Faster Game Solving via Predictive Blackwell Approachability: Connecting Regret Matching and Mirror Descent

Abstract

Blackwell approachability is a framework for reasoning about repeated games with vector-valued payoffs. We introduce *predictive* Blackwell approachability, where an estimate of the next payoff vector is given, and the decision maker tries to achieve better performance based on the accuracy of that estimator. In order to derive algorithms that achieve predictive Blackwell approachability, we start by showing a powerful connection between four well-known algorithms. *Follow-the-regularized-leader (FTRL)* and *online mirror descent (OMD)* are the most prevalent regret minimizers in online convex optimization. In spite of this prevalence, the *regret matching (RM)* and *regret matching*$^+$ *(RM*$^+$*)* algorithms have been preferred in the practice of solving large-scale games (as the local regret minimizers within the counterfactual regret minimization framework). We show that RM and RM$^+$ are the algorithms that result from running FTRL and OMD, respectively, to select the halfspace to force at all times in the underlying Blackwell approachability game. By applying the predictive variants of FTRL or OMD to this connection, we obtain predictive Blackwell approachability algorithms, as well as predictive variants of RM and RM$^+$. In experiments across 18 common zero-sum extensive-form benchmark games, we show that predictive RM$^+$ coupled with counterfactual regret minimization converges vastly faster than the fastest prior algorithms (CFR$^+$, DCFR, LCFR) across all games but two of the poker games, sometimes by two or more orders of magnitude.

1 Introduction

Extensive-form games (EFGs) are the standard class of games that can be used to model sequential interaction, outcome uncertainty, and imperfect information. Operationalizing these models requires algorithms for computing game-theoretic equilibria. A recent success of EFGs is the use of Nash equilibrium for several recent poker AI milestones, such as essentially solving the game of limit Texas hold'em (Bowling et al. 2015), and beating top human poker pros in no-limit Texas hold'em with the *Libratus* AI (Brown and Sandholm 2017). A central component of all recent poker AIs has been a fast iterative method for computing approximate Nash equilibrium at scale. The leading approach is the *counterfactual regret minimization (CFR)* framework, where the problem of minimizing regret over a player's strategy space of an EFG is decomposed into a set of regret-minimization problems over probability simplexes (Zinkevich et al. 2007; Farina, Kroer, and Sandholm 2019c). Each simplex represents the probability over actions at a given decision point. The CFR setup can be combined with any regret minimizer for the simplexes. If both players in a zero-sum EFG repeatedly play each other using a CFR algorithm, the average strategies converge to a Nash equilibrium. Initially *regret matching* (RM) was the prevalent simplex regret minimizer used in CFR. Later, it was found that by alternating strategy updates between the players, taking linear averages of strategy iterates over time, and using a variation of RM called *regret-matching*$^+$ *(RM*$^+$*)* (Tammelin 2014) leads to significantly faster convergence in practice. This variation is called CFR$^+$. Both CFR and CFR$^+$ guarantee convergence to Nash equilibrium at a rate of $T^{-1/2}$. CFR$^+$ has been used in every milestone in developing poker AIs in the last decade (Bowling et al. 2015; Moravčík et al. 2017; Brown and Sandholm 2017, 2019b). This is in spite of the fact that its theoretical rate of convergence is the same as that of CFR with RM (Tammelin 2014; Farina, Kroer, and Sandholm 2019a; Burch, Moravcik, and Schmid 2019), and there exist algorithms which converge at a faster rate of T^{-1} (Hoda et al. 2010; Kroer et al. 2020; Farina, Kroer, and Sandholm 2019b). In spite of this theoretically-inferior convergence rate, CFR$^+$ has repeatedly performed favorably against T^{-1} methods for EFGs Kroer, Farina, and Sandholm (2018); Kroer et al. (2020); Farina, Kroer, and Sandholm (2019b); Gao, Kroer, and Goldfarb (2019). Similarly, the *follow-the-regularized-leader (FTRL)* and *online mirror descent (OMD)* regret minimizers, the two most prominent algorithms in online convex optimization, can be instantiated to have a better dependence on dimensionality than RM$^+$ and RM, yet RM$^+$ has been found to be superior (Brown, Kroer, and Sandholm 2017).

There has been some interest in connecting RM to the more prevalent (and more general) online convex optimization algorithms such as OMD and FTRL, as well as classical first-order methods. Waugh and Bagnell (2015) showed that RM is equivalent to Nesterov's dual averaging algorithm (which is an offline version of FTRL), though this equivalence requires specialized step sizes that are proven correct by invoking the correctness of RM itself. Burch (2018) studies RM and RM$^+$, and contrasts them with mirror descent and other prox-based methods.

We show a strong connection between RM, RM$^+$, and FTRL, OMD. This connection arises via *Blackwell approachability*, a framework for playing games with vector-valued payoffs, where the goal is to get the average payoff to approach some convex target set. Blackwell originally showed that this can be achieved by repeatedly *forcing* the payoffs to lie in a sequence of halfspaces containing the target set Blackwell (1956). Our results are based on extending an equivalence between approachability and regret minimization Abernethy, Bartlett, and Hazan (2011). We show that RM and RM$^+$ are the algorithms that result from running FTRL and OMD, respectively, to select the halfspace to force at all times in the underlying Blackwell approachability game. The equivalence holds for any constant step size. Thus, RM and RM$^+$, the two premier regret minimizers in EFG solving, turn out to follow exactly from the two most prevalent regret minimizers from online optimization theory. This is surprising for several reasons:

- RM$^+$ was originally discovered as a heuristic modification of RM in order to avoid accumulating large negative regrets. In contrast, OMD and FTRL were developed separately from each other.
- When applying FTRL and OMD directly to the strategy space of each player, Farina, Kroer, and Sandholm (2019b, 2020) found that FTRL seems to perform better than OMD, even when using stochastic gradients. This relationship is reversed here, as RM$^+$ is *vastly* faster numerically than RM.
- The dual averaging algorithm (whose simplest variant is an offline version of FTRL), was originally developed in order to have increasing weight put on more recent gradients, as opposed to OMD which has constant or decreasing weight (Nesterov 2009). Here this relationship is reversed: OMD (which we show has a close link to RM$^+$) thresholds away old negative regrets, whereas FTRL keeps them around. Thus OMD ends up being *more* reactive to recent gradients in our setting.
- FTRL and OMD both have a step-size parameter that needs to be set according to the magnitude of gradients, while RM and RM$^+$ are parameter free (which is a desirable feature from a practical perspective). To reconcile this seeming contradiction, we show that the step-size parameter does not affect which halfspaces are forced, so any choice of step size leads to RM and RM$^+$.

Leveraging our connection, we study the algorithms that result from applying predictive variants of FTRL and OMD to choosing which halfspace to force. By applying predictive OMD we get the first predictive variant of RM$^+$, that is, one that has regret that depends on how good the sequence of predicted regret vectors is (as a side note of their paper, Brown and Sandholm (2019a) also tried a heuristic for optimism/predictiveness by counting the last regret vector twice in RM$^+$, but this does not yield a predictive algorithm). We call our regret minimizer *predictive regret matching*$^+$ (PRM$^+$). We go on to instantiate CFR with PRM$^+$ using the two standard techniques—alternation and quadratic averaging—-and find that it often converges much faster than CFR$^+$ and every other prior CFR variant, sometimes by several orders of magnitude. We show this on a large suite of common benchmark EFGs. However, we find that on poker games (except shallow ones), *discounted CFR (DCFR)* (Brown and Sandholm 2019a) is the fastest. We conclude that our algorithm based on PRM$^+$ yields the new state-of-the-art convergence rate for the remaining games. Our results also highlight the need to test on EFGs other than poker, as our non-poker results invert the superiority of prior algorithms as compared to recent results on poker.

2 Online Linear Optimization, Regret Minimizers, and Predictions

At each time t, an oracle for the *online linear optimization (OLO)* problem supports the following two operations, in order: NEXTSTRATEGY returns a point $\boldsymbol{x}^t \in \mathcal{D} \subseteq \mathbb{R}^n$, and OBSERVELOSS receives a *loss vector* $\boldsymbol{\ell}^t$ that is meant to evaluate the strategy $\boldsymbol{x}^t$ that was last output. Specifically, the oracle incurs a loss equal to $\langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle$. The loss vector $\boldsymbol{\ell}^t$ can depend on all past strategies that were output by the oracle. The oracle operates *online* in the sense that each strategy $\boldsymbol{x}^t$ can depend only on the decision $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{t-1}$ output in the past, as well as the loss vectors $\boldsymbol{\ell}^1, \ldots, \boldsymbol{\ell}^{t-1}$ that were observed in the past. No information about the future losses $\boldsymbol{\ell}^t, \boldsymbol{\ell}^{t+1}, \ldots$ is available to the oracle at time t. The objective of the oracle is to make sure the *regret*

$$R^T(\hat{\boldsymbol{x}}) := \sum_{t=1}^{T} \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle - \sum_{t=1}^{T} \langle \boldsymbol{\ell}^t, \hat{\boldsymbol{x}} \rangle = \sum_{t=1}^{T} \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t - \hat{\boldsymbol{x}} \rangle,$$

which measures the difference between the total loss incurred up to time T compared to always using the *fixed* strategy $\hat{\boldsymbol{x}}$, does not grow too fast as a function of time T. Oracles that guarantee that $R^T(\hat{\boldsymbol{x}})$ grow sublinearly in T in the worst case for all $\hat{\boldsymbol{x}} \in \mathcal{D}$ (no matter the sequence of losses $\boldsymbol{\ell}^1, \ldots, \boldsymbol{\ell}^T$ observed) are called *regret minimizers*. While most theory about regret minimizers is developed under the assumption that the domain $\mathcal{D}$ is *convex* and *compact*, in this paper we will need to consider sets $\mathcal{D}$ that are convex and closed, but unbounded (hence, not compact).

Incorporating Predictions

A recent trend in online learning has been concerned with constructing oracles that can incorporate *predictions* of the next loss vector $\boldsymbol{\ell}^t$ in the decision making (Chiang et al. 2012; Rakhlin and Sridharan 2013a,b). Specifically, a *predictive* oracle differs from a regular (that is, non-predictive) oracle for OLO in that the NEXTSTRATEGY function receives a *prediction* $\boldsymbol{m}^t \in \mathbb{R}^n$ of the next loss $\boldsymbol{\ell}^t$ at all times t. Conceptually, a "good" predictive regret minimizer should guarantee a superior regret bound than a non-predictive regret minimizer if $\boldsymbol{m}^t \approx \boldsymbol{\ell}^t$ at all times t. Algorithms exist that can guarantee this. For instance, it is always possible to construct an oracle that guarantees that $R^T = O(1+\sum_{t=1}^{T} \|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|^2)$, which implies that the regret stays constant when $\boldsymbol{m}^t$ is clairvoyant. In fact, even stronger regret bounds can be attained: for example, Syrgkanis et al.

Algorithm 1: (Predictive) FTRL

$\boldsymbol{L}^0 \leftarrow \mathbf{0} \in \mathbb{R}^n$

function NEXTSTRATEGY($\boldsymbol{m}^t$)

  ▷ Set $\boldsymbol{m}^t = \mathbf{0}$ for non-predictive version

  return $\underset{\hat{\boldsymbol{x}} \in \mathcal{D}}{\arg\min} \left\{ \langle \boldsymbol{L}^{t-1} + \boldsymbol{m}^t, \hat{\boldsymbol{x}} \rangle + \frac{1}{\eta}\varphi(\hat{\boldsymbol{x}}) \right\}$

function OBSERVELOSS($\boldsymbol{\ell}^t$)

  $\boldsymbol{L}^t \leftarrow \boldsymbol{L}^{t-1} + \boldsymbol{\ell}^t$

Algorithm 2: (Predictive) OMD

$\boldsymbol{z}^0 \in \mathcal{D}$ such that $\nabla\varphi(\boldsymbol{z}^0) = \mathbf{0}$

function NEXTSTRATEGY($\boldsymbol{m}^t$)

  ▷ Set $\boldsymbol{m}^t = \mathbf{0}$ for non-predictive version

  return $\underset{\hat{\boldsymbol{x}} \in \mathcal{D}}{\arg\min} \left\{ \langle \boldsymbol{m}^t, \hat{\boldsymbol{x}} \rangle + \frac{1}{\eta} D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^{t-1}) \right\}$

function OBSERVELOSS($\boldsymbol{\ell}^t$)

  $\boldsymbol{z}^t \leftarrow \underset{\hat{\boldsymbol{z}} \in \mathcal{D}}{\arg\min} \left\{ \langle \boldsymbol{\ell}^t, \hat{\boldsymbol{z}} \rangle + \frac{1}{\eta} D_\varphi(\hat{\boldsymbol{z}} \,\|\, \boldsymbol{z}^{t-1}) \right\}$

(2015) show that the sharper *Regret bounded by Variation in Utilities (RVU)* condition can be attained, while Farina et al. (2019) focus on *stable-predictivity*.

FTRL, OMD, and their Predictive Variants

Follow-the-regularized-leader (FTRL) (Shalev-Shwartz and Singer 2007) and *online mirror descent (OMD)* are the two best known oracles for the online linear optimization problem. Their *predictive* variants are relatively new and can be traced back to the works by Rakhlin and Sridharan (2013a) and Syrgkanis et al. (2015). Since the original FTRL and OMD algorithms correspond to predictive FTRL and predictive OMD when the prediction $\boldsymbol{m}^t$ is set to the $\mathbf{0}$ vector at all t, the implementation of FTRL in Algorithm 1 and OMD in Algorithm 2 captures both algorithms. In both algorithm, $\eta > 0$ is an arbitrary step size parameter, $\mathcal{D} \subseteq \mathbb{R}^n$ is a convex and closed set, and $\varphi : \mathcal{D} \to \mathbb{R}_{\geq 0}$ is a 1-strongly convex differentiable regularizer (with respect to some norm $\|\cdot\|$). The symbol $D_\varphi(\,\|\,)$ used in OMD denotes the *Bregman divergence* associated with φ, defined as $D_\varphi(\boldsymbol{x} \,\|\, \boldsymbol{c}) \coloneqq \varphi(\boldsymbol{x}) - \varphi(\boldsymbol{c}) - \langle \nabla\varphi(\boldsymbol{c}), \boldsymbol{x} - \boldsymbol{c} \rangle$ for all $\boldsymbol{x}, \boldsymbol{c} \in \mathcal{D}$.

