixed strategy $\sigma^0 \in \Sigma_i$ on the original tree. Therefore

$$R^{+,w}_{i,\text{full}}(I_{\text{root}}) = (R_{i,w})^+ \geq R_{i,w}. \tag{24}$$

Applying Theorem 6 at I_{root} gives

$$R_{i,w} \;\leq\; R^{+,w}_{i,\text{full}}(I_{\text{root}}) \;\leq\; \sum_{I' \in \mathcal{D}_i(I_{\text{root}})} R^{+,w}_{i,\text{imm}}(I'). \tag{25}$$

Since the root term is zero and every original information set of player i lies below I_{root}, $\mathcal{D}_i(I_{\text{root}}) = \{I_{\text{root}}\} \cup \mathcal{J}_i$. Removing the root term yields

$$R_{i,w} \;\leq\; \sum_{I \in \mathcal{J}_i} R^{+,w}_{i,\text{imm}}(I). \tag{26}$$

This is the claim. □

Theorem 8 (Low regret implies equilibrium in two-player zero-sum games; weighted form of Theorem 2 of Zinkevich et al. (2007), proof following Xu et al. (2024))**.** *Let $(\bar{\boldsymbol{x}}^w, \bar{\boldsymbol{y}}^w)$ be the w-weighted average sequence-form profile generated by the played strategies. Then*

$$e(\bar{\boldsymbol{x}}^w, \bar{\boldsymbol{y}}^w) \;\leq\; \tfrac{1}{2}\big(R_{1,w} + R_{2,w}\big). \tag{27}$$

In particular, if both players' weighted overall regret is at most ε, the weighted average profile is a Nash equilibrium up to ε.

Proof. Write δ_1, δ_2 for the two incentives to deviate from the average profile (Section 2). Their sum telescopes, the payoff of the average profile appearing with opposite signs:

$$\delta_1 + \delta_2 \;=\; \max_{\dot{\boldsymbol{y}}' \in \mathcal{Y}} \bar{\boldsymbol{x}}^{w\top} \boldsymbol{A} \dot{\boldsymbol{y}}' \;-\; \min_{\dot{\boldsymbol{x}}' \in \mathcal{X}} \dot{\boldsymbol{x}}'^{\top} \boldsymbol{A} \bar{\boldsymbol{y}}^w. \tag{28}$$

We now write the two overall regrets in the same terms. Under perfect recall, behavioral strategies and sequence-form realizations determine each other (Zinkevich et al., 2007), so the maximization over fixed deviations σ^0 in equation 9 becomes, in sequence form, the maximization over $\mathcal{X}$ for player 1 and over $\mathcal{Y}$ for player 2. Since $\boldsymbol{A}$ encodes player 1's losses, the two players' utilities are $u_1(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}}) = -\dot{\boldsymbol{x}}^\top \boldsymbol{A} \dot{\boldsymbol{y}}$ and $u_2(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}}) = +\dot{\boldsymbol{x}}^\top \boldsymbol{A} \dot{\boldsymbol{y}}$. Let $\bar{u} = \frac{1}{W_T} \sum_{t=1}^T w_t \, \dot{\boldsymbol{x}}^{t\top} \boldsymbol{A} \dot{\boldsymbol{y}}^t$ be the average played payoff, at once player 2's average payoff and player 1's average loss. Expanding equation 9 with these utilities,

$$\begin{aligned}
R_{1,w} &= \max_{\dot{\boldsymbol{x}}' \in \mathcal{X}} \frac{1}{W_T} \sum_{t=1}^T w_t \left(\dot{\boldsymbol{x}}^t - \dot{\boldsymbol{x}}'\right)^\top \boldsymbol{A} \dot{\boldsymbol{y}}^t \;=\; \bar{u} \;-\; \min_{\dot{\boldsymbol{x}}' \in \mathcal{X}} \dot{\boldsymbol{x}}'^{\top} \boldsymbol{A} \bar{\boldsymbol{y}}^w, \\
R_{2,w} &= \max_{\dot{\boldsymbol{y}}' \in \mathcal{Y}} \frac{1}{W_T} \sum_{t=1}^T w_t \, \dot{\boldsymbol{x}}^{t\top} \boldsymbol{A} \big(\dot{\boldsymbol{y}}' - \dot{\boldsymbol{y}}^t\big) \;=\; \max_{\dot{\boldsymbol{y}}' \in \mathcal{Y}} \bar{\boldsymbol{x}}^{w\top} \boldsymbol{A} \dot{\boldsymbol{y}}' \;-\; \bar{u},
\end{aligned} \tag{29}$$

