Minimizing Weighted Counterfactual Regret with Optimistic Online Mirror Descent

Hang Xu[1,2] , **Kai Li**[1,2,∗] , **Bingyun Liu**[1,2] , **Haobo Fu**[6] ,
Qiang Fu[6] , **Junliang Xing**[5,∗] and **Jian Cheng**[1,3,4]

[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]School of Future Technology, University of Chinese Academy of Sciences
[4]AiRiA
[5]Tsinghua University
[6]Tencent AI Lab
{xuhang2020, kai.li, liubingyun2021, jian.cheng}@ia.ac.cn,
{haobofu, leonfu}@tencent.com, jlxing@tsinghua.edu.cn

Abstract

Counterfactual regret minimization (CFR) is a family of algorithms for effectively solving imperfect-information games. It decomposes the total regret into counterfactual regrets, utilizing local regret minimization algorithms, such as Regret Matching (RM) or RM+, to minimize them. Recent research establishes a connection between Online Mirror Descent (OMD) and RM+, paving the way for an optimistic variant PRM+ and its extension PCFR+. However, PCFR+ assigns uniform weights for each iteration when determining regrets, leading to substantial regrets when facing dominated actions. This work explores minimizing weighted counterfactual regret with optimistic OMD, resulting in a novel CFR variant PDCFR+. It integrates PCFR+ and Discounted CFR (DCFR) in a principled manner, swiftly mitigating negative effects of dominated actions and consistently leveraging predictions to accelerate convergence. Theoretical analyses prove that PDCFR+ converges to a Nash equilibrium, particularly under distinct weighting schemes for regrets and average strategies. Experimental results demonstrate PDCFR+'s fast convergence in common imperfect-information games. The code is available at https://github.com/rpSebastian/PDCFRPlus.

1 Introduction

Imperfect-information games (IIGs) model strategic interactions between players with hidden information. Solving such games is challenging since players must reason under uncertainty about opponents' private information. The hidden information plays an essential role in real-world situations such

as medical treatment [Sandholm, 2015], negotiation [Gratch *et al.*, 2016], and security [Lisy *et al.*, 2016], making the research on IIGs theoretically and practically crucial.

In this work, we focus on solving two-player zero-sum (2p0s) IIGs. The typical goal in these games is to find an (approximate) Nash equilibrium (NE) in which no player can benefit from deviating unilaterally from the equilibrium. The common iterative approach minimizes total regrets of both players so that their average strategies over time converge to a NE. The family of counterfactual regret minimization (CFR) algorithms [Zinkevich *et al.*, 2007] decomposes the total regret into the sum of counterfactual regrets associated with decision nodes. Then it employs a local regret minimization algorithm, such as Regret Matching (RM) [Hart and Mas-Colell, 2000] or its variant RM+ [Tammelin, 2014], at each decision node to effectively minimize counterfactual regret. Due to the sound theoretical guarantee and strong empirical performance, CFR and its variants have enabled several major breakthroughs in this field [Bowling *et al.*, 2015; Moravčík *et al.*, 2017; Brown and Sandholm, 2018; Brown and Sandholm, 2019b].

Besides CFR, Online Mirror Descent (OMD) [Beck and Teboulle, 2003] stands out as a prominent and general regret minimization algorithm. It has promising theoretical results but remains less competitive than CFRs when directly applied to solving IIGs. Recently, researchers have tried to build connections between CFR and OMD [Farina *et al.*, 2021; Liu *et al.*, 2022]. The study by [Farina *et al.*, 2021] demonstrates that RM+ is essentially a specialized form of OMD when employed to minimize counterfactual regret at each decision node. This connection inspires an optimistic variant of RM+, named Predictive RM+ (PRM+), seeking to benefit from the predictability of slowly-changing counterfactual losses over time. Its extension Predictive CFR+ (PCFR+) exhibits extremely fast convergence on non-poker IIGs.

Despite the notable success of PCFR+, a limitation arises from its practice of assigning uniform weights to each iteration when determining regrets. The algorithm becomes

particularly challenging when dealing with dominated actions, leading to high regrets for other actions and requiring a substantial number of iterations to mitigate this negative impact[1]. Assigning more weights to recent iterations has shown to be important for fast convergence [Tammelin, 2014; Brown and Sandholm, 2019a; Grand-Clément and Kroer, 2023]. Notably, Discounted CFR (DCFR) [Brown and Sandholm, 2019a] is a family of algorithms that discounts prior iterations when determining both regrets and average strategy. Thus, it is desirable to design a CFR variant that not only allocates more weights to recent iteration, thereby alleviating high regrets from earlier iterations, but also leverages predictions to accelerate convergence.

To this end, we delve into minimizing weighted counterfactual regret with OMD and optimistic variant:

- We demonstrate that by directly employing OMD to minimize weighted counterfactual regret, we obtain a CFR variant, DCFR+, which incorporates regret discounting similar to DCFR and clips negative regrets akin to CFR+. Remarkably, the CFR variant previously discovered through evolutionary search [Hang *et al.*, 2022] is a special case of DCFR+, showcasing faster convergence than DCFR.

- Furthermore, by applying optimistic OMD, we derive a novel CFR variant, PDCFR+, which leverages predicted regrets to compute new strategy like PCFR+, while updating regrets similar to DCFR+. It swiftly mitigates negative effects of dominated actions and consistently leverages predictions to speed up convergence.

- We observe that the updating rule of OMD with increasing weighted loss is equivalent to OMD with an increasing learning rate or a decreasing regularization term. This observation offers a partial explanation for the superior performance of DCFR.

- We theoretically prove that DCFR+ and PDCFR+ converge to a NE under distinct weighting schemes for regrets and average strategies, requiring the weighting scheme for regrets to be more aggressive than that for average strategies.

- Finally, we conduct extensive experimental evaluations of CFR variants on commonly used IIGs in the research community. We find that PDCFR+ outperforms other CFR variants by 4-8 orders of magnitude on non-poker games and small poker games, while DCFR+ converges fastest on large-scale poker games.

2 Preliminaries

In this section, we present the notations for formulating IIGs, introduce typical regret minimization algorithms, and discuss the vanilla CFR algorithm along with its variants.

2.1 Sequential Decision Process

Sequential decision process [Farina *et al.*, 2019a] is a tree-based formalism for describing each player's decision-making in IIGs, consisting of two types of nodes: decision nodes and observation nodes. At each **decision node**

---

[1]We provide a motivating example in section 3.

$j \in \mathcal{J}$, the player selects an **action** based on a **local strategy** $\boldsymbol{x}_j \in \Delta^{n_j}$, where Δ^{n_j} is a simplex over the action set $\mathcal{A}_j \subseteq \mathcal{A}$ of size n_j. After taking the action a, the player proceeds to an observation node $k = \rho(j, a)$. At each **observation node** $k \in \mathcal{K}$, the player receives a **signal** $s \in S_k$ and then reaches another decision node $j' = \rho(k, s)$. The set $C_{j,a} = \left\{ \rho(\rho(j, a), s) : s \in S_{\rho(j,a)} \right\}$ denotes all decision nodes that are earliest reachable after taking action a at j. We introduce a dummy root decision node o with one action to ensure a unique root. An illustration is given in Appendix A.