We state regret guarantees for (predictive) FTRL and (predictive) OMD in Proposition 1. Our statements are slightly more general than those by Syrgkanis et al. (2015), in that we (i) do not assume that the domain is a simplex, and (ii) do not use quantities that might be unbounded in non-compact domains $\mathcal{D}$. A proof of the regret bounds is in Appendix A for FTRL and Appendix B for OMD.

Proposition 1. *At all times T, the regret cumulated by (predictive) FTRL (Algorithm 1) and (predictive) OMD (Algorithm 2) compared to* any *strategy $\hat{\boldsymbol{x}} \in \mathcal{D}$ is bounded as*

$$R^T(\hat{\boldsymbol{x}}) \leq \frac{\varphi(\hat{\boldsymbol{x}})}{\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{c\eta} \sum_{t=1}^{T-1} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2, \tag{1}$$

where $c = 4$ for FTRL and $c = 8$ for OMD, and where $\|\cdot\|_*$ denotes the dual of the norm $\|\cdot\|$ with respect to which φ is 1*-strongly convex.*

Proposition 1 implies that, by appropriately setting the step size parameter (for example, $\eta = T^{-1/2}$), (predictive) FTRL and (predictive) OMD guarantee $R^T(\hat{\boldsymbol{x}}) = O(T^{1/2})$ for all $\hat{\boldsymbol{x}}$. Hence, (predictive) FTRL and (predictive) OMD are regret minimizers.

3 Blackwell Approachability

Blackwell approachability (Blackwell 1956) generalizes the problem of playing a repeated two-player game to games whose utilites are vectors instead of scalars. In a Blackwell approachability game, at all times t, two players interact in this order: first, Player 1 selects an action $\boldsymbol{x}^t \in \mathcal{X}$; then, Player 2 selects an action $\boldsymbol{y}^t \in \mathcal{Y}$; finally, Player 1 incurs the vector-valued payoff $\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) \in \mathbb{R}^d$, where $\boldsymbol{u}$ is a biaffine function. The sets $\mathcal{X}, \mathcal{Y}$ of player actions are assumed to be compact convex sets. Player 1's objective is to guarantee that the average payoff converges to some desired closed convex *target set* $S \subseteq \mathbb{R}^d$. Formally, given target set $S \subseteq \mathbb{R}^d$, Player 1's goal is to pick actions $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots \in \mathcal{X}$ such that no matter the actions $\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots \in \mathcal{Y}$ played by Player 2,

$$\min_{\hat{\boldsymbol{s}} \in S} \left\| \hat{\boldsymbol{s}} - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) \right\|_2 \to 0 \quad \text{as} \quad T \to \infty. \tag{2}$$

A central concept in the theory of Blackwell approachability is the following.

Definition 1 (Approachable halfspace, forcing function)**.** *Let $(\mathcal{X}, \mathcal{Y}, \boldsymbol{u}(\cdot,\cdot), S)$ be a Blackwell approachability game as described above and let $H \subseteq \mathbb{R}^d$ be a halfspace, that is, a set of the form $H = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{a}^\top \boldsymbol{x} \leq b\}$ for some $\boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R}$. The halfspace H is said to be* forceable *if there exists a strategy of Player 1 that guarantees that the payoff is in H no matter the actions played by Player 2. In symbols, H is forceable if there exists $\boldsymbol{x}^* \in \mathcal{X}$ such that for all $\boldsymbol{y} \in \mathcal{Y}$, $\boldsymbol{u}(\boldsymbol{x}^*, \boldsymbol{y}) \in H$. When this is the case, we call action $\boldsymbol{x}^*$ a* forcing action *for H.*

Blackwell's *approachability theorem* (Blackwell 1956) states that goal (2) can be attained if and only if all halfspaces $H \supseteq S$ are forceable. Blackwell approachability has a number of applications and connections to other problems in the online learning and game theory literature (e.g., (Blackwell 1954; Foster 1999; Hart and Mas-Colell 2000)).

In this paper we leverage the Blackwell approachability formalism to draw new connections between FTRL and OMD with RM and RM$^+$, respectively. We also introduce predictive Blackwell approachability, and show that it can be used to develop new state-of-the-art algorithms for simplex domains and imperfect-information extensive-form zero-sum games.

Algorithm 3: From OLO to (predictive) approachability

Data: $\mathcal{D} \subseteq \mathbb{R}^n$ convex and closed, s.t. $\mathcal{K} := C^\circ \cap \mathbb{B}_2^n \subseteq \mathcal{D} \subseteq C^\circ$
$\mathcal{L}$ online linear optimization algorithm for domain $\mathcal{D}$

```
function NextStrategy(v^t)
    ▷ Set v^t = 0 for non-predictive version
    θ^t ← L.NextStrategy(−v^t)
    return x^t ∈ X

function ReceivePayoff(u(x^t, y^t))
    L.ObserveLoss(−u(x^t, y^t))
```

Figure 1: Reduction from an OLO oracle to a strategy for playing a Blackwell approachability game.

4 From Online Linear Optimization to Blackwell Approachability

Abernethy, Bartlett, and Hazan (2011) showed that it is always possible to convert a regret minimizer into an algorithm for a Blackwell approachability game (that is, an algorithm that chooses actions $\boldsymbol{x}^t$ at all times t in such a way that goal (2) holds no matter the actions $\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots$ played by the opponent). In this section, we slightly extend their constructive proof by allowing more flexibility in the choice of the domain of the regret minimizer. This extra flexibility will be needed to show that RM and RM$^+$ can be obtained directly from FTRL and OMD, respectively.

We start from the case where the target set in the Blackwell approachability game is a closed convex cone $C \subseteq \mathbb{R}^n$. As Proposition 2 shows, Algorithm 3 provides a way of playing the Blackwell approachability game that guarantees that (2) is satisfied (the proof is in Appendix C). In broad strokes, Algorithm 3 works as follows (see also Figure 2): the regret minimizer has as its decision space the polar cone to C (or a subset thereof), and its decision is used as the normal vector in choosing a halfspace to force. At time t, the algorithm plays a forcing action $\boldsymbol{x}^t$ for the halfspace H_t induced by the last decision $\boldsymbol{\theta}^t$ output by the OLO oracle $\mathcal{L}$. Then, $\mathcal{L}$ incurs the loss $-\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t)$, where $\boldsymbol{u}$ is the payoff function of the Blackwell approachability game.

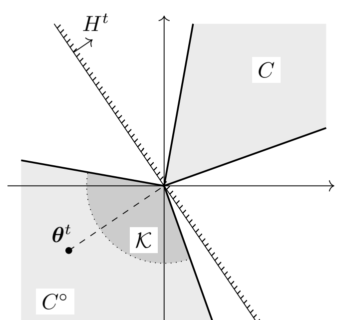


Figure 2: Pictorial depiction of Algorithm 3's inner working: at all times t, the algorithm plays a forcing action for the halfspace H^t induced by the last decision output by the OLO oracle $\mathcal{L}$.

Proposition 2. *Let $(\mathcal{X}, \mathcal{Y}, \boldsymbol{u}(\cdot,\cdot), C)$ be an approachability game, where $C \subseteq \mathbb{R}^n$ is a closed convex cone, such that each halfspace $H \supseteq C$ is approachable (Definition 1). Let $\mathcal{K} := C^\circ \cap \mathbb{B}_2^n$, where $C^\circ = \{\boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{x}, \boldsymbol{y}\rangle \le 0 \, \forall \boldsymbol{y} \in C\}$ denotes the* polar cone *to C and $\mathbb{B}_2^n := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \le 1\}$ is the unit ball. Finally, let $\mathcal{L}$ be an oracle for the OLO problem (for example, the FTRL or OMD algorithm) whose domain of decisions is any closed convex set $\mathcal{D}$, such that $\mathcal{K} \subseteq \mathcal{D} \subseteq C^\circ$. Then, at all times T, the distance between the average payoff cumulated by Algorithm 3 and the target cone C is upper bounded as*

$$\min_{\hat{\boldsymbol{s}} \in C} \left\| \hat{\boldsymbol{s}} - \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) \right\|_2 = \frac{1}{T} \max_{\hat{\boldsymbol{x}} \in \mathcal{K}} R_{\mathcal{L}}^T(\hat{\boldsymbol{x}}),$$

where $R_{\mathcal{L}}^T(\hat{\boldsymbol{x}})$ is the regret cumulated by $\mathcal{L}$ up to time T compared to always playing $\hat{\boldsymbol{x}} \in \mathcal{K}$.

As $\mathcal{K}$ is compact, by virtue of $\mathcal{L}$ being a regret minimizer, $1/T \cdot \max_{\hat{\boldsymbol{x}} \in \mathcal{K}} R^T(\hat{\boldsymbol{x}}) \to 0$ as $T \to \infty$, Algorithm 3 satisfies the Blackwell approachability goal (2). The fact that Proposition 2 applies only to conic target sets does not limit its applicability. Indeed, Abernethy, Bartlett, and Hazan (2011) showed that any Blackwell approachability game with a non-conic target set can be efficiently transformed to another one with a conic target set. In this paper, we only need to focus on conic target sets.

The construction by Abernethy, Bartlett, and Hazan (2011) coincides with Proposition 2 in the special case where the domain $\mathcal{D}$ is set to $\mathcal{D} = \mathcal{K}$. In the next section, we will need our added flexibility in the choice of $\mathcal{D}$: in order to establish the connection between RM$^+$ and OMD, it is necessary to set $\mathcal{D} = C^\circ \neq \mathcal{K}$.

5 Connecting FTRL, OMD with RM, RM$^+$

Constructing a regret minimizer for a simplex domain $\Delta^n := \{\boldsymbol{x} \in \mathbb{R}_{\ge 0} : \|\boldsymbol{x}\|_1 = 1\}$ can be reduced to constructing an algorithm for a particular Blackwell approachability game $\Gamma := (\Delta^n, \mathbb{R}^n, \boldsymbol{u}(\cdot,\cdot), \mathbb{R}^n_{\le 0})$ that we now describe Hart and Mas-Colell (2000). For all $i \in \{1, \ldots, n\}$, the i-th component of the vector-valued payoff function $\boldsymbol{u}$ measures the change in regret incurred at time t, compared to always playing the i-th vertex $\boldsymbol{e}_i$ of the simplex. Formally, $\boldsymbol{u} : \Delta^n \times \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) = \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t, \tag{3}$$

where $\mathbf{1}$ is the n-dimensional vector whose components are all 1. It is known that Γ is such that the halfspace $H_{\boldsymbol{a}} := \{\boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{x}, \boldsymbol{a}\rangle \le 0\} \supseteq \mathbb{R}^n_{\le 0}$ is forceable (Definition 1) for all $\boldsymbol{a} \in \mathbb{R}^n_{\ge 0}$. A forcing action for $H_{\boldsymbol{a}}$ is given by $\boldsymbol{g}(\boldsymbol{a}) := \boldsymbol{a}/\|\boldsymbol{a}\|_1 \in \Delta^n$ when $\boldsymbol{a} \neq \mathbf{0}$; when $\boldsymbol{a} = \mathbf{0}$, any $\boldsymbol{x} \in \Delta^n$ is a forcing action. The following is known.

Lemma 1. *The regret $R^T(\hat{\boldsymbol{x}}) = \frac{1}{T}\sum_{t=1}^{T} \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t - \hat{\boldsymbol{x}} \rangle$ cumulated up to any time T by the decisions $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^T \in \Delta^n$ compared to any $\hat{\boldsymbol{x}} \in \Delta^n$ is related to the distance of the average Blackwell payoff from the target cone $\mathbb{R}^n_{\le 0}$ as*

$$\frac{1}{T} R^T(\hat{\boldsymbol{x}}) \le \min_{\hat{\boldsymbol{s}} \in \mathbb{R}^n_{\le 0}} \left\| \hat{\boldsymbol{s}} - \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \right\|_2 . \tag{4}$$

Algorithm 4: (Predictive) regret matching

1 $\boldsymbol{r}^0 \leftarrow \mathbf{0} \in \mathbb{R}^n,\ \boldsymbol{x}^0 \leftarrow \mathbf{1}/n \in \Delta^n$
2 **function** NEXTSTRATEGY($\boldsymbol{m}^t$)
  ▷ Set $\boldsymbol{m}^t = \mathbf{0}$ for non-predictive version
3   $\boldsymbol{\theta}^t \leftarrow [\boldsymbol{r}^{t-1} + \langle \boldsymbol{m}^t, \boldsymbol{x}^{t-1} \rangle \mathbf{1} - \boldsymbol{m}^t]^+$
4   **if** $\boldsymbol{\theta}^t \neq \mathbf{0}$ **return** $\boldsymbol{x}^t \leftarrow \boldsymbol{\theta}^t \,/\, \|\boldsymbol{\theta}^t\|_1$
5   **else** **return** $\boldsymbol{x}^t \leftarrow$ arbitrary point in Δ^n
6 **function** OBSERVELOSS($\boldsymbol{\ell}^t$)
7   $\boldsymbol{r}^t \leftarrow \boldsymbol{r}^t + \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t$

Algorithm 5: (Predictive) regret matching$^+$

1 $\boldsymbol{z}^0 \leftarrow \mathbf{0} \in \mathbb{R}^n,\ \boldsymbol{x}^0 \leftarrow \mathbf{1}/n \in \Delta^n$
2 **function** NEXTSTRATEGY($\boldsymbol{m}^t$)
  ▷ Set $\boldsymbol{m}^t = \mathbf{0}$ for non-predictive version
3   $\boldsymbol{\theta}^t \leftarrow [\boldsymbol{z}^{t-1} + \langle \boldsymbol{m}^t, \boldsymbol{x}^{t-1} \rangle \mathbf{1} - \boldsymbol{m}^t]^+$
4   **if** $\boldsymbol{\theta}^t \neq \mathbf{0}$ **return** $\boldsymbol{x}^t \leftarrow \boldsymbol{\theta}^t \,/\, \|\boldsymbol{\theta}^t\|_1$
5   **else** **return** $\boldsymbol{x}^t \leftarrow$ arbitrary point in Δ^n
6 **function** OBSERVELOSS($\boldsymbol{\ell}^t$)
7   $\boldsymbol{z}^t \leftarrow [\boldsymbol{z}^{t-1} + \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t]^+$

So, a strategy for the Blackwell approachability game Γ is a regret-minimizing strategy for the simplex domain Δ^n.