For player 1, the second equality averages the played $\dot{\boldsymbol{y}}^t$ inside the bilinear form, so $\frac{1}{W_T} \sum_{t=1}^T w_t \, \dot{\boldsymbol{x}}'^{\top} \boldsymbol{A} \dot{\boldsymbol{y}}^t = \dot{\boldsymbol{x}}'^{\top} \boldsymbol{A} \bar{\boldsymbol{y}}^w$, and the constant $\bar{u}$ can be pulled outside the maximization. For player 2, the same step averages the played $\dot{\boldsymbol{x}}^t$ to $\bar{\boldsymbol{x}}^w$, and the constant $\bar{u}$ drops out. Adding the two lines in equation 29, $\bar{u}$ cancels and what remains is equation 28: $R_{1,w} + R_{2,w} = \delta_1 + \delta_2$. Averaging, $e = (\delta_1 + \delta_2)/2 \leq (R_{1,w} + R_{2,w})/2$. If both regrets are at most ε, then $e \leq \varepsilon$, which is the definition of a Nash equilibrium up to ε. □

Theorem 9 (Weighted averaging; parameterized form of Theorem 1 of Brown & Sandholm (2019))**.** *Suppose each player runs a regret-matching+ accumulator $\boldsymbol{R}_j^t = (\boldsymbol{R}_j^{t-1} + \boldsymbol{r}_j^t)^+$, with $\boldsymbol{R}_j^0 = \boldsymbol{0}$, at every information set j, where $\boldsymbol{r}_j^t$ is the instantaneous counterfactual regret at j in round t, and suppose that at round T every accumulated-regret coordinate obeys $\boldsymbol{R}_j^T[a] \leq B_j$ for a constant*

$B_j \geq 0$ depending on j but not on a. Then for any non-decreasing weights $w_t \geq 0$, not all zero, the w-weighted average strategy profile is a Nash equilibrium up to

$$\frac{w_T}{W_T}\,\frac{1}{2}\left(\sum_{j\in\mathcal{J}_1} B_j + \sum_{j\in\mathcal{J}_2} B_j\right). \tag{30}$$

Brown & Sandholm (2019) state this theorem with the standard RM$^+$ accumulator bound from their Lemmas 2 and 3 (equivalently, the RM$^+$ bound of Tammelin (2014)). Here we write the theorem in a more general form: assume only that each information set j has some final accumulator bound B_j. The proof first turns B_j into a bound on the weighted immediate regret at j, then applies Theorems 7 and 8 to obtain the equilibrium bound. Appendix C later proves the corresponding bounds for IRR-RM$^+$ and IRR-PRM$^+$ (Lemmas 2 and 3). For the DCFR specialization of Brown & Sandholm (2019), one inserts

$$B_j = G\sqrt{n_j T}, \tag{31}$$

where G is the payoff range in our notation. Then Theorem 9 gives, for each player i,

$$R_{i,w} \;\leq\; \frac{w_T}{W_T}\sum_{j\in\mathcal{J}_i} G\sqrt{n_j T} \;\leq\; \frac{w_T}{W_T}\,G\,|\mathcal{J}_i|\,\sqrt{nT}, \tag{32}$$

where $n = \max_j n_j$ and $|\mathcal{I}| = |\mathcal{J}_1| + |\mathcal{J}_2|$. Thus the weighted average profile is a $\frac{w_T}{W_T}\,G\,|\mathcal{I}|\,\sqrt{nT}$-Nash equilibrium, which is the DCFR formula of Brown & Sandholm (2019) in our notation.

Proof. Fix an information set j of player i and an action a. The clip can only add mass, $\boldsymbol{R}_j^t[a] \geq \boldsymbol{R}_j^{t-1}[a] + \boldsymbol{r}_j^t[a]$, so $\boldsymbol{r}_j^t[a] \leq \boldsymbol{R}_j^t[a] - \boldsymbol{R}_j^{t-1}[a]$. Summing with weights and applying partial summation gives

$$\sum_{t=1}^{T} w_t\,\boldsymbol{r}_j^t[a] \;\leq\; \sum_{t=1}^{T} w_t\big(\boldsymbol{R}_j^t[a] - \boldsymbol{R}_j^{t-1}[a]\big) \;=\; w_T\,\boldsymbol{R}_j^T[a] + \sum_{t=1}^{T-1}(w_t - w_{t+1})\,\boldsymbol{R}_j^t[a] \;\leq\; w_T\,B_j, \tag{33}$$

where the last inequality uses $w_t \leq w_{t+1}$ and $\boldsymbol{R}_j^t[a] \geq 0$. Dividing by W_T gives, for each action a at j,

$$\frac{1}{W_T}\sum_{t=1}^{T} w_t\,\boldsymbol{r}_j^t[a] \;\leq\; \frac{w_T}{W_T}\,B_j. \tag{34}$$