For a local strategy set $\{\boldsymbol{x}_j\}_{j \in \mathcal{J}}$, we can construct a **sequence-form strategy** $\dot{\boldsymbol{x}}$ represented as a vector indexed over $\{(j, a) : j \in \mathcal{J}, a \in \mathcal{A}_j\}$. Each entry in the vector corresponds to a pair (j, a), with the value representing the product of probabilities of all actions along the path from the root to (j, a). The sequence-form strategy space is denoted by $\mathcal{X}$.

In a 2p0s IIG, where player 1 and player 2 have sequence-from strategy space $\mathcal{X}$ and $\mathcal{Y}$, and the set of decision nodes are $\mathcal{J}_x$ and $\mathcal{J}_y$, the problem of finding a **NE** can be formulated as a bilinear saddle point problem

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

where $\boldsymbol{A}$ is a sparse payoff matrix encoding the losses for player 1. For a strategy profile $(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}})$, let $\delta_1(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}})$ denote the incentive for player 1 to unilaterally choose another strategy: $\delta_1(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}}) = \dot{\boldsymbol{x}}^\top \boldsymbol{A} \dot{\boldsymbol{y}} - \min_{\dot{\boldsymbol{x}}' \in \mathcal{X}} \dot{\boldsymbol{x}}'^\top \boldsymbol{A} \dot{\boldsymbol{y}}$. Similarly, $\delta_2(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}}) = \max_{\dot{\boldsymbol{y}}' \in \mathcal{Y}} \dot{\boldsymbol{x}}^\top \boldsymbol{A} \dot{\boldsymbol{y}}' - \dot{\boldsymbol{x}}^\top \boldsymbol{A} \dot{\boldsymbol{y}}$. An ϵ-**NE** satisfies $\delta_i(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}}) \leq \epsilon, \forall i \in \{1, 2\}$. The **exploitability** of $(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}})$ measures its distance from equilibrium and is defined by $e(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}}) = \sum_{i \in \{1,2\}} \delta_i(\dot{\boldsymbol{x}}, \dot{\boldsymbol{y}})/2$.

2.2 Regret Minimization

A **regret minimization** algorithm [Zinkevich, 2003] repeatedly plays against an unknown environment and starts with a decision $\boldsymbol{x}^1 \in \mathcal{D} \subseteq \mathbb{R}^n$, where $\mathcal{D}$ is the decision space. In each iteration t, it observes a loss $\boldsymbol{\ell}^t \in \mathbb{R}^n$ from the environment and computes its next decision $\boldsymbol{x}^{t+1}$ based on past decisions $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^t$ and previous losses $\boldsymbol{\ell}^1, \ldots, \boldsymbol{\ell}^t$. The objective of the algorithm is to minimize the **total regret**

$$R^T = \max_{\boldsymbol{x}' \in \mathcal{D}} \sum_{t=1}^{T} \left\langle \boldsymbol{\ell}^t, \boldsymbol{x}^t - \boldsymbol{x}' \right\rangle.$$

Online Mirror Descent (OMD) [Beck and Teboulle, 2003] stands out as a famous regret minimization algorithm compatible with arbitrary decision spaces. It updates the decision according to $\boldsymbol{x}^1 = \operatorname{argmin}_{\boldsymbol{x}' \in \mathcal{D}} \psi(\boldsymbol{x}')$, and

$$\boldsymbol{x}^{t+1} = \operatorname*{argmin}_{\boldsymbol{x}' \in \mathcal{D}} \left\{ \left\langle \boldsymbol{\ell}^t, \boldsymbol{x}' \right\rangle + \frac{1}{\eta} \mathcal{B}_\psi(\boldsymbol{x}' \| \boldsymbol{x}^t) \right\},$$

where $\psi : \mathcal{D} \to \mathbb{R}$ is a regularizer, $\mathcal{B}_\psi(\boldsymbol{x}' \| \boldsymbol{x}) = \psi(\boldsymbol{x}') - \psi(\boldsymbol{x}) - \langle \nabla \psi(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x} \rangle$ is the Bregman divergence associated with ψ, and $\eta > 0$ is an arbitrary step size. Optimistic OMD [Syrgkanis *et al.*, 2015] seeks to benefit from the predictability of slowly-changing loss $\boldsymbol{\ell}^t$. In iteration t, it predicts the next iteration's loss $\boldsymbol{\ell}^{t+1}$ as $\boldsymbol{m}^{t+1}$ and updates the

Algorithms	Cumulative Regret R_j^t	New Strategy x_j^{t+1}	Cumulative Strategy X^t
CFR	$R_j^{t-1} + r_j^t$	$[R_j^t]^+ / \left\| [R_j^t]^+ \right\|_1$	$X^{t-1} + \dot{x}^t$
CFR+	$[R_j^{t-1} + r_j^t]^+$	$R_j^t / \|R_j^t\|_1$	$X^{t-1} + t * \dot{x}^t$
Linear CFR	$R_j^{t-1} + t * r_j^t$	$[R_j^t]^+ / \left\| [R_j^t]^+ \right\|_1$	$X^{t-1} + t * \dot{x}^t$
DCFR	$R_j^t = R_j^{t-1} \odot d_j^{t-1} + r_j^t$, where $d_j^t[a] = \begin{cases} \frac{t^\alpha}{t^\alpha+1} & \text{if } R_j^t[a] > 0 \\ \frac{t^\beta}{t^\beta+1} & \text{otherwise} \end{cases}$	$[R_j^t]^+ / \left\| [R_j^t]^+ \right\|_1$	$X^{t-1}(\frac{t-1}{t})^\gamma + \dot{x}^t$
DCFR+	$\left[R_j^{t-1} \frac{(t-1)^\alpha}{(t-1)^\alpha+1} + r_j^t \right]^+$	$R_j^t / \|R_j^t\|_1$	$X^{t-1}(\frac{t-1}{t})^\gamma + \dot{x}^t$
PCFR+	$[R_j^{t-1} + r_j^t]^+$	$\tilde{R}_j^{t+1} / \left\| \tilde{R}_j^{t+1} \right\|_1$, where $\tilde{R}_j^{t+1} = [R_j^t + v_j^{t+1}]^+$	$X^{t-1} + t^2 \dot{x}^t$
PDCFR+	$\left[R_j^{t-1} \frac{(t-1)^\alpha}{(t-1)^\alpha+1} + r_j^t \right]^+$	$\tilde{R}_j^{t+1} / \left\| \tilde{R}_j^{t+1} \right\|_1$, where $\tilde{R}_j^{t+1} = \left[R_j^t \frac{t^\alpha}{t^\alpha+1} + v_j^{t+1} \right]^+$	$X^{t-1}(\frac{t-1}{t})^\gamma + \dot{x}^t$

Table 1: Comparison of existing CFR variants with our proposed **DCFR+** and **PDCFR+** (α, β, γ are hyperparameters).