When the approachability game Γ is solved by means of the constructive proof of Blackwell's approachability theorem (Blackwell 1956), one recovers a particular regret minimizer for the domain Δ^n known as the *regret matching (RM)* algorithm Hart and Mas-Colell (2000). The same cannot be said for the closely related RM$^+$ algorithm Tammelin (2014), which converges significantly faster in practice than RM, as has been reported many times.

We now uncover deep and surprising connections between RM, RM$^+$ and the OLO algorithms FTRL, OMD by solving Γ using Algorithm 3. Let $\mathcal{L}_\eta^{\text{ftrl}}$ be the FTRL algorithm instantiated over the conic domain $\mathcal{D} = \mathbb{R}^n_{\geq 0}$ with the 1-strongly convex regularizer $\varphi(\boldsymbol{x}) = {}^1\!/\!_2\,\|\boldsymbol{x}\|_2^2$ and an arbitrary step size parameter η. Similarly, let $\mathcal{L}_\eta^{\text{omd}}$ be the OMD algorithm instantiated over the same domain $\mathcal{D} = \mathbb{R}^n_{\geq 0}$ with the same convex regularizer $\varphi(\boldsymbol{x}) = {}^1\!/\!_2\,\|\boldsymbol{x}\|_2^2$. Since $\mathbb{R}^n_{\geq 0} = (\mathbb{R}^n_{\leq 0})^\circ$, $\mathcal{D}$ satisfies the requirements of Proposition 2. So, $\mathcal{L}_\eta^{\text{ftrl}}$ and $\mathcal{L}_\eta^{\text{omd}}$ can be plugged into Algorithm 3 to compute a strategy for the Blackwell approachability game Γ. When that is done, the following can be shown (all proofs for this section are in Appendix D).

Theorem 1 (FTRL reduces to RM)**.** *For all $\eta > 0$, when Algorithm 3 is set up with $\mathcal{D} = \mathbb{R}^n_{\geq 0}$ and regret minimizer $\mathcal{L}_\eta^{\text{ftrl}}$ to play Γ, it produces the same iterates as the RM algorithm.*

Theorem 2 (OMD reduces to RM$^+$)**.** *For all $\eta > 0$, when Algorithm 3 is set up with $\mathcal{D} = \mathbb{R}^n_{\geq 0}$ and regret minimizer $\mathcal{L}_\eta^{\text{omd}}$ to play Γ, it produces the same iterates as the RM$^+$ algorithm.*

Pseudocode for RM and RM$^+$ is given in Algorithms 4 and 5 (when $\boldsymbol{m}^t = \mathbf{0}$). In hindsight, the equivalence between RM and RM$^+$ with FTRL and OMD is clear. The computation of $\boldsymbol{\theta}^t$ on Line 3 in both PRM and PRM$^+$ corresponds to the closed-form solution for the minimization problems of Line 4 in FTRL and Line 3 in OMD, respectively, in accordance with Line 2 of Algorithm 3. Next, Lines 4 and 5 in both PRM and PRM$^+$ compute the forcing action required in Line 3 of Algorithm 3 using the function $\boldsymbol{g}$ defined above. Finally, in accordance with Line 6 of Algorithm 3, Line 7 of PRM corresponds to Line 6 of FTRL, and Line 7 of PRM$^+$ to Line 5 of OMD.

6 Predictive Blackwell Approachability, and Predictive RM and RM$^+$

It is natural to wonder whether it is possible to devise an algorithm for Blackwell approachability games that is able to guarantee faster convergence to the target set when good predictions of the next vector payoff are available. We call this setup *predictive Blackwell approachability*. We answer the question in the positive by leveraging Proposition 2. Since the loss incurred by the regret minimizer is $\boldsymbol{\ell}^t := -\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t)$ (Line 5 in Algorithm 3), any prediction $\boldsymbol{v}^t$ of the payoff $\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is naturally a prediction about the next loss incurred by the underlying regret minimizer $\mathcal{L}$ used in Algorithm 3. Hence, as long as the prediction is propagated as in Line 2 in Algorithm 3, Proposition 2 holds verbatim. In particular, we prove the following. All proofs for this section are in Appendix E.

Proposition 3. *Let $(\mathcal{X}, \mathcal{Y}, \boldsymbol{u}(\cdot,\cdot), S)$ be a Blackwell approachability game, where every halfspace $H \supseteq S$ is approachable (Definition 1). For all T, given predictions $\boldsymbol{v}^t$ of the payoff vectors, there exist algorithms for playing the game (that is, pick $\boldsymbol{x}^t \in \mathcal{X}$ at all t) that guarantee*

$$\min_{\hat{\boldsymbol{s}} \in S} \left\| \hat{\boldsymbol{s}} - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) \right\|_2 \leq \frac{1}{\sqrt{T}} \left(1 + \frac{2}{T} \sum_{t=1}^{T} \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) - \boldsymbol{v}^t\|_2^2 \right).$$

We now focus on how predictive Blackwell approachability ties into our discussion of RM and RM$^+$. In Section 5 we showed that when Algorithm 3 is used in conjunction with FTRL and OMD on the Blackwell approachability game Γ of Section 5, the iterates coincide with those of RM and RM$^+$, respectively. In the rest of this section we investigate the use of *predictive* FTRL and *predictive* OMD in that framework. Specifically, we use predictive FTRL and preditictive OMD as the regret regret minimizers to solve the Blackwell approachability game introduced in Section 5, and coin the resulting predictive regret minimization algorithms for simplex domains *predictive regret matching (PRM)* and *predictive regret matching$^+$ (PRM$^+$)*, respectively. Ideally, starting from the prediction $\boldsymbol{m}^t$ of the next loss, we would want the prediction $\boldsymbol{v}^t$ of the next utility in the equivalent Blackwell game Γ (Section 5) to be $\boldsymbol{v}^t = \langle \boldsymbol{m}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{m}^t$ to maintain symmetry with (3). However, $\boldsymbol{v}^t$ is computed before $\boldsymbol{x}^t$ is computed, and $\boldsymbol{x}^t$ depends on $\boldsymbol{v}^t$, so the previous expression requires the computation of a fixed point. To sidestep this issue, we let

$$\boldsymbol{v}^t := \langle \boldsymbol{m}^t, \boldsymbol{x}^{t-1} \rangle \mathbf{1} - \boldsymbol{m}^t$$

instead. We give pseudocode for PRM and PRM$^+$ as Algorithms 4 and 5. In the rest of this section, we discuss formal guarantees for PRM and PRM$^+$.

Theorem 3 (Correctness of PRM, PRM$^+$)**.** *Let $\mathcal{L}_\eta^{\text{ftrl}*}$ and $\mathcal{L}_\eta^{\text{omd}*}$ denote the predictive FTRL and predictive OMD algorithms instantiated with the same choice of regularizer and domain as in Section 5, and predictions $\boldsymbol{v}^t$ as defined above for the Blackwell approachability game Γ. For all $\eta > 0$, when Algorithm 3 is set up with $\mathcal{D} = \mathbb{R}^n_{\geq 0}$, the regret minimizer $\mathcal{L}_\eta^{\text{ftrl}*}$ (resp., $\mathcal{L}_\eta^{\text{omd}*}$) to play Γ, it produces the same iterates as the PRM (resp., PRM$^+$) algorithm. Furthermore, PRM and PRM$^+$ are regret minimizer for the domain Δ^n, and at all times T satisfy the regret bound*

$$R^T(\hat{\boldsymbol{x}}) \leq \sqrt{2}\left(\sum_{t=1}^{T} \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^t\|_2^2\right)^{1/2}.$$

At a high level, the main insight behind the regret bound of Theorem 3 is to combine Proposition 2 with the guarantees of predictive FTRL and predictive OMD (Proposition 1). In particular, combining (4) with Proposition 2, we find that the regret R^T cumulated by the strategies $\boldsymbol{x}^1, \dots, \boldsymbol{x}^T$ produced up to time T by PRM and PRM$^+$ satisfies

$$\frac{1}{T} \max_{\hat{\boldsymbol{x}} \in \Delta^n} R^T(\hat{\boldsymbol{x}}) \leq \frac{1}{T} \max_{\hat{\boldsymbol{x}} \in \mathbb{R}^n_{\geq 0} \cap \mathbb{B}^n_2} R^T_{\mathcal{L}}(\hat{\boldsymbol{x}}), \tag{5}$$

where $\mathcal{L} = \mathcal{L}_\eta^{\text{ftrl}*}$ for PRM and $\mathcal{L} = \mathcal{L}_\eta^{\text{omd}*}$ for PRM$^+$. Since the domain of the maximization on the right hand side is a subset of the domain $\mathcal{D} = \mathbb{R}^n_{\geq 0}$ of $\mathcal{L}$, the bound in Proposition 1 holds, and in particular

$$\begin{aligned}\max_{\hat{\boldsymbol{x}} \in \Delta^n} R^T(\hat{\boldsymbol{x}}) &\leq \max_{\hat{\boldsymbol{x}} \in \mathbb{R}^n_{\geq 0} \cap \mathbb{B}^n_2} \left\{\frac{\|\hat{\boldsymbol{x}}\|_2^2}{2\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^t\|_2^2\right\} \\ &\leq \left(\frac{1}{2\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^t\|_2^2\right),\end{aligned} \tag{6}$$

where in the first inequality we used the fact that $\varphi(\hat{\boldsymbol{x}}) = \|\hat{\boldsymbol{x}}\|_2^2/2$ by construction and in the second inequality we used the definition of unit ball $\mathbb{B}^n_2$. Finally, using the fact that the iterates produced by PRM and PRM$^+$ do not depend on the chosen step size $\eta > 0$ (first part of Theorem 3), we conclude that (6) must hold true for any $\eta > 0$, and so in particular also the $\eta > 0$ that minimizes the right hand side:

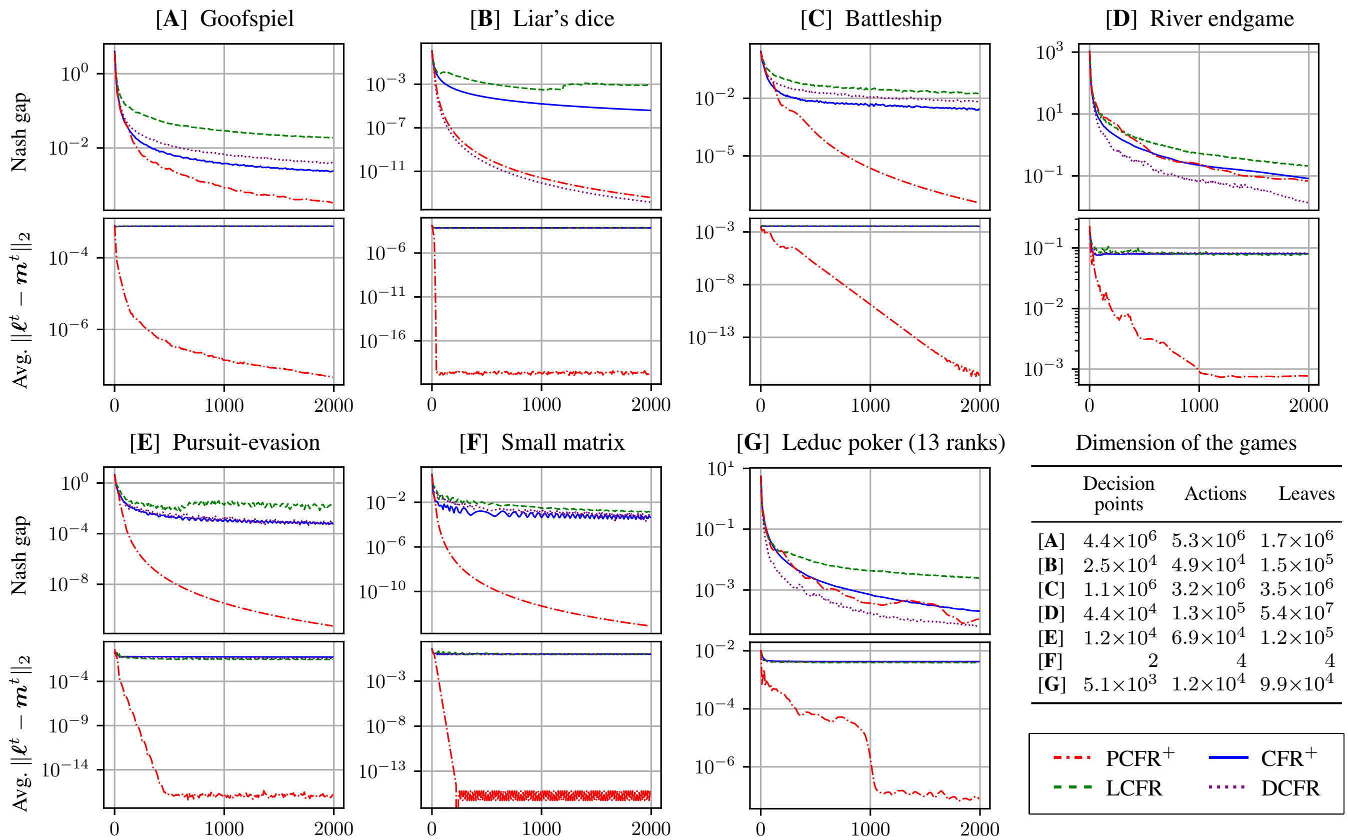
$$\begin{aligned}\max_{\hat{\boldsymbol{x}} \in \Delta^n} R^T(\hat{\boldsymbol{x}}) &\leq \inf_{\eta > 0} \left\{\frac{1}{2\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^t\|_2^2\right\} \\ &= \sqrt{2}\left(\sum_{t=1}^{T} \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^2\|_2^2\right)^{1/2}.\end{aligned}$$

7 Experiments

We conduct experiments on solving two-player zero-sum games. As mentioned previously, for EFGs the CFR framework is used for decomposing regrets into local regret minimization problems at each simplex corresponding to a decision point in the game (Zinkevich et al. 2007; Farina, Kroer, and Sandholm 2019a), and we do the same. However, as the regret minimizer for each local decision point, we use PRM$^+$ instead of RM. In addition, we apply two heuristics that usually lead to better practical performance: we use quadratic averaging of the strategy iterates, that is, we average the sequence-form strategies $\boldsymbol{x}^1, \dots, \boldsymbol{x}^T$ using the formula $\frac{6}{T(T+1)(2T+1)} \sum_{t=1}^{T} t^2 \boldsymbol{x}^t$, and we use the *alternating updates* scheme. We call this algorithm PCFR$^+$. We compare PCFR$^+$ to the prior state-of-the-art CFR variants: CFR$^+$ (Tammelin 2014), *Discounted CFR (DCFR)* with its recommended parameters (Brown and Sandholm 2019a), and *Linear CFR (LCFR)* (Brown and Sandholm 2019a).