Maximizing over a yields the weighted immediate regret at j, $R_{i,\mathrm{imm}}^{+,w}(j) \leq (w_T/W_T)\,B_j$. Therefore, for each player i,

$$R_{i,w} \;\leq\; \sum_{j\in\mathcal{J}_i} R_{i,\mathrm{imm}}^{+,w}(j) \;\leq\; \frac{w_T}{W_T}\sum_{j\in\mathcal{J}_i} B_j, \tag{35}$$

by Theorem 7. Applying Theorem 8 to the two players gives

$$e(\bar{\boldsymbol{x}}^w, \bar{\boldsymbol{y}}^w) \;\leq\; \frac{1}{2}(R_{1,w} + R_{2,w}) \;\leq\; \frac{w_T}{W_T}\,\frac{1}{2}\left(\sum_{j\in\mathcal{J}_1} B_j + \sum_{j\in\mathcal{J}_2} B_j\right). \tag{36}$$

This is the claim. □

C Omitted proofs

Setup. Throughout, fix an information set j and let τ be its reduction round. We abbreviate the round-τ objects by $\boldsymbol{x} = \boldsymbol{x}_j^\tau$, $\boldsymbol{\ell} = \boldsymbol{\ell}_j^\tau$, and $\boldsymbol{r} = \boldsymbol{r}_j^\tau$. By definition of the reduction round, the accumulated regret is still zero before the update, $\boldsymbol{R}_j^{\tau-1} = \mathbf{0}$, and the round is played at the uniform strategy $\boldsymbol{x} = \mathbf{1}/n_j$. Under PCFR+, the predictive readout does not change this: for $\tau > 1$, one has $\boldsymbol{m}_j^\tau = \boldsymbol{r}_j^{\tau-1}$, while $\boldsymbol{m}_j^1 = \mathbf{0}$; since τ is the first round whose instantaneous regret has a positive component, $\boldsymbol{m}_j^\tau$ has no positive component, and therefore $[\boldsymbol{R}_j^{\tau-1} + \boldsymbol{m}_j^\tau]^+ = \mathbf{0}$ as well. Write

$O = \sum_a [\boldsymbol{r}[a]]^+ > 0$, let $\boldsymbol{x}^{\tau+1} = [\boldsymbol{r}]^+/O$ be the next strategy, and set $\boldsymbol{\sigma} = \lambda \boldsymbol{x}^{\tau+1} + (1-\lambda)\boldsymbol{x}$ to be the interpolated baseline of equation 2. Also write

$$S' = \sum_a [\langle \boldsymbol{\sigma}, \boldsymbol{\ell} \rangle - \boldsymbol{\ell}[a]]^+, \tag{37}$$

so that $s = S'/O$; if $S' = 0$, Algorithm 1 sets $s := 1$. The counterfactual loss $\boldsymbol{\ell}$ generates the instantaneous regret $\boldsymbol{r} = \langle \boldsymbol{x}, \boldsymbol{\ell} \rangle \mathbf{1} - \boldsymbol{\ell}$, so an action carries positive regret exactly when its loss is below the average loss of the current strategy,

$$\boldsymbol{r}[a] > 0 \iff \boldsymbol{\ell}[a] < \langle \boldsymbol{x}, \boldsymbol{\ell} \rangle. \tag{38}$$

We also use

$$\langle \boldsymbol{x}, \boldsymbol{r} \rangle = \langle \boldsymbol{x}, \boldsymbol{\ell} \rangle \langle \boldsymbol{x}, \mathbf{1} \rangle - \langle \boldsymbol{x}, \boldsymbol{\ell} \rangle = 0, \tag{39}$$

because $\boldsymbol{x}$ is a distribution.

Theorem 1. *The scale factor of Algorithm 1 satisfies $s \leq 1$.*

Proof of Theorem 1. The setup gives $O > 0$; if the re-evaluated mass is zero ($S' = 0$), Algorithm 1 sets $s := 1$, so suppose $S' > 0$. By equation 38, if $\boldsymbol{x}^{\tau+1}[a] > 0$ then $\boldsymbol{\ell}[a] < \langle \boldsymbol{x}, \boldsymbol{\ell} \rangle$. Since $\boldsymbol{x}^{\tau+1}$ is a distribution,

$$\langle \boldsymbol{x}^{\tau+1}, \boldsymbol{\ell} \rangle = \sum_{a:\, \boldsymbol{x}^{\tau+1}[a] > 0} \boldsymbol{x}^{\tau+1}[a]\, \boldsymbol{\ell}[a] < \sum_{a:\, \boldsymbol{x}^{\tau+1}[a] > 0} \boldsymbol{x}^{\tau+1}[a]\, \langle \boldsymbol{x}, \boldsymbol{\ell} \rangle = \langle \boldsymbol{x}, \boldsymbol{\ell} \rangle. \tag{40}$$