decision according to $x^1 = z^0 = \operatorname{argmin}_{x' \in \mathcal{D}} \psi(x')$, and

$$z^t = \underset{z' \in \mathcal{D}}{\operatorname{argmin}} \ \langle \ell^t, z' \rangle + \frac{1}{\eta} \mathcal{B}_\psi(z' \,\|\, z^{t-1}),$$

$$x^{t+1} = \underset{x' \in \mathcal{D}}{\operatorname{argmin}} \ \left\{ \langle m^{t+1}, x' \rangle + \frac{1}{\eta} \mathcal{B}_\psi(x' \,\|\, z^t) \right\}.$$

Regret Matching (RM) [Hart and Mas-Colell, 2000] and Regret Matching+ (RM+) [Tammelin, 2014] are two regret minimization algorithms operating on the simplex Δ^n. RM typically starts with a uniform random strategy x^1. On each iteration t, RM calculates the **instantaneous regret** $r^t = \langle x^t, \ell^t \rangle \mathbf{1} - \ell^t$ and then accumulates it to obtain the **cumulative regret** $R^t = \sum_{k=1}^t r^k$. The next strategy is proportional to the positive cumulative regret, *i.e.*, $x^{t+1} = [R^t]^+ / \left\| [R^t]^+ \right\|_1$, where $[\cdot]^+ = \max\{\cdot, 0\}$. For convenience, we define $\mathbf{0}/0$ as the uniform distribution. RM+ is a straightforward variant of RM. It sets any action with negative cumulative regret to zero in each iteration so that it promptly reuses an action showing promise of performing well. Formally, $R^t = [R^{t-1} + r^t]^+$ and $x^{t+1} = R^t / \|R^t\|_1$.

2.3 Counterfactual Regret Minimization

Counterfactual Regret Minimization (CFR) [Zinkevich *et al.*, 2007] is one of the most popular equilibrium-finding algorithms for solving IIGs. It operates as a regret minimization algorithm within the sequence-form strategy space. Given a sequence-form strategy $\dot{x}^t \in \mathcal{X}$ and a loss $\dot{\ell}^t = A\dot{y}^t$, the key concept involves constructing a **counterfactual loss**

$\ell_j^t \in \mathbb{R}^{n_j}$ for each decision node $j \in \mathcal{J}$:

$$\ell_j^t[a] = \dot{\ell}^t[j, a] + \sum_{j' \in C_{j,a}} \langle \ell_{j'}^t, x_{j'}^t \rangle.$$

CFR then employs RM to minimize the total counterfactual regret $R_j^T = \max_{x_j' \in \Delta^{n_j}} \sum_{t=1}^T \langle \ell_j^t, x_j^t - x_j' \rangle$ at each decision node with respect to counterfactual loss. It is guaranteed that the total regret R^T is bounded by the sum of the total counterfactual regrets under each decision node, *i.e.*, $R^T \leq \sum_{j \in \mathcal{J}} [R_j^T]^+$.

The computational procedure of CFR on each iteration t is summarized as follows: (1) decomposes the sequence-form strategy $\dot{x}^t$ into local strategies x_j^t for each decision node $j \in \mathcal{J}$; (2) recursively traverses the game tree to calculate the counterfactual loss ℓ_j^t; (3) accumulates the **instantaneous counterfactual regret** $r_j^t = \langle x_j^t, \ell_j^t \rangle \mathbf{1} - \ell_j^t$ to obtain the **cumulative counterfactual regret** $R_j^t = R_j^{t-1} + r_j^t$; (4) computes the new local strategy $x_j^{t+1} = [R_j^t]^+ / \left\| [R_j^t]^+ \right\|_1$ for the next iteration; (5) constructs $\dot{x}^{t+1}$ based on local strategies; (6) computes the **average strategy** $\bar{x}^t = X^t / \|X^t\|_1$, where $X^t = X^{t-1} + \dot{x}^t$ is the **cumulative strategy**. If both players employ CFR, the average strategy profile $\langle \bar{x}^T, \bar{y}^T \rangle$ converges to a $\Delta(|\mathcal{J}_x| + |\mathcal{J}_y|)\sqrt{|\mathcal{A}|}/\sqrt{T}$-NE after T iterations, where Δ is the range of losses.

2.4 CFR Variants

Since the birth of CFR, researchers have proposed many novel CFR variants, greatly improving the convergence rate

of the vanilla CFR. CFR+ [Tammelin, 2014; Bowling *et al.*, 2015] incorporates three small yet effective modifications, resulting in an order of magnitude faster convergence compared to CFR. (1) CFR+ employs RM+ instead of RM, *i.e.*, $\boldsymbol{R}_j^t = \left[\boldsymbol{R}_j^{t-1} + \boldsymbol{r}_j^t\right]^+$. (2) CFR+ adopts a linearly weighted average strategy where iteration t is weighted by t, *i.e.*, $\boldsymbol{X}^t = \boldsymbol{X}^{t-1} + t\dot{\boldsymbol{x}}^t$. (3) CFR+ uses the alternating-updates technique.

DCFR [Brown and Sandholm, 2019a] is a family of algorithms which discounts prior iterations' cumulative regrets and dramatically accelerates convergence especially in games where some actions are very costly mistakes. Specifically,

$$\boldsymbol{R}_j^t = \boldsymbol{R}_j^{t-1} \odot \boldsymbol{d}_j^{t-1} + \boldsymbol{r}_j^t, \text{where } \boldsymbol{d}_j^t[a] = \begin{cases} \frac{t^\alpha}{t^\alpha + 1} & \text{if } \boldsymbol{R}_j^t[a] > 0 \\ \frac{t^\beta}{t^\beta + 1} & \text{otherwise} \end{cases}$$

$$\boldsymbol{X}^t = \boldsymbol{X}^{t-1} \left(\frac{t-1}{t}\right)^\gamma + \dot{\boldsymbol{x}}^t.$$

Linear CFR [Brown and Sandholm, 2019a] is a special case of DCFR where iteration t's contribution to cumulative regrets and cumulative strategy is proportional to t.

2.5 The Predictive CFR Variant

Recently, [Farina *et al.*, 2021] demonstrates that a general regret minimization algorithm can be adapted for constructing a regret minimization algorithm tailored to a simplex. To illustrate, let us focus on the objective of minimizing the total counterfactual regret R_j^T at a decision node j. Additionally, we have a general regret minimization algorithm $\mathbb{A}$, which we apply to a decision space $\mathbb{R}_{\geq 0}^{n_j}$. At each iteration t, upon obtaining the counterfactual loss $\boldsymbol{\ell}_j^t$, we initially compute a modified loss $\tilde{\boldsymbol{\ell}}_j^t = -\boldsymbol{r}_j^t = \boldsymbol{\ell}_j^t - \langle \boldsymbol{\ell}_j^t, \boldsymbol{x}_j^t \rangle \mathbf{1}$. It is then fed into $\mathbb{A}$, and we receive its next decision $\tilde{\boldsymbol{x}}_j^{t+1} \in \mathbb{R}_{\geq 0}^{n_j}$. Finally, we obtain the next strategy as $\boldsymbol{x}_j^{t+1} = \tilde{\boldsymbol{x}}_j^{t+1} / \left\| \tilde{\boldsymbol{x}}_j^{t+1} \right\|_1$.