We conduct the experiments on common benchmark games. We show results on seven games in the main body of the paper. An additional 11 games are shown in the appendix. The experiments shown in the main body are representative of those in the appendix. A description of all the games is in Appendix G, and the results are shown in Figure 3. The x-axis shows the number of iterations of each algorithm. Every algorithm pays almost exactly the same cost per iteration, since the predictions require only one additional thresholding step in PCFR$^+$. For each game, the top plot shows on the y-axis the Nash gap, while the bottom plot shows the accuracy in our predictions of the regret vector, measured as the average ℓ_2 norm of the difference between the actual loss $\boldsymbol{\ell}^t$ received and its prediction $\boldsymbol{m}^t$ across all regret minimizers at all decision points in the game. For all non-predictive algorithms (CFR$^+$, LCFR, and DCFR), we let $\boldsymbol{m}^t = \mathbf{0}$. For our predictive algorithm, we set $\boldsymbol{m}^t = \boldsymbol{\ell}^{t-1}$ at all times $t \geq 2$ and $\boldsymbol{m}^1 = \mathbf{0}$. Both y-axes are in log scale. On Battleship and Pursuit-evasion, PCFR$^+$ is faster than the other algorithms by 3-6 orders of magnitude already after 500 iterations, and around 10 orders of magnitude after 2000 iterations. On Goofspiel, PCFR$^+$ is also significantly faster than the other algorithms, by 0.5-1 order of magnitude. Finally, in the River endgame, our only poker experiment here, PCFR$^+$ is slightly faster than CFR$^+$, but slower than DCFR. Finally, PRM$^+$ converges very rapidly on the *smallmatrix* game, a 2-by-2 matrix game where CFR$^+$ and other RM-based methods converge at a rate slower than T^{-1} Farina, Kroer, and Sandholm (2019b). Across all non-poker games in the appendix, we also find that PCFR$^+$ beats the other algorithms, often by several orders of magnitude. We conclude that PCFR$^+$ seems to be the fastest method for solving non-poker EFGs. The only exception to the non-poker-game empirical rule is Liar's Dice (game [**B**]), where our predictive method performs comparably to DCFR. In the appendix, we also test CFR$^+$ with quadratic averaging (as opposed to the linear averaging that CFR$^+$ normally uses). This does not change any of our conclusions, except that for Liar's Dice, CFR$^+$ performs comparably to DCFR and PCFR$^+$ when using quadratic averaging (in fact, quadratic averaging hurts CFR$^+$ in every game except poker and Liar's Dice).

We tested on three poker games, the River endgame shown here (which is a real endgame encountered by the *Libratus* AI (Brown and Sandholm 2017) in the man-machine "Brains vs. Artificial Intelligence: Upping the Ante" competition), as well as Kuhn and Leduc poker in the appendix.



Dimension of the games

	Decision points	Actions	Leaves
[A]	4.4×10^6	5.3×10^6	1.7×10^6
[B]	2.5×10^4	4.9×10^4	1.5×10^5
[C]	1.1×10^6	3.2×10^6	3.5×10^6
[D]	4.4×10^4	1.3×10^5	5.4×10^7
[E]	1.2×10^4	6.9×10^4	1.2×10^5
[F]	2	4	4
[G]	5.1×10^3	1.2×10^4	9.9×10^4

Figure 3: Performance of PCFR$^+$, CFR$^+$, DCFR, and LCFR on five EFGs. In all plots, the x axis is the number of iterations of each algorithm. For each game, the top plot shows that the Nash gap on the y axis (on a log scale), the bottom plot shows and the average prediction error (on a log scale).

On Kuhn poker, PCFR$^+$ is extremely fast and the fastest of the algorithms. That game is known to be significantly easier than deeper EFGs for predictive algorithms (Farina, Kroer, and Sandholm 2019b). On Leduc poker as well as the River endgame, the predictions in PCFR$^+$ do not seem to help as much as in other games. On the River endgame, the performance is essentially the same as that of CFR$^+$. On Leduc poker, it leads to a small speedup over CFR$^+$. On both of those games, DCFR is fastest. In contrast, DCFR actually performs worse than CFR$^+$ in our non-poker experiments, though it is sometimes on par with CFR$^+$. In the appendix, where we try quadratic averaging in CFR$^+$, we find that for poker games this does speed up CFR$^+$, and allows it to be slightly faster than PCFR$^+$ on the River endgame and Leduc poker. We conclude that PCFR$^+$ is much faster than CFR$^+$ and DCFR on non-poker games, whereas on poker games DCFR is the fastest.

The convergence rate of PCFR$^+$ is closely related to how good the predictions $\boldsymbol{m}^t$ of $\boldsymbol{\ell}^t$ are. On Battleship and Pursuit-evasion, the predictions become extremely accurate very rapidly, and PCFR$^+$ converges at an extremely fast rate. On Goofspiel, the predictions are fairly accurate (the error is of the order 10^{-5}) and PCFR$^+$ is still significantly faster than the other algorithms. On the River endgame, the average prediction error is of the order 10^{-3}, and PCFR$^+$ performs on par with CFR$^+$, and slower than DCFR. Similar trends prevail in the experiments in the appendix. Additional experimental insights are described in the appendix.

8 Conclusions and Future Research

We extended Abernethy, Bartlett, and Hazan (2011)'s reduction of Blackwell approachability to regret minimization beyond the compact setting. This extended reduction allowed us to show that FTRL applied to the decision of which halfspace to force in Blackwell approachability is equivalent to the regret matching algorithm. OMD applied to the same problem turned out to be equivalent to RM$^+$. Then, we showed that the predictive variants of FTRL and OMD yield predictive algorithms for Blackwell approachability, as well as predictive variants of RM and RM$^+$. Combining PRM$^+$ with CFR, we introduced the PCFR$^+$ algorithm for solving EFGs. Experiments across many common benchmark games showed that PCFR$^+$ outperforms the prior state-of-the-art algorithms on non-poker games by orders of magnitude.

This work also opens future directions. Can PRM$^+$ guarantee T^{-1} convergence on matrix games like optimistic FTRL and OMD, or do the less stable updates prevent that? Can one develop a predictive variant of DCFR, which is faster on poker domains? Can one combine DCFR and PCFR$^+$, so DCFR would be faster initially but PCFR$^+$ would overtake? If the cross-over point could be approximated, this might yield a best-of-both-worlds algorithm.

9 Broader Impact

In this paper, we contributed several theoretical and algorithmic results. The most direct impact is practical advancements in equilibrium computation: in most cases, the regret minimizers we introduce converge to equilibrium in extensive-form imperfect-information games faster than the prior state of the art.

The downstream applications of our results are hard to predict. For one, our results could be used to compute strong, game-theoretic strategies in strategic interactions between rational agents. If all agents in the interaction have comparable access to equilibrium computation technology, this can result in improved social welfare and economic efficiency. On the other hand, if only a small subset of agents have access to technology that is able to compute strong strategies, those strategies could be used to maximally exploit agents that do not have access to such technology. This risk is especially true in zero-sum interactions, where a gain in value for one agent has to be compensated by a loss in value from one (or more) of the other agents.

We believed that publishing this paper and disseminating fast algorithms for equilibrium computation is a first step towards mitigating the risk of unequal access to equilibrium-finding technology.

References

Abernethy, J.; Bartlett, P. L.; and Hazan, E. 2011. Blackwell Approachability and No-Regret Learning are Equivalent. In *COLT*, 27–46.

Blackwell, D. 1954. Controlled random walks. In *Proceedings of the international congress of mathematicians*, volume 3, 336–338.

Blackwell, D. 1956. An analog of the minmax theorem for vector payoffs. *Pacific Journal of Mathematics* 6: 1–8.

Bowling, M.; Burch, N.; Johanson, M.; and Tammelin, O. 2015. Heads-up Limit Hold'em Poker is Solved. *Science* 347(6218).

Brown, N.; Kroer, C.; and Sandholm, T. 2017. Dynamic Thresholding and Pruning for Regret Minimization. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Brown, N.; and Sandholm, T. 2017. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* eaao1733.

Brown, N.; and Sandholm, T. 2019a. Solving imperfect-information games via discounted regret minimization. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Brown, N.; and Sandholm, T. 2019b. Superhuman AI for multiplayer poker. *Science* 365(6456): 885–890.

Burch, N. 2018. Time and space: Why imperfect information games are hard .

Burch, N.; Moravcik, M.; and Schmid, M. 2019. Revisiting CFR+ and alternating updates. *Journal of Artificial Intelligence Research* 64: 429–443.

Chiang, C.-K.; Yang, T.; Lee, C.-J.; Mahdavi, M.; Lu, C.-J.; Jin, R.; and Zhu, S. 2012. Online optimization with gradual variations. In *Conference on Learning Theory*, 6–1.

Farina, G.; Kroer, C.; Brown, N.; and Sandholm, T. 2019. Stable-Predictive Optimistic Counterfactual Regret Minimization. In *International Conference on Machine Learning (ICML)*.

Farina, G.; Kroer, C.; and Sandholm, T. 2019a. Online Convex Optimization for Sequential Decision Processes and Extensive-Form Games. In *AAAI Conference on Artificial Intelligence*.

Farina, G.; Kroer, C.; and Sandholm, T. 2019b. Optimistic Regret Minimization for Extensive-Form Games via Dilated Distance-Generating Functions. In *Advances in Neural Information Processing Systems*, 5222–5232.

Farina, G.; Kroer, C.; and Sandholm, T. 2019c. Regret Circuits: Composability of Regret Minimizers. In *International Conference on Machine Learning*, 1863–1872.

Farina, G.; Kroer, C.; and Sandholm, T. 2020. Stochastic regret minimization in extensive-form games. *arXiv preprint arXiv:2002.08493* .

Foster, D. P. 1999. A proof of calibration via Blackwell's approachability theorem. *Games and Economic Behavior* 29(1-2): 73–78.

Gao, Y.; Kroer, C.; and Goldfarb, D. 2019. Increasing Iterate Averaging for Solving Saddle-Point Problems. *arXiv preprint arXiv:1903.10646* .

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

Hoda, S.; Gilpin, A.; Peña, J.; and Sandholm, T. 2010. Smoothing Techniques for Computing Nash Equilibria of Sequential Games. *Mathematics of Operations Research* 35(2).

Kroer, C.; Farina, G.; and Sandholm, T. 2018. Solving Large Sequential Games with the Excessive Gap Technique. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Kroer, C.; Waugh, K.; Kılınç-Karzan, F.; and Sandholm, T. 2020. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming* .

Moravčík, M.; Schmid, M.; Burch, N.; Lisý, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; and Bowling, M. 2017. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* .

Nesterov, Y. 2009. Primal-dual subgradient methods for convex problems. *Mathematical programming* 120(1): 221–259.

Rakhlin, A.; and Sridharan, K. 2013a. Online Learning with Predictable Sequences. In *Conference on Learning Theory*, 993–1019.

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

Shalev-Shwartz, S.; and Singer, Y. 2007. A primal-dual perspective of online learning algorithms. *Machine Learning* 69(2-3): 115–142.

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

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

Waugh, K.; and Bagnell, D. 2015. A Unified View of Large-scale Zero-sum Equilibrium Computation. In *Computer Poker and Imperfect Information Workshop at the AAAI Conference on Artificial Intelligence (AAAI)*.

Zinkevich, M.; Bowling, M.; Johanson, M.; and Piccione, C. 2007. Regret Minimization in Games with Incomplete Information. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Additional Bibliographic Remarks

1. Gordon's Lagrangian Hedging framework (Gordon 2005, 2007) partially overlaps with the construction by Abernethy, Bartlett, and Hazan (2011) that we used in the paper. It appears that Abernethy et al. were not aware of Gordon's results. We did not investigate to what extent the *predictive* point of view we adopted in the paper could apply to Gordon's result.
2. In his PhD thesis, Burch (2018) mentions an algorithm that he coins "optimistic RM$^+$". No theory is provided, and unfortunately Burch never defined the algorithm formally, so it is not clear whether his algorithm is the same as PRM$^+$ as defined in Algorithm 5 in our paper. Brown, Kroer, and Sandholm (2017) gave an interpretation of optimistic RM$^+$ by Burch that would imply it is different from PRM$^+$. We indend to check with Burch directly for the final version of this paper.

References

Gordon, G. J. 2005. No-regret algorithms for structured prediction problems. Technical report, Carnegie-Mellon University, Computer Science Department, Pittsburgh PA USA.

Gordon, G. J. 2007. No-regret algorithms for online convex programs. In *Advances in Neural Information Processing Systems*, 489–496.

A Analysis of (Predictive) FTRL

In the proof of Proposition 5 we will use the following technical lemma (see, e.g, Farina, Kroer, and Sandholm (2019)).