Hence

$$\langle \boldsymbol{\sigma}, \boldsymbol{\ell} \rangle = (1-\lambda)\, \langle \boldsymbol{x}, \boldsymbol{\ell} \rangle + \lambda\, \langle \boldsymbol{x}^{\tau+1}, \boldsymbol{\ell} \rangle \;<\; \langle \boldsymbol{x}, \boldsymbol{\ell} \rangle. \tag{41}$$

Lowering the average loss lowers the excess of every action, and the positive part is monotone:

$$[\langle \boldsymbol{\sigma}, \boldsymbol{\ell} \rangle \mathbf{1} - \boldsymbol{\ell}]^+ \;\leq\; [\langle \boldsymbol{x}, \boldsymbol{\ell} \rangle \mathbf{1} - \boldsymbol{\ell}]^+ = [\boldsymbol{r}]^+. \tag{42}$$

Summing the coordinates gives $S' \leq O$, hence $s = S'/O \leq 1$. This is exactly the claim. □

No-regret guarantee on the RM+ base (Theorem 2). Let the algorithm face an arbitrary (possibly adaptive) sequence of loss vectors $\boldsymbol{\ell}^1, \ldots, \boldsymbol{\ell}^T$: write $\boldsymbol{x}^t$ for its current strategy, $\boldsymbol{r}^t = \langle \boldsymbol{x}^t, \boldsymbol{\ell}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t$ for the induced instantaneous regret ($\|\boldsymbol{r}^t\|_\infty \leq G$), and $\boldsymbol{R}^t$ for the accumulated regret: $\boldsymbol{R}^t = (\boldsymbol{R}^{t-1} + \boldsymbol{r}^t)^+$ at ordinary rounds and $\boldsymbol{R}^\tau = s\,(\boldsymbol{r}^\tau)^+$ at the reduction round, which occurs at most once; $\kappa \in \{0, 1\}$ indicates whether it occurred.

Lemma 1 (The RM$^+$ accumulator inequality of Tammelin (2014), plus the reduction term)**.** *For every action a and horizon T,*

$$\sum_{t=1}^{T} \boldsymbol{r}^t[a] \;\leq\; \boldsymbol{R}^T[a] \;+\; \kappa\,(1-s)\,(\boldsymbol{r}^\tau[a])^+, \tag{43}$$

where the second term on the right is present only when the reduction occurred ($\kappa = 1$).

Proof. At an ordinary round the positive part can only increase the accumulator, $\boldsymbol{R}^t[a] = [\boldsymbol{R}^{t-1}[a] + \boldsymbol{r}^t[a]]^+ \geq \boldsymbol{R}^{t-1}[a] + \boldsymbol{r}^t[a]$, so $\boldsymbol{r}^t[a] \leq \boldsymbol{R}^t[a] - \boldsymbol{R}^{t-1}[a]$. At the reduction round $\boldsymbol{R}^{\tau-1} = \mathbf{0}$ and $\boldsymbol{R}^\tau[a] = s\,(\boldsymbol{r}^\tau[a])^+$, so $\boldsymbol{r}^\tau[a] \leq \boldsymbol{R}^\tau[a] + (1-s)\,(\boldsymbol{r}^\tau[a])^+$. Summing over $t = 1, \ldots, T$ and telescoping (the accumulator starts at zero) gives the claim. □

Lemma 2 (Accumulated-regret bound; the RM$^+$ potential bound of Tammelin (2014))**.** *For every horizon T,* $\max_a \boldsymbol{R}^T[a] \leq \|\boldsymbol{R}^T\|_2 \leq G\sqrt{|\mathcal{A}|T}$.

Proof. The first inequality is immediate. For the second, run the RM$^+$ potential $\Phi^t = \|\boldsymbol{R}^t\|_2^2$. At an ordinary round $\boldsymbol{R}^t = (\boldsymbol{R}^{t-1} + \boldsymbol{r}^t)^+$, so

$$\Phi^t \;\leq\; \|\boldsymbol{R}^{t-1} + \boldsymbol{r}^t\|_2^2 \;=\; \Phi^{t-1} + 2\,\langle \boldsymbol{R}^{t-1}, \boldsymbol{r}^t \rangle + \|\boldsymbol{r}^t\|_2^2, \tag{44}$$

and the cross term vanishes: $\boldsymbol{R}^{t-1} \geq \mathbf{0}$ and $\boldsymbol{x}^t = \boldsymbol{R}^{t-1}/\|\boldsymbol{R}^{t-1}\|_1$ (when $\boldsymbol{R}^{t-1} = \mathbf{0}$ the term is trivially zero), so $\langle \boldsymbol{R}^{t-1}, \boldsymbol{r}^t \rangle = \|\boldsymbol{R}^{t-1}\|_1 \langle \boldsymbol{x}^t, \boldsymbol{r}^t \rangle = 0$ by identity equation 39. At the reduction