Based on the construction, [Farina *et al.*, 2021] illustrates that employing OMD with $\psi = \frac{1}{2} \|\cdot\|_2^2$ as the algorithm $\mathbb{A}$ yields the same strategy as RM+ in each iteration. Additionally, by using Optimistic OMD with $\psi = \frac{1}{2} \|\cdot\|_2^2$ and $\boldsymbol{m}^{t+1} = -\boldsymbol{v}_j^{t+1}$, where $\boldsymbol{v}_j^{t+1} = \langle \boldsymbol{m}_j^{t+1}, \boldsymbol{x}_j^t \rangle \mathbf{1} - \boldsymbol{m}_j^{t+1}$ is the prediction of $\boldsymbol{r}_j^{t+1}$ and $\boldsymbol{m}_j^{t+1} = \boldsymbol{\ell}_j^t$ is the prediction of $\boldsymbol{\ell}_j^{t+1}$, Predictive RM+ (PRM+) is obtained, *i.e.*,

$$\boldsymbol{R}_j^t = [\boldsymbol{R}_j^{t-1} + \boldsymbol{r}_j^t]^+, \tilde{\boldsymbol{R}}_j^{t+1} = [\boldsymbol{R}_j^t + \boldsymbol{v}_j^{t+1}]^+, \boldsymbol{x}_j^{t+1} = \tilde{\boldsymbol{R}}_j^{t+1} / \left\| \tilde{\boldsymbol{R}}_j^{t+1} \right\|_1.$$

Moreover, [Farina *et al.*, 2021] introduces Predictive CFR+ (PCFR+), which uses PRM+ as the regret minimization algorithm in each decision node and incorporates a quadratic weighted average strategy, defined as $\boldsymbol{X}^t = \boldsymbol{X}^{t-1} + t^2 \dot{\boldsymbol{x}}^t$.

We summarized CFR and typical variants in Table 1.

3 Performance Decline of PCFR+ in the Presence of Dominated Actions

Although PCFR+ successfully employs prediction to significantly accelerate convergence, we show that its performance deteriorates notably when facing dominated actions. Consider a 2×2 zero-sum normal-form game *NFG (2)*, represented by $\max_{\boldsymbol{x} \in \Delta^2} \min_{\boldsymbol{y} \in \Delta^2} \boldsymbol{x}^T \boldsymbol{U} \boldsymbol{y}$, where $\boldsymbol{U} = ((1, 0), (0, 2))$ is
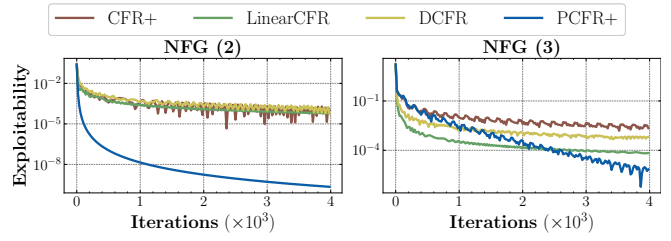

Figure 1: Convergence results of four CFR variants on two games.

player 1's utility matrix. As shown in the left plot of Figure 1, PCFR+ converges extremely faster than other CFR variants, benefiting from prediction during strategy updates. However, when we introduce a dominated action for each player, resulting in *NFG (3)*, where $\boldsymbol{U} = ((1, 0, 5), (0, 2, 0), (0, 0, 100))$, PCFR+'s performance declines greatly as depicted in the right plot. For player 2, action 3 is dominated by action 1, allowing its elimination. Thus, player 1's action 3 can also be eliminated. Consequently, *NFG (3)* reduces to *NFG (2)*, with the NE $\boldsymbol{x} = \boldsymbol{y} = \left(\frac{2}{3}, \frac{1}{3}, 0\right)$. After the first iteration of PCFR+, player 2 has cumulative regrets $\boldsymbol{R}^1 = \left(\frac{100}{3}, \frac{100}{3}, 0\right)$. The substantial loss of action 3 leads to high cumulative regrets for player 2's actions 1 and 2, making it challenging to distinguish between these two actions. Due to the small instantaneous regrets per iteration, PCFR+ requires about 1500 iterations to achieve a 2:1 cumulative regrets for actions 1 and 2, producing an approximate NE.

While LinearCFR and DCFR initially converge faster than PCFR+ in *NFG (3)* by assigning more weights to recent iterations and alleviating high regrets from earlier iterations, PCFR+ leverages predictions to surpass them after 2000 iterations. It will be a best-of-both-worlds algorithm if we design a CFR variant that not only allocates more weights to recent iterations but also employs predictions to compute a new strategy, therefore greatly speeding up convergence.

4 Minimizing Weighted Counterfactual Regret with OMD and Optimistic Variant

In various situations, it makes sense to consider recent iterations as more crucial than earlier ones. Take equilibrium-finding algorithms as an example, earlier iterations are prone to selecting incorrect actions, resulting in substantial losses and causing significant regrets [Brown and Sandholm, 2019a]. Additionally, it is typical for recent iterations to generate strategies that are closer to a NE [Perolat *et al.*, 2021]. Therefore, assigning more weights to recent iterations is a natural choice, potentially leading to fast convergence [Abernethy *et al.*, 2018; Wang and Abernethy, 2018].

Moreover, it is a prevalent practice to permit distinct weighting sequences for regrets and average strategies to facilitate fast convergence. For instance, CFR+ adopts a linearly weighted average strategy while employing a uniform weighting sequence for regrets. Besides, DCFR takes this approach a step further by using distinct weighting sequences for both regrets and average strategy.

In a weighted regret minimization algorithm applied to the sequence-form strategy space, the loss $\tilde{\boldsymbol{\ell}}^t$ incurred at iteration

Algorithm 1: Construction of a weighed CFR variant using a general regret minimization algorithm from player 1's perspective.

Input: game G, total iterations T, general regret minimization algorithm $\mathbb{A}$, weighting sequences $\boldsymbol{w}$ and $\boldsymbol{\tau}$.