Lemma 2. *Let $\varphi : \mathcal{D} \to \mathbb{R}_{\geq 0}$ be a 1-strongly convex differentiable regularizer with respect to some norm $\|\cdot\|$, and let $\|\cdot\|_*$ be the dual norm to $\|\cdot\|$. Finally, let $\boldsymbol{\psi} : \mathbb{R}^n \to \mathcal{D}$ be the function*

$$\boldsymbol{\psi} : \boldsymbol{g} \mapsto \underset{\hat{\boldsymbol{x}} \in \mathcal{D}}{\arg\min} \left\{ \langle \boldsymbol{g}, \hat{\boldsymbol{x}} \rangle + \frac{1}{\eta}\varphi(\hat{\boldsymbol{x}}) \right\}.$$

Then, $\boldsymbol{\psi}$ is η-Lipschitz continuous with respect to the dual norm, in the sense that

$$\|\boldsymbol{\psi}(\boldsymbol{g}) - \boldsymbol{\psi}(\boldsymbol{g'})\| \leq \eta \, \|\boldsymbol{g} - \boldsymbol{g}'\|_* \quad \forall \boldsymbol{g}, \boldsymbol{g}' \in \mathbb{R}^n.$$

Proposition 4. *Let $\varphi : \mathcal{D} \to \mathbb{R}_{\geq 0}$ be a 1-strongly regularizer with respect to some norm $\|\cdot\|$, and let $\|\cdot\|_*$ be the dual norm to $\|\cdot\|$. For all $\hat{\boldsymbol{x}} \in \mathcal{D}$, all $\eta > 0$, and all times T, the regret cumulated by (predictive) FTRL (Algorithm 1) compared to any fixed strategy $\hat{\boldsymbol{x}} \in \mathcal{D}$ is bounded as*

$$R^T(\hat{\boldsymbol{x}}) \leq \frac{\varphi(\hat{\boldsymbol{x}})}{\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{4\eta} \sum_{t=1}^{T-1} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2. \tag{7}$$

Proof. We combine several techniques and insights from the original works of Rakhlin and Sridharan (2013) and Syrgkanis et al. (2015). Let $\boldsymbol{\psi} : \mathbb{R}^n \to \mathcal{D}$ be the function that maps

$$\boldsymbol{\psi} : \boldsymbol{g} \mapsto \underset{\hat{\boldsymbol{x}} \in \mathcal{D}}{\arg\min} \left\{ \langle \boldsymbol{g}, \hat{\boldsymbol{x}} \rangle + \frac{1}{\eta}\varphi(\hat{\boldsymbol{x}}) \right\}.$$

With that notation, at all times t, predictive FTRL outputs the decision $\boldsymbol{x}^t = \boldsymbol{\psi}(\boldsymbol{L}^{t-1} + \boldsymbol{m}^t)$, where $\boldsymbol{L}^{t-1} = \sum_{\tau=1}^{t-1} \boldsymbol{\ell}^\tau$. For the purpose of this proof, we also introduce the sequence $\boldsymbol{w}^t := \boldsymbol{\psi}(\boldsymbol{L}^t)$ for $t = 1, 2, \ldots$. For any $\hat{\boldsymbol{x}} \in \mathcal{D}$,

$$R^T(\hat{\boldsymbol{x}}) = \sum_{t=1}^{T} \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t - \hat{\boldsymbol{x}} \rangle = \underbrace{\sum_{t=1}^{T} \langle \boldsymbol{m}^t, \boldsymbol{x}^t - \boldsymbol{w}^t \rangle + \langle \boldsymbol{\ell}^t, \boldsymbol{w}^t - \hat{\boldsymbol{x}} \rangle}_{\textcircled{A}} + \underbrace{\sum_{t=1}^{T} \langle \boldsymbol{\ell}^t - \boldsymbol{m}^t, \boldsymbol{x}^t - \boldsymbol{w}^t \rangle}_{\textcircled{B}}$$

We now bound each of the three terms on the right-hand side:

Ⓐ A critical observation to bound Ⓐ is the following. Since $\boldsymbol{\psi}(\boldsymbol{g})$ is a minimizer of $\langle \boldsymbol{g}, \hat{\boldsymbol{x}} \rangle + \frac{1}{\eta}\varphi(\hat{\boldsymbol{x}})$, then by the fist-order optimality conditions,

$$\left\langle \boldsymbol{g} + \frac{1}{\eta} \nabla\varphi(\boldsymbol{\psi}(\boldsymbol{g})),\ \boldsymbol{\xi} - \boldsymbol{\psi}(\boldsymbol{g}) \right\rangle \geq 0 \quad \forall \boldsymbol{g} \in \mathbb{R}^n, \boldsymbol{\xi} \in \mathcal{D}. \tag{8}$$

Using the hypothesis on the 1-strongly convexity of φ and applying (8), for all $\boldsymbol{\xi}$ we obtain

$$\frac{1}{\eta}\varphi(\boldsymbol{\xi}) + \langle \boldsymbol{g}, \boldsymbol{\xi} \rangle \geq \frac{1}{\eta}\varphi(\boldsymbol{\psi}(\boldsymbol{g})) + \langle \boldsymbol{g}, \boldsymbol{\psi}(\boldsymbol{g}) \rangle + \left\langle \boldsymbol{g} + \frac{1}{\eta} \nabla\varphi(\boldsymbol{\psi}(\boldsymbol{g})),\ \boldsymbol{\xi} - \boldsymbol{\psi}(\boldsymbol{g}) \right\rangle + \frac{1}{2\eta}\|\boldsymbol{\xi} - \boldsymbol{\psi}(\boldsymbol{g})\|^2$$

$$\geq \frac{1}{\eta}\varphi(\boldsymbol{\psi}(\boldsymbol{g})) + \langle \boldsymbol{g}, \boldsymbol{\psi}(\boldsymbol{g})\rangle + \frac{1}{2\eta}\|\boldsymbol{\xi} - \boldsymbol{\psi}(\boldsymbol{g})\|^2. \tag{9}$$

By applying (9) to the two choices $(\boldsymbol{g}, \boldsymbol{\xi}) = (\boldsymbol{L}^{t-1}, \boldsymbol{x}^t), (\boldsymbol{L}^{t-1} + \boldsymbol{m}^t, \boldsymbol{w}^t)$, respectively, we have the two inequalities

$$\frac{1}{\eta}\varphi(\boldsymbol{x}^t) + \langle \boldsymbol{L}^{t-1}, \boldsymbol{x}^t\rangle \geq \frac{1}{\eta}\varphi(\boldsymbol{w}^{t-1}) + \langle \boldsymbol{L}^{t-1}, \boldsymbol{w}^{t-1}\rangle + \frac{1}{2\eta}\|\boldsymbol{x}^t - \boldsymbol{w}^{t-1}\|^2$$
$$\frac{1}{\eta}\varphi(\boldsymbol{w}^t) + \langle \boldsymbol{L}^{t-1} + \boldsymbol{m}^t, \boldsymbol{w}^t\rangle \geq \frac{1}{\eta}\varphi(\boldsymbol{x}^t) + \langle \boldsymbol{L}^{t-1} + \boldsymbol{m}^t, \boldsymbol{x}^t\rangle + \frac{1}{2\eta}\|\boldsymbol{w}^t - \boldsymbol{x}^t\|^2.$$

Summing the two above inequalities and rearranging terms yields

$$\langle \boldsymbol{m}^t, \boldsymbol{x}^t - \boldsymbol{w}^t\rangle \leq \frac{1}{\eta}(\varphi(\boldsymbol{w}^t) - \varphi(\boldsymbol{w}^{t-1})) + \langle \boldsymbol{L}^{t-1}, \boldsymbol{w}^t - \boldsymbol{w}^{t-1}\rangle - \frac{1}{2\eta}\Big(\|\boldsymbol{x}^t - \boldsymbol{w}^{t-1}\|^2 + \|\boldsymbol{w}^t - \boldsymbol{x}^t\|^2\Big).$$

Summing over $t = 1, \dots, T$ and simplifying telescopic terms,

$$\begin{aligned}
\sum_{t=1}^{T}\langle \boldsymbol{m}^t, \boldsymbol{x}^t - \boldsymbol{w}^t\rangle &\leq \frac{1}{\eta}(\varphi(\boldsymbol{w}^T) - \varphi(\boldsymbol{w}^0)) + \sum_{t=1}^{T}\langle \boldsymbol{L}^{t-1}, \boldsymbol{w}^t - \boldsymbol{w}^{t-1}\rangle - \sum_{t=1}^{T}\frac{1}{2\eta}\Big(\|\boldsymbol{x}^t - \boldsymbol{w}^{t-1}\|^2 + \|\boldsymbol{w}^t - \boldsymbol{x}^t\|^2\Big) \\
&\leq \frac{1}{\eta}(\varphi(\boldsymbol{w}^T) - \varphi(\boldsymbol{w}^0)) + \sum_{t=1}^{T}\langle \boldsymbol{L}^{t-1}, \boldsymbol{w}^t - \boldsymbol{w}^{t-1}\rangle - \sum_{t=1}^{T-1}\frac{1}{2\eta}\Big(\|\boldsymbol{x}^{t+1} - \boldsymbol{w}^t\|^2 + \|\boldsymbol{w}^t - \boldsymbol{x}^t\|^2\Big) \\
&\leq \frac{1}{\eta}(\varphi(\boldsymbol{w}^T) - \varphi(\boldsymbol{w}^0)) + \sum_{t=1}^{T}\langle \boldsymbol{L}^{t-1}, \boldsymbol{w}^t - \boldsymbol{w}^{t-1}\rangle - \sum_{t=1}^{T-1}\frac{1}{4\eta}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2,
\end{aligned}$$

where the second inequality follows by removing a term from the last parenthesis and rearranging, and the third from the parallelogram inequality $\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 \geq \frac{1}{2}\|\boldsymbol{a} + \boldsymbol{b}\|^2$ valid for all choices of vectors $\boldsymbol{a}, \boldsymbol{b}$ and norm $\|\cdot\|$.
In order to recognize $\textcircled{A}$ on the left-hand side, we add the quantity $\sum_{t=1}^{T}\langle \boldsymbol{\ell}^t, \boldsymbol{w}^t - \hat{\boldsymbol{x}}\rangle$ on both sides, and obtain

$$\begin{aligned}
\textcircled{A} &\leq \frac{1}{\eta}(\varphi(\boldsymbol{w}^T) - \varphi(\boldsymbol{w}^0)) + \sum_{t=1}^{T}\Big(\langle \boldsymbol{\ell}^t, \boldsymbol{w}^t - \hat{\boldsymbol{x}}\rangle + \langle \boldsymbol{L}^{t-1}, \boldsymbol{w}^t - \boldsymbol{w}^{t-1}\rangle\Big) - \frac{1}{4\eta}\sum_{t=1}^{T-1}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 \\
&= \frac{1}{\eta}(\varphi(\boldsymbol{w}^T) - \varphi(\boldsymbol{w}^0)) + \sum_{t=1}^{T}\Big(\langle \boldsymbol{L}^t, \boldsymbol{w}^t\rangle - \langle \boldsymbol{L}^{t-1}, \boldsymbol{w}^{t-1}\rangle - \langle \boldsymbol{\ell}^t, \hat{\boldsymbol{x}}\rangle\Big) - \frac{1}{4\eta}\sum_{t=1}^{T-1}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 \\
&= \frac{1}{\eta}(\varphi(\boldsymbol{w}^T) - \varphi(\boldsymbol{w}^0)) + \langle \boldsymbol{L}^T, \boldsymbol{w}^T - \hat{\boldsymbol{x}}\rangle - \frac{1}{4\eta}\sum_{t=1}^{T-1}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2,
\end{aligned} \tag{10}$$

where we simplified the telescopic sum $\sum_{t=1}^{T}\langle \boldsymbol{L}^t, \boldsymbol{w}^t\rangle - \langle \boldsymbol{L}^{t-1}, \boldsymbol{w}^{t-1}\rangle = \langle \boldsymbol{L}^T, \boldsymbol{w}^T\rangle$ in the last step. Finally, using Equation (9) with $\boldsymbol{g} = \boldsymbol{L}^T, \boldsymbol{\xi} = \hat{\boldsymbol{x}}$, we can write

$$\frac{1}{\eta}\varphi(\hat{\boldsymbol{x}}) + \langle \boldsymbol{L}^T, \hat{\boldsymbol{x}}\rangle \geq \frac{1}{\eta}\varphi(\boldsymbol{w}^T) + \langle \boldsymbol{L}^T, \boldsymbol{w}^T\rangle \implies \frac{1}{\eta}\varphi(\boldsymbol{w}^T) + \langle \boldsymbol{L}^T, \boldsymbol{w}^T - \hat{\boldsymbol{x}}\rangle \leq \frac{1}{\eta}\varphi(\hat{\boldsymbol{x}}),$$

and substituting the last expression into (10), we obtain

$$\textcircled{A} \leq \frac{1}{\eta}(\varphi(\hat{\boldsymbol{x}}) - \varphi(\boldsymbol{w}^0)) - \sum_{t=1}^{T-1}\frac{1}{4\eta}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 \leq \frac{\varphi(\hat{\boldsymbol{x}})}{\eta} - \frac{1}{4\eta}\sum_{t=1}^{T-1}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2. \tag{11}$$

$\textcircled{B}$ By applying the generalized Cauchy-Schwarz inequality and Lemma 2,

$$\langle \boldsymbol{\ell}^t - \boldsymbol{m}^t, \boldsymbol{x}^t - \boldsymbol{w}^t\rangle \leq \|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_* \|\boldsymbol{x}^t - \boldsymbol{w}^t\| \leq \eta\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2.$$

Hence,

$$\textcircled{B} = \sum_{t=1}^{T}\langle \boldsymbol{\ell}^t - \boldsymbol{m}^t, \boldsymbol{x}^t - \boldsymbol{w}^t\rangle \leq \eta\sum_{t=1}^{T}\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2. \tag{12}$$

Finally, summing the bounds for $\textcircled{A}$ (11) and for $\textcircled{B}$ (12), we obtain the statement. □

B Analysis of (Predictive) OMD

In the proof of Proposition 5 we will use the two following technical lemmas.