round $\boldsymbol{R}^{\tau-1} = \mathbf{0}$ and $\Phi^\tau = s^2\|(\boldsymbol{r}^\tau)^+\|_2^2 \le \|\boldsymbol{r}^\tau\|_2^2$, so the same increment bound holds. Hence, with $\Phi^0 = 0$ since the accumulator starts at zero, $\Phi^T \le \sum_{t=1}^T \|\boldsymbol{r}^t\|_2^2 \le |\mathcal{A}|TG^2$ and $\|\boldsymbol{R}^T\|_2 = \sqrt{\Phi^T} \le G\sqrt{|\mathcal{A}|T}$. □

Theorem 2 (No-regret guarantee for IRR-RM+). *Let RM+ face an arbitrary (possibly adaptive) sequence of loss vectors $\boldsymbol{\ell}^1, \dots, \boldsymbol{\ell}^T$ on action set $\mathcal{A}$, whose induced instantaneous regrets satisfy $\|\boldsymbol{r}^t\|_\infty \le G$ for all t. Let $\kappa \in \{0, 1\}$ indicate whether the reduction occurred; if $\kappa = 1$, let τ be the reduction round and $s \in (0, 1]$ the scale used there. The resulting algorithm remains no-regret: its T-round regret $R_{\text{IRR}}^T = \max_a \sum_{t=1}^T \boldsymbol{r}^t[a]$ obeys*

$$R_{\text{IRR}}^T \le G\sqrt{|\mathcal{A}|T} + G\,\kappa, \tag{45}$$

so the excess over the standard RM+ bound is a constant independent of T.

Proof of Theorem 2. Combining Lemmas 1 and 2,

$$R_{\text{IRR}}^T \;=\; \max_a \sum_{t=1}^T \boldsymbol{r}^t[a] \;\le\; \max_a \boldsymbol{R}^T[a] \;+\; \kappa\,(1-s)\,\big\|(\boldsymbol{r}^\tau)^+\big\|_\infty \;\le\; G\sqrt{|\mathcal{A}|T} + G\,\kappa, \tag{46}$$

since $(1-s)\,\|(\boldsymbol{r}^\tau)^+\|_\infty \le \|\boldsymbol{r}^\tau\|_\infty \le G$ when $\kappa = 1$; if $\kappa = 0$, the second term is absent. A reduction with $S' = 0$ sets $s = 1$ and contributes nothing. The excess $G\kappa$ is a constant independent of T, so $R_{\text{IRR}}^T/T \to 0$ as $T \to \infty$: the algorithm is no-regret. □

No-regret guarantee on the PRM+ base (Theorem 3). Viewing one information set as a standalone online learner on local action set $\mathcal{A}$, we suppress the subscript j below. The accumulator obeys the same recurrence as above, so Lemma 1 is unchanged. Write $\boldsymbol{m}^1 = \mathbf{0}$ and $\boldsymbol{m}^t = \boldsymbol{r}^{t-1}$ for $t \ge 2$, except that after the single replacement of $\boldsymbol{r}^\tau$ by $s\,\boldsymbol{r}^\tau$, the same scaled vector feeds both the accumulator and the next prediction (Algorithm 2).

Lemma 3 (Predicted-readout accumulator bound)**.** *For every horizon T, under the prediction sequence above,* $\max_a \boldsymbol{R}^T[a] \le \|\boldsymbol{R}^T\|_2 \le \sqrt{3}\,G\sqrt{|\mathcal{A}|T}$.

Proof. Run the potential $\Phi^t = \|\boldsymbol{R}^t\|_2^2$. At an ordinary round, $\boldsymbol{R}^t = [\boldsymbol{R}^{t-1} + \boldsymbol{r}^t]^+$, so

$$\Phi^t \;\le\; \|\boldsymbol{R}^{t-1} + \boldsymbol{r}^t\|_2^2 \;=\; \Phi^{t-1} + 2\,\langle \boldsymbol{R}^{t-1}, \boldsymbol{r}^t\rangle + \|\boldsymbol{r}^t\|_2^2. \tag{47}$$