1 **for** $j \in \mathcal{J}$ **do**
2     $\mathbb{A}_j \leftarrow$ instantiate $\mathbb{A}$ with $\mathcal{D} = \mathbb{R}_{\geq 0}^{n_j}$;
3     $\boldsymbol{x}_j^1 \leftarrow \mathbb{A}_j.\text{first\_decision}()$;
4 construct $\dot{\boldsymbol{x}}^1$ by $\{\boldsymbol{x}_j^1 \mid j \in \mathcal{J}\}$;
5 **for** $t = 1 \rightarrow T$ **do**
6   decompose $\dot{\boldsymbol{x}}^t$ into $\{\boldsymbol{x}_j^t \mid j \in \mathcal{J}\}$;
7   calculate the counterfactual loss $\{\boldsymbol{\ell}_j^t \mid j \in \mathcal{J}\}$
      based on the loss $\dot{\boldsymbol{\ell}}^t = A\dot{\boldsymbol{y}}^t$;
8   **for** $j \in \mathcal{J}$ **do**
9        $\boldsymbol{r}_j^t \leftarrow \langle \boldsymbol{\ell}_j^t, \boldsymbol{x}_j^t \rangle \mathbf{1} - \boldsymbol{\ell}_j^t$;
10        $\tilde{\boldsymbol{\ell}}_j^t \leftarrow -w^t \boldsymbol{r}_j^t$;
11        $\mathbb{A}_j.\text{observe\_loss}(\tilde{\boldsymbol{\ell}}_j^t)$;
12        $\tilde{\boldsymbol{x}}_j^{t+1} \leftarrow \mathbb{A}_j.\text{next\_decision}()$;
13        $\boldsymbol{x}_j^{t+1} \leftarrow \tilde{\boldsymbol{x}}_j^{t+1} / \|\tilde{\boldsymbol{x}}_j^{t+1}\|_1$;
14   construct $\dot{\boldsymbol{x}}^{t+1}$ by $\{\boldsymbol{x}_j^{t+1} \mid j \in \mathcal{J}\}$;
15     $\boldsymbol{X}^t \leftarrow \boldsymbol{X}^{t-1} + \tau^t \dot{\boldsymbol{x}}^t$;
16     $\bar{\boldsymbol{x}}^t \leftarrow \boldsymbol{X}^t / \|\boldsymbol{X}^t\|_1$;

Output: The final average strategy $\bar{\boldsymbol{x}}^T$.

t is scaled by the weight w_t, forming a sequence of weights denoted as $\boldsymbol{w}$. The total weighted regret by the sequence $\boldsymbol{w}$ is

$$R_{\boldsymbol{w}}^T = \max_{\dot{\boldsymbol{x}}' \in \mathcal{X}} \sum_{t=1}^{T} \left\langle w^t \dot{\boldsymbol{\ell}}^t, \dot{\boldsymbol{x}}^t - \dot{\boldsymbol{x}}' \right\rangle.$$

Moreover, define $\bar{\boldsymbol{x}}_{\boldsymbol{\tau}}^t$ to be the weighted average strategy scaled by the weighting sequence $\boldsymbol{\tau}$:

$$\bar{\boldsymbol{x}}_{\boldsymbol{\tau}}^t = \boldsymbol{X}_{\boldsymbol{\tau}}^t / \|\boldsymbol{X}_{\boldsymbol{\tau}}^t\|_1, \; \boldsymbol{X}_{\boldsymbol{\tau}}^t = \boldsymbol{X}_{\boldsymbol{\tau}}^{t-1} + \tau^t \dot{\boldsymbol{x}}^t.$$

To minimize the total weighted regret $R_{\boldsymbol{w}}^T$, we first decompose it into the sum of the total weighted counterfactual regrets $R_{j,\boldsymbol{w}}^T$ under each decision node by constructing the counterfactual loss $\boldsymbol{\ell}_j^t$ at each iteration t, where

$$R_{j,\boldsymbol{w}}^T = \max_{\boldsymbol{x}_j' \in \Delta^{n_j}} \sum_{t=1}^{T} \left\langle w^t \boldsymbol{\ell}_j^t, \boldsymbol{x}_j^t - \boldsymbol{x}_j' \right\rangle,$$

so that we have the guarantee $R_{\boldsymbol{w}}^T \leq \sum_{j \in \mathcal{J}} [R_{j,\boldsymbol{w}}^T]^+$ and transform the objective to minimize the total weighted counterfactual regret $R_{j,\boldsymbol{w}}^T$ at each decision node.

To develop an algorithm for minimizing weighted counterfactual regret by leveraging a general regret minimization algorithm, we propose a construction method similar to the one described in [Farina *et al.*, 2021], but with a modified loss $\tilde{\boldsymbol{\ell}}_j^t = -w^t \boldsymbol{r}_j^t$. A concise summary of this approach is presented in Algorithm 1.

As demonstrated in Theorem 1 (all proofs of the theorems are in Appendix C), when employing OMD with $\psi = \frac{1}{2} \|\cdot\|_2^2$ as the algorithm $\mathbb{A}$, it simplifies to the WCFR+ algorithm:

$$\boldsymbol{R}_j^t = [\boldsymbol{R}_j^{t-1} + w^t \boldsymbol{r}_j^t]^+, \boldsymbol{x}_j^{t+1} = \|\boldsymbol{R}_j^t\| / \|\boldsymbol{R}_j^t\|_1, \boldsymbol{X}^t = \boldsymbol{X}^{t-1} + \tau^t \dot{\boldsymbol{x}}^t.$$

Similarly, when employing optimistic OMD with $\psi = \frac{1}{2} \|\cdot\|_2^2$, it reduces to the PWCFR+ algorithm:

$$\boldsymbol{R}_j^t = [\boldsymbol{R}_j^{t-1} + w^t \boldsymbol{r}_j^t]^+, \tilde{\boldsymbol{R}}_j^{t+1} = [\boldsymbol{R}_j^t + w^{t+1} \boldsymbol{v}_j^{t+1}]^+,$$

$$\boldsymbol{x}_j^{t+1} = \|\tilde{\boldsymbol{R}}_j^{t+1}\| / \|\tilde{\boldsymbol{R}}_j^{t+1}\|_1, \boldsymbol{X}^t = \boldsymbol{X}^{t-1} + \tau^t \dot{\boldsymbol{x}}^t$$

Theorem 1. *For all $\eta > 0$, when employing OMD and optimistic OMD with $\psi = \frac{1}{2} \|\cdot\|_2^2$ as the algorithm $\mathbb{A}$, they reduce to WCFR+ and PWCFR+, respectively.*

As illustrated in Theorem 2, when both players employ a weighted regret minimization algorithm in a 2p0s IIG, the weighted average strategy profile converges to a NE.

Theorem 2. *Assuming both players employ a weighted regret minimization algorithm with the weighting sequence $\boldsymbol{\tau}$ for the average strategy in a 2p0s IIG, and given that the two players have total weighted regrets $R_{\boldsymbol{\tau},x}^T$ and $R_{\boldsymbol{\tau},y}^T$ respectively, the weighted average strategy profile $(\bar{\boldsymbol{x}}_{\boldsymbol{\tau}}^T, \bar{\boldsymbol{y}}_{\boldsymbol{\tau}}^T)$ after T iterations forms a $\frac{R_{\boldsymbol{\tau},x}^T + R_{\boldsymbol{\tau},y}^T}{\sum_{t=1}^{T} \tau^t}$-NE.*

Although WCFR+ and PWCFR+ algorithms aim to minimize the total weight regret $R_{\boldsymbol{w}}^T$, we show that they also minimize the total weighted regret $R_{\boldsymbol{\tau}}^T$ as long as the weighting sequence $\boldsymbol{w}$ is more aggressive than $\boldsymbol{\tau}$.