Lemma 3. *For any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ and $\rho > 0$, it holds that $\langle \boldsymbol{a}, \boldsymbol{b}\rangle \le \frac{\rho}{2}\|\boldsymbol{a}\|_*^2 + \frac{1}{2\rho}\|\boldsymbol{b}\|^2$.*

Proof. By the arithmetic mean-geometric mean inequality, we have

$$\frac{\rho}{2}\|\boldsymbol{a}\|_*^2 + \frac{1}{2\rho}\|\boldsymbol{b}\|^2 = \frac{1}{2}\left(\rho\|\boldsymbol{a}\|_*^2 + \frac{1}{\rho}\|\boldsymbol{b}\|^2\right) \ge \sqrt{\|\boldsymbol{a}\|_*^2 \cdot \|\boldsymbol{b}\|^2} = \|\boldsymbol{a}\|_* \cdot \|\boldsymbol{b}\| \ge \langle \boldsymbol{a}, \boldsymbol{b}\rangle,$$

where we used the generalized Cauchy-Schwarz inequality in the last step. □

Lemma 4. *Let $\mathcal{D} \subseteq \mathbb{R}^d$ be closed and convex, let $\boldsymbol{g} \in \mathbb{R}^n, \boldsymbol{c} \in \mathcal{D}$, and Let $\varphi : \mathcal{D} \to \mathbb{R}_{\ge 0}$ be a 1-strongly convex differentiable regularizer with respect to some norm $\|\cdot\|$, and let $\|\cdot\|_*$ be the dual norm to $\|\cdot\|$. Then,*

$$\boldsymbol{a}^* := \operatorname*{arg\,min}_{\hat{\boldsymbol{a}}\in\mathcal{D}} \left\{ \langle \boldsymbol{g}, \hat{\boldsymbol{a}}\rangle + \frac{1}{\eta} D_\varphi(\hat{\boldsymbol{a}} \,\|\, \boldsymbol{c}) \right\}$$

is well defined (that is, the minimizer exists and is unique), and for all $\hat{\boldsymbol{a}} \in \mathcal{D}$ satisfies the inequality

$$\langle \boldsymbol{g}, \boldsymbol{a}^* - \hat{\boldsymbol{a}}\rangle \le \frac{1}{\eta}\Big(D_\varphi(\hat{\boldsymbol{a}} \,\|\, \boldsymbol{c}) - D_\varphi(\hat{\boldsymbol{a}} \,\|\, \boldsymbol{a}^*) - D_\varphi(\boldsymbol{a}^* \,\|\, \boldsymbol{c})\Big).$$

Proof. The necessary first-order optimality conditions for the argmin problem in the statement is

$$\left\langle \nabla_{\boldsymbol{a}}\left[\langle \boldsymbol{g}, \boldsymbol{a}\rangle + \frac{1}{\eta}D_\varphi(\boldsymbol{a} \,\|\, \boldsymbol{c})\right](\boldsymbol{a}^*), \boldsymbol{a}^* - \hat{\boldsymbol{a}} \right\rangle \ge 0 \quad \forall\, \hat{\boldsymbol{a}} \in \mathcal{D}.$$

Expanding the gradient, we have that for all $\hat{\boldsymbol{a}} \in \mathcal{D}$

$$\left\langle \boldsymbol{g} - \frac{1}{\eta}\Big(\nabla\varphi(\boldsymbol{a}^*) - \nabla\varphi(\boldsymbol{c})\Big), \boldsymbol{a}^* - \hat{\boldsymbol{a}} \right\rangle \ge 0 \iff \langle \boldsymbol{g}, \boldsymbol{a}^* - \hat{\boldsymbol{a}}\rangle \le \frac{1}{\eta}\Big\langle \nabla\varphi(\boldsymbol{a}^*) - \nabla\varphi(\boldsymbol{c}), \boldsymbol{a}^* - \hat{\boldsymbol{a}}\Big\rangle.$$

Finally, noting that

$$\begin{aligned}\Big\langle \nabla\varphi(\boldsymbol{a}^*) - \nabla\varphi(\boldsymbol{c}), \boldsymbol{a}^* - \hat{\boldsymbol{a}}\Big\rangle &= \Big(\varphi(\hat{\boldsymbol{a}}) - \varphi(\boldsymbol{c}) - \langle \nabla\varphi(\boldsymbol{c}), \hat{\boldsymbol{a}} - \boldsymbol{c}\rangle\Big) \\ &\quad - \Big(\varphi(\hat{\boldsymbol{a}}) - \varphi(\boldsymbol{a}^*) - \langle \nabla\varphi(\boldsymbol{a}^*), \hat{\boldsymbol{a}} - \boldsymbol{a}^*\rangle\Big) \\ &\quad - \Big(\varphi(\boldsymbol{a}^*) - \varphi(\boldsymbol{c}) - \langle \nabla\varphi(\boldsymbol{c}), \boldsymbol{a}^* - \boldsymbol{c}\rangle\Big) \\ &= D_\varphi(\hat{\boldsymbol{a}} \,\|\, \boldsymbol{c}) - D_\varphi(\hat{\boldsymbol{a}} \,\|\, \boldsymbol{a}^*) - D_\varphi(\boldsymbol{a}^* \,\|\, \boldsymbol{c})\end{aligned}$$

yields the statement. □

Proposition 5. *Let $\varphi : \mathcal{D} \to \mathbb{R}_{\ge 0}$ be a 1-strongly convex differentiable regularizer with respect to some norm $\|\cdot\|$, and let $\|\cdot\|_*$ be the dual norm to $\|\cdot\|$. For all $\hat{\boldsymbol{x}} \in \mathcal{D}$, all $\eta > 0$, and all times T, the regret cumulated by (predictive) OMD (Algorithm 2) compared to any fixed strategy $\hat{\boldsymbol{x}} \in \mathcal{D}$ is bounded as*

$$R^T(\hat{\boldsymbol{x}}) \le \frac{D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^0)}{\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{8\eta}\sum_{t=1}^{T-1}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2. \tag{13}$$

Proof. We combine several techniques and insights from the original works of Rakhlin and Sridharan (2013) and Syrgkanis et al. (2015). For any $\hat{\boldsymbol{x}} \in \mathcal{D}$,

$$R^T(\hat{\boldsymbol{x}}) = \sum_{t=1}^{T}\langle \boldsymbol{\ell}^t, \boldsymbol{x}^t - \hat{\boldsymbol{x}}\rangle = \sum_{t=1}^{T}\Bigg(\underbrace{\langle \boldsymbol{\ell}^t - \boldsymbol{m}^t, \boldsymbol{x}^t - \boldsymbol{z}^t\rangle}_{\text{Ⓐ}} + \underbrace{\langle \boldsymbol{m}^t, \boldsymbol{x}^t - \boldsymbol{z}^t\rangle}_{\text{Ⓑ}} + \underbrace{\langle \boldsymbol{\ell}^t, \boldsymbol{z}^t - \hat{\boldsymbol{x}}\rangle}_{\text{Ⓒ}}\Bigg)$$

We now bound each of the three terms on the right-hand side:

Ⓐ We use Lemma 3 with $\rho = 2\eta$ to bound the first term:

$$\langle \boldsymbol{\ell}^t - \boldsymbol{m}^t, \boldsymbol{x}^t - \boldsymbol{z}^t\rangle \le \eta\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 + \frac{1}{4\eta}\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2.$$

Ⓑ Ⓒ In order to bound these terms, we use Lemma 4:

$$\langle \boldsymbol{m}^t, \boldsymbol{x}^t - \boldsymbol{z}^t \rangle \le \frac{1}{\eta}\Big(D_\varphi(\boldsymbol{z}^t \,\|\, \boldsymbol{z}^{t-1}) - D_\varphi(\boldsymbol{z}^t \,\|\, \boldsymbol{x}^t) - D_\varphi(\boldsymbol{x}^t \,\|\, \boldsymbol{z}^{t-1})\Big)$$
$$\langle \boldsymbol{\ell}^t, \boldsymbol{z}^t - \hat{\boldsymbol{x}} \rangle \le \frac{1}{\eta}\Big(D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^{t-1}) - D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^t) - D_\varphi(\boldsymbol{z}^t \,\|\, \boldsymbol{z}^{t-1})\Big)$$

Hence, combining all bounds, we have that for any $\hat{\boldsymbol{x}} \in \mathcal{D}$,

$$\begin{aligned}
R^T(\hat{\boldsymbol{x}}) &\le \sum_{t=1}^{T} \Bigg(\eta\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 + \frac{1}{4\eta}\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 \\
&\qquad\qquad + \frac{1}{\eta}\Big(D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^{t-1}) - D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^t) - D_\varphi(\boldsymbol{z}^t \,\|\, \boldsymbol{x}^t) - D_\varphi(\boldsymbol{x}^t \,\|\, \boldsymbol{z}^{t-1})\Big)\Bigg) \\
&\le \sum_{t=1}^{T} \Bigg(\eta\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 + \frac{1}{4\eta}\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 + \frac{1}{\eta}\Big(D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^{t-1}) - D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^t)\Big) \\
&\qquad\qquad - \frac{1}{2\eta}\Big(\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 + \|\boldsymbol{x}^t - \boldsymbol{z}^{t-1}\|^2\Big)\Bigg) \\
&= \sum_{t=1}^{T} \Bigg(\eta\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{4\eta}\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 - \frac{1}{2\eta}\|\boldsymbol{x}^t - \boldsymbol{z}^{t-1}\|^2 + \frac{1}{\eta}\Big(D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^{t-1}) - D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^t)\Big)\Bigg) \\
&\le \sum_{t=1}^{T} \Bigg(\eta\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{4\eta}\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 - \frac{1}{4\eta}\|\boldsymbol{x}^t - \boldsymbol{z}^{t-1}\|^2 + \frac{1}{\eta}\Big(D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^{t-1}) - D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^t)\Big)\Bigg)
\end{aligned}$$

where we used the fact that $D_\varphi(\boldsymbol{a} \,\|\, \boldsymbol{b}) \ge \frac{1}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2$ for all $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{D}$ (because φ is 1-strongly convex by hypothesis) in the second inequality. Since the differences of divergences on the right-hand side are telescopic, we further obtain

$$\begin{aligned}
R^T(\hat{\boldsymbol{x}}) &\le \frac{D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^0) - D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^t)}{\eta} + \eta\sum_{t=1}^{T}\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{4\eta}\sum_{t=1}^{T}\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 - \frac{1}{4\eta}\sum_{t=1}^{T}\|\boldsymbol{x}^t - \boldsymbol{z}^{t-1}\|^2 \\
&\le \frac{D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^0)}{\eta} + \eta\sum_{t=1}^{T}\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{4\eta}\sum_{t=1}^{T}\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 - \frac{1}{4\eta}\sum_{t=1}^{T}\|\boldsymbol{x}^t - \boldsymbol{z}^{t-1}\|^2 \\
&= \frac{D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^0)}{\eta} + \eta\sum_{t=1}^{T}\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{4\eta}\sum_{t=1}^{T}\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 - \frac{1}{4\eta}\sum_{t=0}^{T-1}\|\boldsymbol{x}^{t+1} - \boldsymbol{z}^t\|^2 \\
&\le \frac{D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^0)}{\eta} + \eta\sum_{t=1}^{T}\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{4\eta}\sum_{t=1}^{T-1}\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 - \frac{1}{4\eta}\sum_{t=1}^{T-1}\|\boldsymbol{x}^{t+1} - \boldsymbol{z}^t\|^2 \\
&= \frac{D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^0)}{\eta} + \eta\sum_{t=1}^{T}\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{4\eta}\sum_{t=1}^{T-1}\Big(\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 + \|\boldsymbol{x}^{t+1} - \boldsymbol{z}^t\|^2\Big),
\end{aligned}$$

where we used the nonnegativity of divergences in the second inequality, and some trivial manipulation of summation indices in the later steps. Finally, we use the triangle inequality for the norm $\|\cdot\|$ to conclude that at all $t = 1, \ldots, T-1$

$$\|\boldsymbol{x}^t - \boldsymbol{z}^t\|^2 + \|\boldsymbol{x}^{t+1} - \boldsymbol{z}^t\|^2 \ge \frac{1}{2}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2,$$

and hence for all $\hat{\boldsymbol{x}} \in \mathcal{D}$

$$R^T(\hat{\boldsymbol{x}}) \le \frac{D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^0)}{\eta} + \eta\sum_{t=1}^{T}\|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{8\eta}\sum_{t=1}^{T-1}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2.$$

□

When $\nabla\varphi(\boldsymbol{z}^0) = \boldsymbol{0}$ as in Line 1 in Algorithm 2, $D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^0) \le \varphi(\hat{\boldsymbol{x}})$ and so Proposition 5 becomes

Corollary 1. *For all $\hat{\boldsymbol{x}} \in \mathcal{D}$, all $\eta > 0$, and all times T, the regret cumulated by (predictive) OMD (Algorithm 2) compared to any fixed strategy $\hat{\boldsymbol{x}} \in \mathcal{D}$ is bounded as*

$$R^T(\hat{\boldsymbol{x}}) \le \frac{\varphi(\hat{\boldsymbol{x}})}{\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{8\eta} \sum_{t=1}^{T-1} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2. \tag{14}$$

C Online Linear Optimization to Approachability

Proposition 2. *Let $(\mathcal{X}, \mathcal{Y}, \boldsymbol{u}(\cdot,\cdot), C)$ be an approachability game, where $C \subseteq \mathbb{R}^n$ is a closed convex cone, such that each halfspace $H \supseteq C$ is approachable (Definition 1). Let $\mathcal{K} := C^\circ \cap \mathbb{B}_2^n$, where $C^\circ = \{\boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{x}, \boldsymbol{y} \rangle \le 0 \ \forall \boldsymbol{y} \in C\}$ denotes the* polar cone *to C and $\mathbb{B}_2^n := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \le 1\}$ is the unit ball. Finally, let $\mathcal{L}$ be an oracle for the OLO problem (for example, the FTRL or OMD algorithm) whose domain of decisions is any closed convex set $\mathcal{D}$, such that $\mathcal{K} \subseteq \mathcal{D} \subseteq C^\circ$. Then, at all times T, the distance between the average payoff cumulated by Algorithm 3 and the target cone C is upper bounded as*

$$\min_{\hat{\boldsymbol{s}} \in C} \left\| \hat{\boldsymbol{s}} - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) \right\|_2 = \frac{1}{T} \max_{\hat{\boldsymbol{x}} \in \mathcal{K}} R_{\mathcal{L}}^T(\hat{\boldsymbol{x}}),$$

where $R_{\mathcal{L}}^T(\hat{\boldsymbol{x}})$ is the regret cumulated by $\mathcal{L}$ up to time T compared to always playing $\hat{\boldsymbol{x}} \in \mathcal{K}$.