At the reduction round, the same bound holds because $\boldsymbol{R}^{\tau-1} = \mathbf{0}$ gives $\Phi^\tau = s^2\|(\boldsymbol{r}^\tau)^+\|_2^2 \le \|\boldsymbol{r}^\tau\|_2^2$, exactly as in the proof of Lemma 2. Decompose the cross term through the predicted readout $\tilde{\boldsymbol{R}}^t = [\boldsymbol{R}^{t-1} + \boldsymbol{m}^t]^+$ and the clip correction $\boldsymbol{c}^t = \boldsymbol{R}^{t-1} - \tilde{\boldsymbol{R}}^t$:

$$\langle \boldsymbol{R}^{t-1}, \boldsymbol{r}^t\rangle = \langle \tilde{\boldsymbol{R}}^t, \boldsymbol{r}^t\rangle + \langle \boldsymbol{c}^t, \boldsymbol{r}^t\rangle. \tag{48}$$

The first term vanishes exactly as in the proof of Lemma 2, with $\tilde{\boldsymbol{R}}^t$ in place of $\boldsymbol{R}^{t-1}$. The correction is controlled by the prediction componentwise: if $\boldsymbol{R}^{t-1}[a] + \boldsymbol{m}^t[a] > 0$ then $\boldsymbol{c}^t[a] = -\boldsymbol{m}^t[a]$, while otherwise $0 \le \boldsymbol{c}^t[a] = \boldsymbol{R}^{t-1}[a] \le -\boldsymbol{m}^t[a]$; in both cases $|\boldsymbol{c}^t[a]| \le |\boldsymbol{m}^t[a]|$, hence $\|\boldsymbol{c}^t\|_2 \le \|\boldsymbol{m}^t\|_2$. Cauchy–Schwarz and $2ab \le a^2 + b^2$ then give

$$2\,\langle \boldsymbol{R}^{t-1}, \boldsymbol{r}^t\rangle \;\le\; 2\,\big|\langle \boldsymbol{c}^t, \boldsymbol{r}^t\rangle\big| \;\le\; \|\boldsymbol{m}^t\|_2^2 + \|\boldsymbol{r}^t\|_2^2. \tag{49}$$

Summing the increment bound over $t = 1, \dots, T$ (with $\Phi^0 = 0$ as before) therefore gives

$$\Phi^T \;\le\; \sum_{t=1}^T \big(\|\boldsymbol{m}^t\|_2^2 + 2\,\|\boldsymbol{r}^t\|_2^2\big). \tag{50}$$

The prediction is the previous round's regret, scaled by $s \le 1$ once at $t = \tau + 1$: $\boldsymbol{m}^1 = \mathbf{0}$ and $\|\boldsymbol{m}^t\|_2 \le \|\boldsymbol{r}^{t-1}\|_2$ for $t \ge 2$, so $\sum_{t=1}^T \|\boldsymbol{m}^t\|_2^2 \le \sum_{t=1}^{T-1} \|\boldsymbol{r}^t\|_2^2 \le \sum_{t=1}^T \|\boldsymbol{r}^t\|_2^2$. With $\|\boldsymbol{r}^t\|_2 \le \sqrt{|\mathcal{A}|}\,\|\boldsymbol{r}^t\|_\infty \le \sqrt{|\mathcal{A}|}\,G$,

$$\Phi^T \;\le\; 3\sum_{t=1}^T \|\boldsymbol{r}^t\|_2^2 \;\le\; 3\,|\mathcal{A}|\,T\,G^2, \qquad \text{so} \qquad \|\boldsymbol{R}^T\|_2 = \sqrt{\Phi^T} \le \sqrt{3}\,G\sqrt{|\mathcal{A}|T}, \tag{51}$$

and $\max_a \boldsymbol{R}^T[a] \le \|\boldsymbol{R}^T\|_2$. □

Theorem 3 (No-regret guarantee for IRR-PRM+). *Let PRM+ face an arbitrary (possibly adaptive) sequence of loss vectors $\boldsymbol{\ell}^1, \dots, \boldsymbol{\ell}^T$ on action set $\mathcal{A}$, whose induced instantaneous regrets satisfy $\|\boldsymbol{r}^t\|_\infty \le G$ for all t. Let $\kappa \in \{0,1\}$ indicate whether the reduction occurred; if $\kappa = 1$, let τ be the reduction round and $s \in (0,1]$ the scale used there. Run the standard prediction rule $\boldsymbol{m}^1 = \mathbf{0}$ and $\boldsymbol{m}^{t+1} = \boldsymbol{r}^t$, except that if $\kappa = 1$, then $\boldsymbol{m}^{\tau+1} = s\,\boldsymbol{r}^\tau$. The resulting algorithm remains no-regret: its T-round regret $R^T_{\text{IRR}} = \max_a \sum_{t=1}^T \boldsymbol{r}^t[a]$ obeys*

$$R^T_{\text{IRR}} \le \sqrt{3}\,G\sqrt{|\mathcal{A}|T} + G\,\kappa. \tag{52}$$

Proof of Theorem 3. The proof of Theorem 2 is unchanged with Lemma 2 replaced by Lemma 3. Thus

$$R^T_{\text{IRR}} \;\le\; \max_a \boldsymbol{R}^T[a] \;+\kappa\,(1-s)\,\big\|(\boldsymbol{r}^\tau)^+\big\|_\infty \;\le\; \sqrt{3}\,G\sqrt{|\mathcal{A}|T} + G\,\kappa, \tag{53}$$

because $(1-s)\,\|(\boldsymbol{r}^\tau)^+\|_\infty \le \|\boldsymbol{r}^\tau\|_\infty \le G$. The excess $G\kappa$ is again a constant independent of T, so the algorithm is no-regret. □