Theorem 3. *Assuming both players employ the WCFR+ algorithm with the weighting sequence $\boldsymbol{w}$ for loss and $\boldsymbol{\tau}$ for the average strategy in a 2p0s IIG, and $\{\frac{\tau_t}{w_t}\}_{t \leq T}$ is a positive non-increasing sequence, the weighted average strategy profile $(\bar{\boldsymbol{x}}_{\boldsymbol{\tau}}^T, \bar{\boldsymbol{y}}_{\boldsymbol{\tau}}^T)$ after T iterations forms a $\left(\sum_{j \in \mathcal{J}_x \cup \mathcal{J}_y} \sqrt{\frac{\tau^1}{w^1} \sum_{t=1}^{T} \tau^t w^t \|\boldsymbol{r}_j^t\|_2^2} \right) / \sum_{t=1}^{T} \tau^t$-NE.*

Theorem 4. *Assuming both players employ the PWCFR+ algorithm with the weighting sequence $\boldsymbol{w}$ for loss and $\boldsymbol{\tau}$ for the average strategy in a 2p0s IIG, and $\{\frac{\tau_t}{w_t}\}_{t \leq T}$ is a positive non-increasing sequence, the weighted average strategy profile $(\bar{\boldsymbol{x}}_{\boldsymbol{\tau}}^T, \bar{\boldsymbol{y}}_{\boldsymbol{\tau}}^T)$ after T iterations forms a $\left(\sum_{j \in \mathcal{J}_x \cup \mathcal{J}_y} \sqrt{2 \frac{\tau^1}{w^1} \sum_{t=1}^{T} \tau^t w^t \|\boldsymbol{r}_j^t - \boldsymbol{v}_j^t\|_2^2} \right) / \sum_{t=1}^{T} \tau^t$-NE.*

A question remains regarding the selection of suitable weighting sequences in practical applications. Inspired by DCFR [Brown and Sandholm, 2019a] and to avoid potential numerical issues, we adopt a similar less-aggressive discounting sequence, resulting in a specific algorithm DCFR+:

$$\boldsymbol{R}_j^t = \left[\boldsymbol{R}_j^{t-1} \frac{(t-1)^\alpha}{(t-1)^\alpha + 1} + \boldsymbol{r}_j^t \right]^+, \boldsymbol{x}_j^{t+1} = \boldsymbol{R}_j^t / \|\boldsymbol{R}_j^t\|_1,$$

$$\boldsymbol{X}^t = \boldsymbol{X}^{t-1} \left(\frac{t-1}{t} \right)^\gamma + \dot{\boldsymbol{x}}^t.$$

DCFR+ integrates features from both CFR+ and DCFR in a principled manner. It discounts cumulative regrets, assigning

more weights to recent instantaneous regrets. Additionally, it clips the negative part of cumulative regrets, allowing for quick reuse of promising actions. Notably, the CFR variant discovered through evolutionary search [Hang *et al.*, 2022] emerges as a special case of DCFR+ where $\alpha = 1.5$ and $\gamma = 4$, showcasing faster convergence than DCFR.

We adopt a similar weighting sequence for PWCFR+, resulting in the PDCFR+ algorithm. PDCFR+ utilizes predicted cumulative regrets to compute new strategy akin to PCFR+, while updating cumulative regrets similar to DCFR+:

$$\boldsymbol{R}_j^t = \left[\boldsymbol{R}_j^{t-1}\frac{(t-1)^{\alpha}}{(t-1)^{\alpha}+1} + \boldsymbol{r}_j^t\right]^+, \tilde{\boldsymbol{R}}_j^{t+1} = \left[\boldsymbol{R}_j^t\frac{t^{\alpha}}{t^{\alpha}+1} + \boldsymbol{v}_j^{t+1}\right]^+,$$

$$\boldsymbol{x}_j^{t+1} = \tilde{\boldsymbol{R}}_j^{t+1}/\left\|\tilde{\boldsymbol{R}}_j^{t+1}\right\|_1, \boldsymbol{X}^t = \boldsymbol{X}^{t-1}\left(\frac{t-1}{t}\right)^{\gamma} + \dot{\boldsymbol{x}}^t.$$

Interestingly, our connection between OMD and WCFR+ also provides a new perspective to understand the increasing weight w^t in DCFR and DCFR+. When employing OMD with $\psi = \frac{1}{2}\left\|\cdot\right\|_2^2$ as the algorithm $\mathbb{A}$, it updates the decision $\tilde{\boldsymbol{x}}_j^{t+1}$, corresponding to the cumulative regrets $\boldsymbol{R}_j^t$ in WCFR+:

$$\tilde{\boldsymbol{x}}_j^{t+1} = \underset{\tilde{\boldsymbol{x}}_j' \in \mathbb{R}_{\geq 0}^{n_j}}{\operatorname{argmin}} \left\{\left\langle -w^t\boldsymbol{r}_j^t, \tilde{\boldsymbol{x}}_j'\right\rangle + \frac{1}{2\eta}\left\|\tilde{\boldsymbol{x}}_j' - \tilde{\boldsymbol{x}}_j^t\right\|_2^2\right\}.$$

It can be equivalently written as:

$$\tilde{\boldsymbol{x}}_j^{t+1} = \underset{\tilde{\boldsymbol{x}}_j' \in \mathbb{R}_{\geq 0}^{n_j}}{\operatorname{argmin}} \left\{\left\langle -\boldsymbol{r}_j^t, \tilde{\boldsymbol{x}}_j'\right\rangle + \frac{1}{2\eta w^t}\left\|\tilde{\boldsymbol{x}}_j' - \tilde{\boldsymbol{x}}_j^t\right\|_2^2\right\}$$

$$= \underset{\tilde{\boldsymbol{x}}_j' \in \mathbb{R}_{\geq 0}^{n_j}}{\operatorname{argmin}} \left\{\left\langle -\boldsymbol{r}_j^t, \tilde{\boldsymbol{x}}_j'\right\rangle + \frac{1}{2\eta}\frac{1}{w^t}\left\|\tilde{\boldsymbol{x}}_j' - \tilde{\boldsymbol{x}}_j^t\right\|_2^2\right\}.$$

We can interpret the increasing weight w^t in two ways: an increasing learning rate $\eta^t = \eta w^t$ or a decreasing regularization term $\frac{1}{w_t}$. As a result, during the early iterations, cumulative regrets $\boldsymbol{R}_j^t$, *i.e.*, $\tilde{\boldsymbol{x}}_j^{t+1}$, are learned at a gradual pace or regularized by a substantial strength. This interpretation aligns with our intuition, as we aim to control cumulative regrets from growing excessively large in the initial stages.

5 Experiments

In this section, we briefly describe testing games, compare PDCFR+ with other CFR variants, and analyze PDCFR+'s superior performance on normal-form games.

5.1 Testing Games

We use several commonly used IIGs in the research community and provide brief descriptions below. For more details, please refer to the Appendix B. ***Kuhn Poker*** [Kuhn, 1950] is a simplified poker with a three-card deck and one chance to bet for each player. ***Leduc Poker*** [Southey *et al.*, 2005] is a larger game with a 6-card deck and two betting rounds. In ***Liar's Dice (x)*** (x=4, 5) [Lisý *et al.*, 2015], each player gets an x-sided dice, which they roll at the start and take turn placing bets on the outcome. ***Goofspiel (x)*** (x=4, 5) [Ross, 1971] is a card game where each player has x cards and aims to score points by bidding simultaneously in x rounds. ***GoofspielImp***

(x) (x=4, 5) is a imperfect variant of *Goofspiel (x)*, where bid cards remain unrevealed. In ***Battleship (x)*** (x=2, 3) [Farina *et al.*, 2019b], players secretly position a 1×2 ship on separate $2 \times x$ grids and take turns firing at the opponent's ship three times. ***HUNL Subgame (x)*** (x=3, 4) is a heads-up no-limit Texas hold'em (HUNL) subgame generated by the top poker agent Libratus [Brown and Sandholm, 2018].