Proof. Let $\mathcal{K} := C^\circ \cap \mathbb{B}_2^n$. As proved by Abernethy, Bartlett, and Hazan (2011), the distance from the generic point $\boldsymbol{z}$ to the convex cone C can be computed as

$$\min_{\hat{\boldsymbol{s}} \in C} \|\hat{\boldsymbol{s}} - \boldsymbol{z}\|_2 = \max_{\hat{\boldsymbol{\theta}} \in \mathcal{K}} \langle \hat{\boldsymbol{\theta}}, \boldsymbol{z} \rangle.$$

Hence,

$$\begin{aligned}
\min_{\hat{\boldsymbol{s}} \in C} \left\| \hat{\boldsymbol{s}} - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) \right\|_2 &= \max_{\hat{\boldsymbol{\theta}} \in \mathcal{K}} \left\langle \hat{\boldsymbol{\theta}}, \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) \right\rangle \\
&= -\frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{\theta}^t, \boldsymbol{\ell}^t \rangle + \frac{1}{T} \max_{\hat{\boldsymbol{\theta}} \in \mathcal{K}} \left\{ \sum_{t=1}^{T} \langle \boldsymbol{\ell}^t, \boldsymbol{\theta}^t - \hat{\boldsymbol{\theta}} \rangle \right\} && (15) \\
&= -\frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{\theta}^t, \boldsymbol{\ell}^t \rangle + \frac{1}{T} \max_{\hat{\boldsymbol{\theta}} \in \mathcal{K}} R(\hat{\boldsymbol{\theta}}) && (16)
\end{aligned}$$

where the second step uses $\boldsymbol{\ell}^t = -\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t)$. Since $\boldsymbol{\theta}^t \in \mathcal{D} \subseteq C^\circ$, the halfspace $H^t := \{\boldsymbol{z} : \langle \boldsymbol{\theta}^t, \boldsymbol{z} \rangle \le 0\}$ contains C at all times t. Furthermore, by construction $\boldsymbol{x}^t$ forces H^t, and so $\langle \boldsymbol{\theta}^t, \boldsymbol{\ell}^t \rangle = -\langle \boldsymbol{\theta}^t, \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) \rangle \ge 0$, and therefore

$$-\frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{\theta}^t, \boldsymbol{\ell}^t \rangle \le 0. \tag{17}$$

Plugging (17) into (16) yields the statement. □

D Connections between FTRL, OMD and RM, RM$^+$

Lemma 1. *The regret $R^T(\hat{\boldsymbol{x}}) = \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t - \hat{\boldsymbol{x}} \rangle$ cumulated up to any time T by the decisions $\boldsymbol{x}^1, \dots, \boldsymbol{x}^T \in \Delta^n$ compared to any $\hat{\boldsymbol{x}} \in \Delta^n$ is related to the distance of the average Blackwell payoff from the target cone $\mathbb{R}_{\le 0}^n$ as*

$$\frac{1}{T} R^T(\hat{\boldsymbol{x}}) \le \min_{\hat{\boldsymbol{s}} \in \mathbb{R}_{\le 0}^n} \left\| \hat{\boldsymbol{s}} - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \right\|_2. \tag{4}$$

So, a strategy for the Blackwell approachability game Γ is a regret-minimizing strategy for the simplex domain Δ^n.

Proof. The regret $R^T(\hat{\boldsymbol{x}})$ cumulated by PRM and PRM$^+$ satisfies

$$\begin{aligned}
\frac{1}{T} R^T(\hat{\boldsymbol{x}}) &= \frac{1}{T} \sum_{t=1}^{T} \Big(\langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle - \langle \boldsymbol{\ell}^t, \hat{\boldsymbol{x}} \rangle \Big) = \sum_{t=1}^{T} \Big(\langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \langle \mathbf{1}, \hat{\boldsymbol{x}} \rangle - \langle \boldsymbol{\ell}^t, \hat{\boldsymbol{x}} \rangle \Big) \\
&= \left\langle \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t, \hat{\boldsymbol{x}} \right\rangle = \left\langle \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t), \hat{\boldsymbol{x}} \right\rangle
\end{aligned}$$

$$= \min_{\hat{\boldsymbol{s}}\in\mathbb{R}^n_{\le 0}} \left\langle -\hat{\boldsymbol{s}} + \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t), \hat{\boldsymbol{x}} \right\rangle, \tag{18}$$

where we used the fact that $\hat{\boldsymbol{x}} \in \Delta^n$ in the second equality, and the fact that $\min_{\hat{\boldsymbol{s}}\in\mathbb{R}^n_{\le 0}} \langle -\hat{\boldsymbol{s}}, \hat{\boldsymbol{x}}\rangle = 0$ since $\hat{\boldsymbol{x}} \ge \boldsymbol{0}$. Applying the Cauchy-Schwarz inequality to the right-hand side of (24), we obtain

$$\frac{1}{T} R^T(\hat{\boldsymbol{x}}) \le \min_{\hat{\boldsymbol{s}}\in\mathbb{R}^n_{\le 0}} \left\| -\hat{\boldsymbol{s}} + \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \right\|_2 \|\hat{\boldsymbol{x}}\|_2.$$

So, using the fact that $\|\hat{\boldsymbol{x}}\|_2 \le 1$ for any $\hat{\boldsymbol{x}} \in \Delta^n$, and applying Proposition 2,

$$\frac{1}{T} R^T(\hat{\boldsymbol{x}}) \le \min_{\hat{\boldsymbol{s}}\in\mathbb{R}^n_{\le 0}} \left\| -\hat{\boldsymbol{s}} + \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \right\|_2 = \frac{1}{T} \max_{\hat{\boldsymbol{x}}'\in\mathbb{R}^n_{\ge 0}\cap\mathbb{B}^n_2} R^T_{\mathcal{L}}(\hat{\boldsymbol{x}}'), \tag{19}$$

□

Theorem 1 (FTRL reduces to RM)**.** *For all $\eta > 0$, when Algorithm 3 is set up with $\mathcal{D} = \mathbb{R}^n_{\ge 0}$ and regret minimizer $\mathcal{L}^{\text{ftrl}}_\eta$ to play Γ, it produces the same iterates as the RM algorithm.*

Proof. Given the definition of Γ and Algorithm 3, at all times t, $\mathcal{L}^{\text{ftrl}}_\eta$ observes loss $-\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t)$, where $\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \coloneqq \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t\rangle \boldsymbol{1} - \boldsymbol{\ell}^t$ is the vector-valued payoff in Γ and measures the increase of regret at time t relative to each vertex of the simplex. For the specific choice of domain $\mathcal{D} = \mathbb{R}^n_{\ge 0}$ and regularizer $\varphi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$, the computation of the next iterate (Line 3 in non-predictive FTRL, Algorithm 1) reduces to

$$\begin{aligned}
\boldsymbol{\theta}^t &= \operatorname*{arg\,min}_{\hat{\boldsymbol{x}}\in\mathbb{R}^n_{\ge 0}} \left\{ \left\langle -\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t), \hat{\boldsymbol{x}} \right\rangle + \frac{1}{2\eta}\|\hat{\boldsymbol{x}}\|_2^2 \right\} \\
&= \operatorname*{arg\,min}_{\hat{\boldsymbol{x}}\in\mathbb{R}^n_{\ge 0}} \left\{ \left\langle -2\eta\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t), \hat{\boldsymbol{x}} \right\rangle + \|\hat{\boldsymbol{x}}\|_2^2 \right\} \\
&= \operatorname*{arg\,min}_{\hat{\boldsymbol{x}}\in\mathbb{R}^n_{\ge 0}} \left\| \hat{\boldsymbol{x}} - \eta\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \right\|_2^2 = \left[\eta\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \right]^+ = \eta \left[\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \right]^+.
\end{aligned}$$

Now, the value of $\eta > 0$ does not affect the forcing action that needs to be played on Line 3 of Algorithm 3. Indeed, whenever $\boldsymbol{\theta}^t \ne 0$, $\boldsymbol{g}(\boldsymbol{\theta}^t) = \boldsymbol{\theta}^t / \|\boldsymbol{\theta}^t\|_1$, so η cancels out in the fraction and at all t,

$$\boldsymbol{x}^t = \frac{\left[\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t)\right]^+}{\left\| \left[\sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t)\right]^+ \right\|_1}.$$

This is exactly the strategy output by RM. □

Theorem 2 (OMD reduces to RM$^+$)**.** *For all $\eta > 0$, when Algorithm 3 is set up with $\mathcal{D} = \mathbb{R}^n_{\ge 0}$ and regret minimizer $\mathcal{L}^{\text{omd}}_\eta$ to play Γ, it produces the same iterates as the RM$^+$ algorithm.*

Proof. Given the definition of Γ and Algorithm 3, at all times t, $\mathcal{L}^{\text{omd}}_\eta$ observes loss $-\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t)$, where $\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \coloneqq \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t\rangle \boldsymbol{1} - \boldsymbol{\ell}^t$ is the vector-valued payoff in Γ and measures the increase of regret at time t relative to each vertex of the simplex. In the non-predictive version of OMD $\boldsymbol{m}^t = \boldsymbol{0}$, Line 3 in Algorithm 2 is equivalent to $\arg\min D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^{t-1}) = \boldsymbol{z}^{t-1}$. Hence, for the specific choice of domain $\mathcal{D} = \mathbb{R}^n_{\ge 0}$ and regularizer $\varphi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$, the computation of the next iterate (Line 5 in non-predictive OMD, Algorithm 2) reduces to

$$\begin{aligned}
\boldsymbol{\theta}^t = \boldsymbol{z}^{t-1} &= \operatorname*{arg\,min}_{\hat{\boldsymbol{z}}\in\mathbb{R}^n_{\ge 0}} \left\{ \left\langle -\boldsymbol{u}(\boldsymbol{x}^{t-1}, \boldsymbol{\ell}^{t-1}), \hat{\boldsymbol{z}} \right\rangle + \frac{1}{\eta} D_\varphi(\hat{\boldsymbol{z}} \,\|\, \boldsymbol{z}^{t-2}) \right\} \\
&= \operatorname*{arg\,min}_{\hat{\boldsymbol{z}}\in\mathbb{R}^n_{\ge 0}} \left\{ \left\langle -\boldsymbol{u}(\boldsymbol{x}^{t-1}, \boldsymbol{\ell}^{t-1}), \hat{\boldsymbol{z}} \right\rangle + \frac{1}{2\eta}\|\hat{\boldsymbol{z}} - \boldsymbol{z}^{t-2}\|_2^2 \right\} \\
&= \operatorname*{arg\,min}_{\hat{\boldsymbol{z}}\in\mathbb{R}^n_{\ge 0}} \left\| \hat{\boldsymbol{z}} - \boldsymbol{z}^{t-2} - \eta\,\boldsymbol{u}(\boldsymbol{x}^{t-1}, \boldsymbol{\ell}^{t-1}) \right\|_2^2 = \left[\boldsymbol{z}^{t-2} + \eta\,\boldsymbol{u}(\boldsymbol{x}^{t-1}, \boldsymbol{\ell}^{t-1}) \right]^+
\end{aligned}$$

$$= \left[\boldsymbol{\theta}^{t-1} + \eta\, \boldsymbol{u}(\boldsymbol{x}^{t-1}, \boldsymbol{\ell}^{t-1})\right]^+. \tag{20}$$

Since $\boldsymbol{\theta}^1 = \boldsymbol{z}^0 = \mathbf{0}$, the only effect of the step size η is a rescaling of all iterates $\{\boldsymbol{\theta}^t\}$ by a constant. However, the forcing action $\boldsymbol{g}(\boldsymbol{\theta}^t) = \boldsymbol{\theta}^t / \|\boldsymbol{\theta}^t\|_1$ is invariant to positive rescaling of $\boldsymbol{\theta}^t$. For this reason, all choices of $\eta > 0$ result in the same iterates being output by the algorithm. So, in particular we can assume without loss of generality that $\eta = 1$ in (20), which corresponds exactly to the update step in RM$^+$. □

E Predictive Blackwell Approachability and Predictive RM, RM$^+$

Proposition 3. *Let $(\mathcal{X}, \mathcal{Y}, \boldsymbol{u}(\cdot,\cdot), S)$ be a Blackwell approachability game, where every halfspace $H \supseteq S$ is approachable (Definition 1). For all T, given predictions $\boldsymbol{v}^t$ of the payoff vectors, there exist algorithms for playing the game (that is, pick $\boldsymbol{x}^t \in \mathcal{X}$ at all t) that guarantee*

$$\min_{\hat{\boldsymbol{s}} \in S} \left\| \hat{\boldsymbol{s}} - \frac{1}{T} \sum_{t=1}^T \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) \right\|_2 \le \frac{1}{\sqrt{T}} \left(1 + \frac{2}{T} \sum_{t=1}^T \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) - \boldsymbol{v}^t\|_2^2 \right).$$

Proof. As shown by Abernethy, Bartlett, and Hazan (2011), a Blackwell approachability game with a non-conic target set can be converted to a conic target set at the cost of a factor 2 in the distance bound. Hence, we assume that S is a closed convex cone, and use the construction of Algorithm 3 instantiated with the FTRL algorithm with domain $\mathcal{D} = S^\circ$, regularizer $\varphi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$, and step size parameter $\eta > 0$. Proposition 2, along with the aforementioned factor 2 reduction from generic convex target set to conic target set, implies that

$$\begin{aligned}
\min_{\hat{\boldsymbol{s}} \in C} \left\| \hat{\boldsymbol{s}} - \frac{1}{T} \sum_{t=1}^T \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) \right\|_2 &\le \frac{2}{T} \max_{\hat{\boldsymbol{x}} \in S^\circ \cap \mathbb{B}_2^n} R^T(\hat{\boldsymbol{x}}) \\
&\le \frac{2}{T} \max_{\hat{\boldsymbol{x}} \in S^\circ \cap \mathbb{B}_2^n} \left(\frac{\|\hat{\boldsymbol{x}}\|_2^2}{2\eta} + \eta \sum_{t=1}^T \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) - \boldsymbol{v}^t\|_2^2 \right) \\
&\le \frac{2}{T} \left(\frac{1}{2\eta} + \eta \sum_{t=1}^T \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{y}^t) - \boldsymbol{v}^t\|_2^2 \right)
\end{aligned}$$

where the second inequality follows from expanding the regret bound for FTRL (Proposition 4), and the third inequality follows from the fact that $\hat{\boldsymbol{x}} \in \mathbb{B}_2^n$. Setting $\eta = \frac{1}{\sqrt{T}}$ yields the result. □

Theorem 3 (Correctness of PRM, PRM$^+$)**.** *Let $\mathcal{L}_\eta^{\text{ftrl*}}$ and $\mathcal{L}_\eta^{\text{omd*}}$ denote the predictive FTRL and predictive OMD algorithms instantiated with the same choice of regularizer and domain as in Section 5, and predictions $\boldsymbol{v}^t$ as defined above for the Blackwell approachability game Γ. For all $\eta > 0$, when Algorithm 3 is set up with $\mathcal{D} = \mathbb{R}_{\ge 0}^n$, the regret minimizer $\mathcal{L}_\eta^{\text{ftrl*}}$ (resp., $\mathcal{L}_\eta^{\text{omd*}}$) to play Γ, it produces the same iterates as the PRM (resp., PRM$^+$) algorithm. Furthermore, PRM and PRM$^+$ are regret minimizer for the domain Δ^n, and at all times T satisfy the regret bound*

$$R^T(\hat{\boldsymbol{x}}) \le \sqrt{2} \left(\sum_{t=1}^T \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^t\|_2^2 \right)^{1/2}.$$

Proof. Given the definition of Γ and Algorithm 3, at all times t, $\mathcal{L}_\eta^{\text{ftrl*}}$ and $\mathcal{L}_\eta^{\text{omd*}}$ observe loss $-\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t)$, where $\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \coloneqq \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t$ is the vector-valued payoff in Γ and measures the increase of regret at time t relative to each vertex of the simplex. Furthermore, at all t the prediction given to $\mathcal{L}_\eta^{\text{ftrl*}}$ and $\mathcal{L}_\eta^{\text{omd*}}$ is $-\boldsymbol{v}^t$ (Line 2, Algorithm 3). We now break up the analysis according to the OLO oracle used.