Rate in extensive-form games (Corollary 1). Fix non-decreasing weights $w_t \ge 0$, not all zero, and write $W_T = \sum_{t=1}^T w_t$. Let $n = \max_j n_j$. Let j be an information set of one player with n_j actions, and let $\tilde{\boldsymbol{r}}_j^t$ be the sequence fed to the accumulator: $\boldsymbol{r}_j^t$ at ordinary rounds and $s\,\boldsymbol{r}_j^\tau$ at the reduction round, which requires $\boldsymbol{R}_j^{\tau-1} = \mathbf{0}$ and a positive instantaneous regret. The partial-summation step in the proof of Theorem 9 applies verbatim to this sequence. The sequence fed to the accumulator coincides with the counterfactual regret of the played strategies except at τ, where it withholds the share $(1-s)\,\boldsymbol{r}_j^\tau[a]$, so for every action a

$$\sum_{t=1}^T w_t\,\boldsymbol{r}_j^t[a] \;\le\; w_T\,\boldsymbol{R}_j^T[a] \;+\; \kappa_j\,(1-s)\,w_\tau\,\big(\boldsymbol{r}_j^\tau[a]\big)^+, \tag{54}$$

with $\kappa_j \in \{0,1\}$ flagging whether the reduction occurred at j. By Lemmas 2 and 3, $\boldsymbol{R}_j^T[a] \le cG\sqrt{n_j T}$ with $c = 1$ on the RM$^+$ base and $c = \sqrt{3}$ on the PRM$^+$ base, and $\|(\boldsymbol{r}_j^\tau)^+\|_\infty \le G$, so with $w_\tau \le w_T$ and $\sqrt{n_j} \le \sqrt{n}$,

$$\sum_{t=1}^T w_t\,\boldsymbol{r}_j^t[a] \;\le\; w_T\,G\,\big(c\sqrt{nT}+1\big). \tag{55}$$

Dividing by W_T and inserting $B_j = G\,(c\sqrt{nT}+1)$ into Theorem 9 bounds the exploitability of the w-weighted average profile by $\frac{w_T}{W_T}\,|\mathcal{I}|\,G\,(c\sqrt{nT}+1)$, the first claim of the corollary with c absorbed into $O(\cdot)$. For $w_t = t^\gamma$ with $\gamma \ge 0$, $W_T \ge T^{\gamma+1}/(\gamma+1)$ and $w_T = T^\gamma$, so $w_T/W_T \le (\gamma+1)/T$ and the gap is $O\big((\gamma+1)\,G\,|\mathcal{I}|\,(\sqrt{n}+1)/\sqrt{T}\big)$.

From approachability to online mirror descent (Theorem 4). The correspondence used in Section 5.2 has three steps. First, regret matching on the simplex is a Blackwell approachability procedure (Hart & Mas-Colell, 2000). Second, Abernethy et al. (2011) show that for a closed convex cone C, Blackwell approachability of C can be implemented through no-regret online linear optimization over the polar cone C°. In the simplex regret problem the target cone is $\mathbb{R}^{|\mathcal{A}|}_{\le 0}$, so the oracle domain is $C^\circ = \mathbb{R}^{|\mathcal{A}|}_{\ge 0}$. Third, Farina et al. (2021) instantiate that oracle as predictive OMD with regularizer $\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$; after applying the forcing action to the resulting OMD readout, the induced simplex iterates coincide with RM$^+$ and, with prediction $\boldsymbol{m}^{t+1} = \boldsymbol{r}^t$, with PRM$^+$. In this instantiation the mirror step is the projection $[\cdot]^+$ onto $\mathbb{R}^{|\mathcal{A}|}_{\ge 0}$, while the forcing action is the ℓ_1 normalization that maps the nonnegative OMD readout back to the simplex. Theorem 4 proves only the last step of this chain: our one-time reduction is a single deviation from the constant-step-size local OMD update, and it is exact because it occurs at the zero state; the forcing action is then applied as usual. Like the rest of this paragraph, the statement is local to one information set; the extensive-form rate statements use the global quantity $n = \max_j n_j$.