The testing games selected exhibit diverse structures and considerable game sizes, presenting a nontrivial challenge to solve. However, their complexity remains manageable for an algorithm to complete within two weeks, rendering them highly suitable for evaluating algorithmic performance. When extending to larger games, we can capitalize on a variety of well-established techniques, such as abstraction, decomposition [Burch *et al.*, 2014], and subgame-solving [Brown and Sandholm, 2017; Brown *et al.*, 2018]. These approaches effectively mitigate the increased complexity associated with scaling up the size of the games.

5.2 Convergence Results

We compare PDCFR+ with CFR+, LinearCFR, DCFR, PCFR+ and DCFR+ across twelve games. For DCFR, we adopt the hyperparameters suggested by the authors, specifically $\alpha = 1.5$, $\beta = 0$, and $\gamma = 2$. Regarding DCFR+ and PDCFR+, according to Theorems 3 and 4, α and γ must ensure that $\left\{\frac{\tau_t}{w_t}\right\}_{t \leq T}$ forms a positive, non-increasing sequence. However, in practice, we find that the range of values for α and γ can be larger, and they generally yield good convergence results. Based on these observations, for DCFR+, we set $\alpha = 1.5$ and $\gamma = 4$ following [Hang *et al.*, 2022], which is automatically set by an evolutionary search algorithm. We perform a coarse grid search to fine-tune the hyperparameters for PDCFR+, and the best one, *i.e.*, $\alpha = 2.3$ and $\gamma = 5$, is then used across all games. All algorithms utilize the alternating-updates technique.

We run each algorithm for 20,000 iterations in each testing game to observe their long-term behavior, with Figure 2 displaying exploitability curves. We set a minimum reachable exploitability of 10^{-12}, denoting the point where the average strategy profile is considered sufficiently converged to a NE. Across non-poker games (*Goofspiel*, *Battleship*, *Liar's Dice*), PDCFR+ outperforms others by 4-8 orders of magnitude, taking 2,000-12,000 iterations to achieve the minimum reachable exploitability. For poker games, PDCFR+ excels in Kuhn Poker. In Leduc Poker, PDCFR+ initially lags behind DCFR and CFR+ up to 10,000 iterations but exhibits faster convergence thereafter. In the large-scale poker game *HUNL Subgame*, PDCFR+ performs comparably to CFR+ and PCFR+, while DCFR+ emerges as the fastest algorithm. In summary, PDCFR+ stands as an improved version of PCFR+, with a similar property for excellence in non-poker games.

PDCFR+ employs $\gamma = 5$ in the computation of average strategies, while PCFR+ utilizes quadratic averaging, equivalent to $\gamma = 2$ in PDCFR+. In order to assess whether PDCFR+'s improved performance is solely attributable to its specific averaging of strategies, we present the performance of PCFR+ (5), which employs $\gamma = 5$, in Figure 2. PCFR+ (5) outperforms PCFR+ with quadratic averaging in nine games. PDCFR+ performs similarly to PCFR+ (5) in six games but
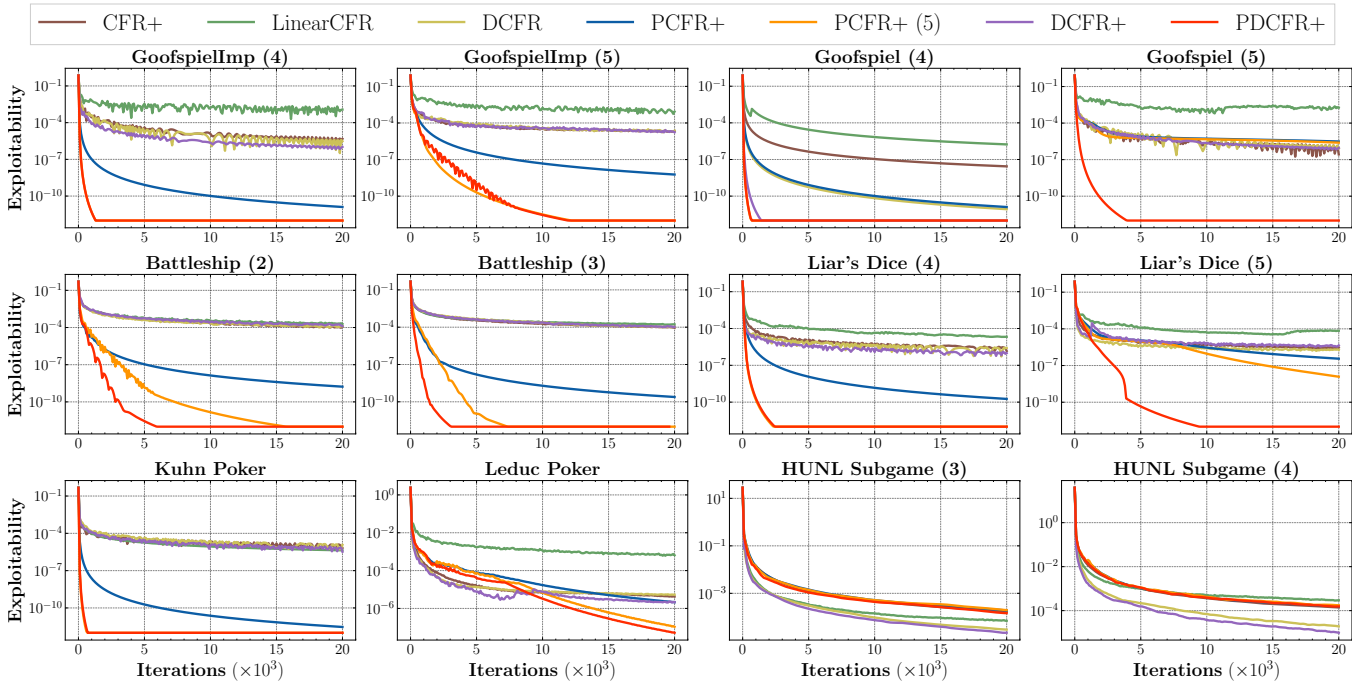
Figure 2: Convergence results of seven CFR variants on twelve testing games. Each algorithm runs for 20,000 iterations to display a long-time behavior. In all plots, the x-axis is the number of iteration, and the y-axis represents exploitability, displayed on a logarithmic scale.
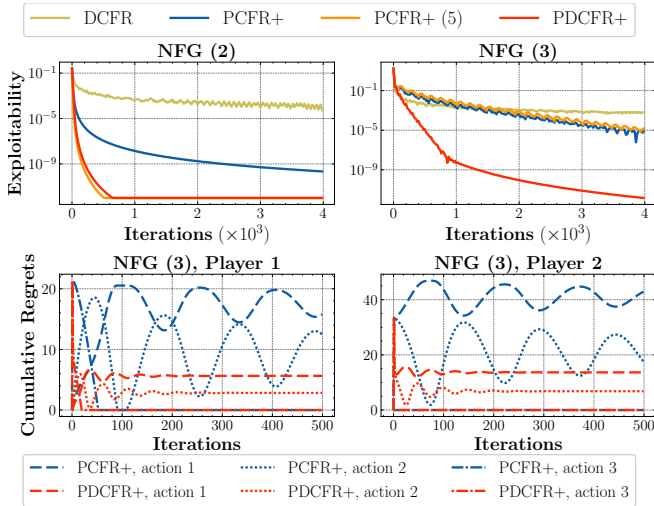

Figure 3: The top plots illustrate the convergence results of four CFR variants on two games. The bottom plots show that PDCFR+ quickly learns stable cumulative regrets in *NFG (3)*.

exhibits significantly faster convergence in five games. Particularly in *GoofSpiel (5)* and *Liar's Dice (5)*, PDCFR+ surpasses PCFR+ (5) by 5-7 orders of magnitude. Thus, we conclude that the acceleration observed in PDCFR+ is not solely a result of using a more aggressive weighting scheme for determining average strategies but is also attributed to the weighting scheme employed in determining regrets.

To further comprehend the enhanced performance of PD-CFR+, consider the normal-form games *NFG (2)* and *NFG*

(3) discussed in section 3. Top plots of Figure 3 show that PDCFR+ converges extremely fast on both games, not affected by the dominated actions. After the first iteration of PDCFR+ in *NFG (3)*, player 2 has high cumulative regrets $R^1 = \left(\frac{100}{3}, \frac{100}{3}, 0\right)$ due to the dominated action 3. As shown in the bottom-right plot of Figure 3, PDCFR+ rapidly discounts cumulative regrets, and at around the 200th iteration, it has already learned stable 2:1 cumulative regrets for actions 1 and 2, producing an approximate NE. In contrast, PCFR+ is still struggling with fluctuations and requires much more iterations to approach a NE. This highlights PDCFR+'s ability to swiftly mitigate the negative effects of dominated actions and consistently leverage predictions to accelerate convergence.

6 Conclusions and Future Research

This work proposes to minimize weighted counterfactual regret with OMD and optimistic variant. It provides a new perspective to understand DCFR's superior performance and derives a novel CFR variant PDCFR+. PDCFR+ discounts cumulative regrets from early iterations to mitigate the negative effects of dominated actions and consistently leverages predictions to accelerate convergence. Experimental results demonstrate that PDCFR+ achieves competitive results compared with other CFR variants. Recent work [Farina *et al.*, 2023] proposes two fixes for PRM+, achieving $O(1/T)$ convergence in normal-form games. It is worthwhile to investigate whether our algorithm can achieve similar convergence by introducing a specific weighting sequence. Besides, combining PDCFR+ with function approximations [Brown *et al.*, 2019] and integrating it with the dynamic discounting framework [Hang *et al.*, 2024] are also promising future works.

Acknowledgments

References

[Abernethy *et al.*, 2018] Jacob Abernethy, Kevin A Lai, Kfir Y Levy, and Jun-Kun Wang. Faster rates for convex-concave games. In *Conference On Learning Theory*, pages 1595–1625, 2018.

[Beck and Teboulle, 2003] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[Bowling *et al.*, 2015] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015.

[Brown and Sandholm, 2017] Noam Brown and Tuomas Sandholm. Safe and nested subgame solving for imperfect-information games. In *Advances in Neural Information Processing Systems*, pages 689–699, 2017.

[Brown and Sandholm, 2018] Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

[Brown and Sandholm, 2019a] Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *AAAI Conference on Artificial Intelligence*, pages 1829–1836, 2019.

[Brown and Sandholm, 2019b] Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456):885–890, 2019.

[Brown *et al.*, 2018] Noam Brown, Tuomas Sandholm, and Brandon Amos. Depth-limited solving for imperfect-information games. In *Advances in Neural Information Processing Systems*, pages 7674–7685, 2018.

[Brown *et al.*, 2019] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *International Conference on Machine Learning*, pages 793–802, 2019.

[Burch *et al.*, 2014] Neil Burch, Michael Johanson, and Michael Bowling. Solving imperfect information games using decomposition. In *AAAI Conference on Artificial Intelligence*, pages 602–608, 2014.

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

[Farina *et al.*, 2019b] 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*, pages 9229–9239, 2019.

[Farina *et al.*, 2021] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. In *AAAI Conference on Artificial Intelligence*, pages 5363–5371, 2021.

[Farina *et al.*, 2023] 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.

[Grand-Clément and Kroer, 2023] Julien Grand-Clément and Christian Kroer. Solving optimization problems with blackwell approachability. *Mathematics of Operations Research*, 2023.

[Gratch *et al.*, 2016] Jonathan Gratch, Zahra Nazari, and Emmanuel Johnson. The misrepresentation game: How to win at negotiation while seeming like a nice guy. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 728–737, 2016.

[Hang *et al.*, 2022] Xu Hang, Li Kai, Fu Haobo, Fu Qiang, and Xing Junliang. Autocfr: Learning to design counterfactual regret minimization algorithms. In *AAAI Conference on Artificial Intelligence*, pages 5244–5251, 2022.

[Hang *et al.*, 2024] Xu Hang, Li Kai, Fu Haobo, Fu Qiang, Xing Junliang, and Jian Cheng. Dynamic discounted counterfactual regret minimization. In *International Conference on Learning Representations*, pages 1–18, 2024.

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

[Kuhn, 1950] Harold W Kuhn. A simplified two-person poker. *Contributions to the Theory of Games*, 1:97–103, 1950.

[Lisỳ *et al.*, 2015] Viliam Lisỳ, 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*, pages 27–36, 2015.

[Lisy *et al.*, 2016] Viliam Lisy, Trevor Davis, and Michael Bowling. Counterfactual regret minimization in sequential security games. In *AAAI Conference on Artificial Intelligence*, pages 544–550, 2016.

[Liu *et al.*, 2022] Weiming Liu, Huacong Jiang, Bin Li, and Houqiang Li. Equivalence analysis between counterfactual regret minimization and online mirror descent. In *International Conference on Machine Learning*, pages 13717–13745, 2022.

[Moravčík *et al.*, 2017] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and

Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[Perolat *et al.*, 2021] Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning*, pages 8525–8535, 2021.

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

[Sandholm, 2015] Tuomas Sandholm. Steering evolution strategically: Computational game theory and opponent exploitation for treatment planning, drug design, and synthetic biology. In *AAAI Conference on Artificial Intelligence*, pages 4057–4061, 2015.

[Southey *et al.*, 2005] 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*, pages 550–558, 2005.

[Syrgkanis *et al.*, 2015] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.

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

[Wang and Abernethy, 2018] Jun-Kun Wang and Jacob D Abernethy. Acceleration through optimistic no-regret dynamics. In *Advances in Neural Information Processing Systems*, pages 3828–3838, 2018.

[Zinkevich *et al.*, 2007] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, pages 1729–1736, 2007.

[Zinkevich, 2003] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003.