Theorem 4 (IRR-PRM+ as a one-time OMD step-size change). *Fix an arbitrary (possibly adaptive) sequence of losses and an information set with local action set $\mathcal{A}$, and let a run of IRR-PRM+ (Algorithm 2) reduce at round τ with scale $s \in (0, 1]$. When the approachability construction of Farina et al. (2021, Algorithm 3) is set up with oracle domain $\mathbb{R}^{|\mathcal{A}|}_{\geq 0}$ and the OMD update of Algorithm 3 as the regret minimizer, with step sizes $\eta_\tau = s$ and $\eta_t = 1$ otherwise, observed payoffs $\boldsymbol{u}^t = \boldsymbol{r}_j^t$, and the uniform strategy at the zero readout, it produces the same strategy sequence as IRR-PRM+.*

Proof of Theorem 4. Fix the loss sequence and run the OMD update of Algorithm 3 and IRR-PRM+ (Algorithm 2) on it. Write $\boldsymbol{R}_j^t$, $\boldsymbol{m}^t$, $\tilde{\boldsymbol{R}}_j^t$, $\boldsymbol{x}_j^t$ for the IRR-PRM+ accumulator, stored prediction, pre-normalization readout, and played strategy. We prove by induction on t that

$$\boldsymbol{z}^t = \boldsymbol{R}_j^t. \tag{56}$$

The strategy statement then follows because, after applying the forcing action to the OMD readout $\boldsymbol{\theta}^t = [\boldsymbol{z}^{t-1} + \boldsymbol{v}^t]^+$, the resulting simplex action agrees with $\boldsymbol{x}_j^t$ whenever $\boldsymbol{\theta}^t$ and $\tilde{\boldsymbol{R}}_j^t$ coincide; both procedures use ℓ_1 normalization for nonzero vectors and the same uniform choice at zero.

Both algorithms start from zero state with zero stored prediction, so the base case holds. For rounds $t < \tau$, no reduction has occurred and both updates are exactly the standard PRM$^+$ recurrence with unit step size, so the invariant holds.

At round τ, the reduction condition gives $\boldsymbol{R}_j^{\tau-1} = \boldsymbol{0} = \boldsymbol{z}^{\tau-1}$. The pre-decision readout still uses the previous stored vector, so $\boldsymbol{v}^\tau = \boldsymbol{u}^{\tau-1} = \boldsymbol{r}_j^{\tau-1} = \boldsymbol{m}^\tau$ (zero if $\tau = 1$), and therefore

$$\boldsymbol{\theta}^\tau = [\boldsymbol{z}^{\tau-1} + \boldsymbol{v}^\tau]^+ = [\boldsymbol{m}^\tau]^+ = \tilde{\boldsymbol{R}}_j^\tau. \tag{57}$$

Thus, after the forcing action, the two procedures choose the same strategy at round τ. For the post-update state,

$$\boldsymbol{z}^\tau = [s\,\boldsymbol{u}^\tau]^+ = [s\,\boldsymbol{r}_j^\tau]^+ = \boldsymbol{R}_j^\tau, \tag{58}$$

because IRR-PRM+ also scales that one regret vector by s before accumulating it.

At round $\tau + 1$, the key point is that the OMD update stores the previous observed vector scaled by its step size: by line 7 of Algorithm 3,

$$\boldsymbol{v}^{\tau+1} = \eta_\tau \boldsymbol{u}^\tau = s\,\boldsymbol{r}_j^\tau. \tag{59}$$

This is exactly the same one-time scaled vector that IRR-PRM+ stores as its next prediction, $\boldsymbol{m}^{\tau+1} = s\,\boldsymbol{r}_j^\tau$. Since also $\boldsymbol{z}^\tau = \boldsymbol{R}_j^\tau$, the two pre-normalization readouts coincide exactly:

$$\boldsymbol{\theta}^{\tau+1} = [\boldsymbol{z}^\tau + \boldsymbol{v}^{\tau+1}]^+ = [\boldsymbol{R}_j^\tau + \boldsymbol{m}^{\tau+1}]^+ = \tilde{\boldsymbol{R}}_j^{\tau+1}. \tag{60}$$

Hence the forcing action again yields the same strategy. The post-update states also coincide, because from this round onward both updates use the fresh unscaled regret vector with unit step size:

$$\boldsymbol{z}^{\tau+1} = [\boldsymbol{z}^\tau + \boldsymbol{u}^{\tau+1}]^+ = [\boldsymbol{R}_j^\tau + \boldsymbol{r}_j^{\tau+1}]^+ = \boldsymbol{R}_j^{\tau+1}. \tag{61}$$

For all rounds $t \geq \tau + 2$, both the accumulator update and the stored prediction again carry fresh unscaled vectors and $\eta_t = 1$, so the two OMD recurrences are identical and the induction closes. □