$\mathcal{L}_\eta^{\text{ftrl*}}$ corresponds to Predictive RM For the specific choice of domain $\mathcal{D} = \mathbb{R}_{\ge 0}^n$ and regularizer $\varphi = \|\cdot\|_2^2$, Line 3 in Algorithm 1 has the closed-form solution

$$\boldsymbol{\theta}^t = \left[-\eta \left(-\sum_{t=1}^T \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^t \right) \right]^+ = \eta \left[\sum_{t=1}^T \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) + \boldsymbol{v}^t \right]^+.$$

Since the forcing action $\boldsymbol{g}(\boldsymbol{\theta}^t) = \boldsymbol{\theta}^t / \|\boldsymbol{\theta}^t\|_1$ is invariant to positive constants, we see that the action $\boldsymbol{x}^t$ picked by Algorithm 3 (Line 3) is the same for all values of $\eta > 0$ and is computed as

$$\boldsymbol{x}^t = \frac{\left[\sum_{t=1}^T \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) + \boldsymbol{v}^t \right]^+}{\left\| \left[\sum_{t=1}^T \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) + \boldsymbol{v}^t \right]^+ \right\|_1}. \tag{21}$$

provided $\boldsymbol{\theta}^t \neq \mathbf{0}$, and is an arbitrary vector $\boldsymbol{x}^t \in \Delta^n$ otherwise, in accordance with the analysis of the approachability of halfspaces in Γ (Section 5). By using the definition of $\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) := \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t$ and $\boldsymbol{v}^t := \langle \boldsymbol{m}^t, \boldsymbol{x}^{t-1} \rangle \mathbf{1} - \boldsymbol{m}^t$, we see that at all times t the iterates produced by Line 4 in Algorithm 4 are exactly as in (21).

$\mathcal{L}_\eta^{\text{omd*}}$ corresponds to Predictive RM$^+$ For the specific choice of domain $\mathcal{D} = \mathbb{R}^n_{\geq 0}$ and regularizer $\varphi = \|\cdot\|_2^2$, as already note in the proof of Theorem 2, Line 5 in Predictive OMD (Algorithm 2) has the closed-form solution

$$\boldsymbol{z}^t = \left[\boldsymbol{z}^{t-1} + \eta\, \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t)\right]^+ \tag{22}$$

at all t. Similarly, Line 3 in Predictive OMD (Algorithm 2) has the closed-form solution

$$\boldsymbol{\theta}^t = \left[\boldsymbol{z}^{t-1} + \eta \boldsymbol{v}^t\right]^+. \tag{23}$$

Since both (22) and (23) are homogeneous in $\eta > 0$ (that is, the only effect of η is to rescale all $\boldsymbol{\theta}^t$ and $\boldsymbol{z}^t$ by the same constant) and the forcing action $\boldsymbol{g}(\boldsymbol{\theta}^t) = \boldsymbol{\theta}^t / \|\boldsymbol{\theta}^t\|_1$ for Γ is invariant to positive rescaling of $\boldsymbol{\theta}^t$, we see that Algorithm 3 outputs the same iterates no matter the choice of step size parameter $\eta > 0$. In particular, we can assume without loss of generality that $\eta = 1$. In that case, Equation (22) corresponds exactly to Line 7 in PRM$^+$ (Algorithm 5), and line Equation (23) corresponds exactly to Line 4.

Regret analysis The regret $R^T(\hat{\boldsymbol{x}})$ cumulated by PRM and PRM$^+$ satisfies

$$\begin{aligned}
\frac{1}{T} R^T(\hat{\boldsymbol{x}}) &= \frac{1}{T} \sum_{t=1}^{T} \Big(\langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle - \langle \boldsymbol{\ell}^t, \hat{\boldsymbol{x}} \rangle \Big) = \sum_{t=1}^{T} \Big(\langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \langle \mathbf{1}, \hat{\boldsymbol{x}} \rangle - \langle \boldsymbol{\ell}^t, \hat{\boldsymbol{x}} \rangle \Big) \\
&= \left\langle \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t, \hat{\boldsymbol{x}} \right\rangle = \left\langle \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t), \hat{\boldsymbol{x}} \right\rangle \\
&= \min_{\hat{\boldsymbol{s}} \in \mathbb{R}^n_{\leq 0}} \left\langle -\hat{\boldsymbol{s}} + \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t), \hat{\boldsymbol{x}} \right\rangle,
\end{aligned} \tag{24}$$

where we used the fact that $\hat{\boldsymbol{x}} \in \Delta^n$ in the second equality, and the fact that $\min_{\hat{\boldsymbol{s}} \in \mathbb{R}^n_{\leq 0}} \langle -\hat{\boldsymbol{s}}, \hat{\boldsymbol{x}} \rangle = 0$ since $\hat{\boldsymbol{x}} \geq \mathbf{0}$. Applying the Cauchy-Schwarz inequality to the right-hand side of (24), we obtain

$$\frac{1}{T} R^T(\hat{\boldsymbol{x}}) \leq \min_{\hat{\boldsymbol{s}} \in \mathbb{R}^n_{\leq 0}} \left\| -\hat{\boldsymbol{s}} + \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \right\|_2 \|\hat{\boldsymbol{x}}\|_2.$$

So, using the fact that $\|\hat{\boldsymbol{x}}\|_2 \leq 1$ for any $\hat{\boldsymbol{x}} \in \Delta^n$, and applying Proposition 2,

$$\frac{1}{T} R^T(\hat{\boldsymbol{x}}) \leq \min_{\hat{\boldsymbol{s}} \in \mathbb{R}^n_{\leq 0}} \left\| -\hat{\boldsymbol{s}} + \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) \right\|_2 = \frac{1}{T} \max_{\hat{\boldsymbol{x}}' \in \mathbb{R}^n_{\geq 0} \cap \mathbb{B}^n_2} R^T_{\mathcal{L}}(\hat{\boldsymbol{x}}'), \tag{25}$$

where $R^T_{\mathcal{L}}$ is the regret cumulated by the OLO oracle used in Algorithm 3—in our case, $\mathcal{L}_\eta^{\text{ftrl*}}$ for PRM and $\mathcal{L}_\eta^{\text{omd*}}$ for PRM$^+$. In either case ($\mathcal{L} = \mathcal{L}_\eta^{\text{ftrl*}}$ or $\mathcal{L} = \mathcal{L}_\eta^{\text{omd*}}$), Proposition 1 offers a bound on $R^T_{\mathcal{L}}(\hat{\boldsymbol{x}})$ that holds for all $\hat{\boldsymbol{x}} \in \mathcal{D} = \mathbb{R}^n_{\geq 0}$. So, in particular the bound holds for all points in $\mathcal{K} = \mathbb{R}^n_{\geq 0} \cap \mathbb{B}^n_2 \subseteq \mathcal{D}$. Consequently,

$$\max_{\hat{\boldsymbol{x}}' \in \mathbb{R}^n_{\geq 0} \cap \mathbb{B}^n_2} R^T_{\mathcal{L}}(\hat{\boldsymbol{x}}') \leq \max_{\hat{\boldsymbol{x}}' \in \mathbb{R}^n_{\geq 0} \cap \mathbb{B}^n_2} \left\{ \frac{\|\hat{\boldsymbol{x}}'\|_2^2}{2\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^t\|_2^2 \right\} \leq \frac{1}{2\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^t\|_2^2, \tag{26}$$

where we used the fact that $\hat{\boldsymbol{x}}' \in \mathbb{B}^n_2$ in the last step. Substituting (26) into (25), we have

$$R^T(\hat{\boldsymbol{x}}) \leq \frac{1}{2\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^t\|_2^2.$$

Since we have shown above that the iterates produced by the algorithm are independent of $\eta > 0$, we can minimize the right-hand side over $\eta > 0$, obtaining the bound

$$R^T(\hat{\boldsymbol{x}}) \leq \sqrt{2} \left(\sum_{t=1}^{T} \|\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) - \boldsymbol{v}^t\|_2^2 \right)^{1/2}.$$

Finally, expanding the definition of $\boldsymbol{u}(\boldsymbol{x}^t, \boldsymbol{\ell}^t) := \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t$ and $\boldsymbol{v}^t := \langle \boldsymbol{m}^t, \boldsymbol{x}^{t-1} \rangle \mathbf{1} - \boldsymbol{m}^t$, we obtain the statement. □

F Extensive-Form Games and Counterfactual Regret Minimization

An extensive-form game is a game played on a game tree. Each player in an extensive-form game faces a sequential decision process. A sequential decision process is a tree consisting of two types of nodes: *decision nodes* and *observation nodes*. We denote the set of decision nodes as $\mathcal{J}$, and the set of observation nodes with $\mathcal{K}$. At each decision node $j \in \mathcal{J}$, the agent picks an action according to a distribution $\boldsymbol{x}_j \in \Delta^{n_j}$ over the set A_j of $n_j = |A_j|$ actions available at that decision node, and the process moves to the observation node that is reached by following the edge corresponding to the selected action at j, if any. At each observation point $k \in \mathcal{K}$, the agent receives one out of n_k possible signals; the set of signals that the agent can observe is denoted as S_k. After the signal is received, the process moves to the decision node that is reached by following the edge corresponding to the signal at k.

The observation node that is reached by the agent after picking action $a \in A_j$ at decision point $j \in \mathcal{J}$ is denoted by $\rho(j, a)$. Likewise, the decision node reached by the agent after observing signal $s \in S_k$ at observation point $k \in \mathcal{K}$ is denoted by $\rho(k, s)$. The set of all observation points reachable from $j \in \mathcal{J}$ is denoted as $\mathcal{C}_j \coloneqq \{\rho(j,a) : a \in A_j\}$. Similarly, the set of all decision points reachable from $k \in \mathcal{K}$ is denoted as $\mathcal{C}_k \coloneqq \{\rho(k,s) : s \in S_k\}$. To ease the notation, sometimes we will use the notation $\mathcal{C}_{ja}$ to mean $\mathcal{C}_{\rho(j,a)}$.

Pairs $z = (j, a)$ with $j \in \mathcal{J}, a \in A_j$ for which $\rho(j, a) = \emptyset$ are called *terminal sequences* and have an associated payoff vector $(u(z), -u(z))$ (that is, we assume the game is zero sum). We denote the set of all terminal sequences (also called *leaves*) with Z.

Sequence Form for Sequential Decision Processes Given a strategy $\{\boldsymbol{x}_j\}_{j\in\mathcal{J}}$ for the player, its sequence-form representation (von Stengel 1996), denoted $\mu(\boldsymbol{x})$ is defined as the vector indexed over $\{(j, a) : j \in \mathcal{J}, a \in A_j\}$ whose entry corresponding to a generic pair (j, a) is the product of the probability of all actions on the path from the root of the decision process to (j, a). We denote the range of μ, that is the set of all possible sequence-form strategies as the $\boldsymbol{x}_j$ vary arbitrarily over $\Delta^{|A_j|}$ as Q. We call Q the sequence-form strategy space of the player.

It is well-known that a Nash equilibrium in a two-player zero-sum extensive form game can be expressed as a bilinear saddle point problem

$$\min_{\boldsymbol{q}_1\in Q_1} \max_{\boldsymbol{q}_2\in Q_2} \boldsymbol{q}_1^\top \boldsymbol{A} \boldsymbol{q}_2,$$

where Q_1 and Q_2 are the sequence-form strategy spaces of Player 1 and 2, respectively, and $\boldsymbol{A}$ is a suitable game-dependent matrix. It is also common knowledge that by letting regret minimizers for Q_1 and Q_2 play against each other, we can sole the bilinear saddle point above (e.g., Farina, Kroer, and Sandholm (2018)). So, we now focus on the task of constructing a regret minimizer for a sequence-form strategy space.

Counterfactual Regret Minimization

The counterfactual regret minimization framework (Zinkevich et al. 2007) provides a way of constructing a regret minimization for the sequence-form strategy space of a player by combining independent regret minimizers *local* to each of the player's decision points $j \in \mathcal{J}$. At each $j \in \mathcal{J}$, the corresponding regret minimizer—denoted $\mathcal{R}_j$—is responsible for selecting the strategy $\boldsymbol{x}_j^t$ at all times t.

CFR achieves its goal by setting the losses observed by the local regret minimizers in a specific way. In particular, let $\boldsymbol{\ell}^t$ be the loss at time t relative to the whole sequence-form strategy space Q of the player. Then, for each decision point $j \in \mathcal{J}$, the regret minimizer $\mathcal{R}_j$ local at j is fed the loss vector $\boldsymbol{\ell}_j^t \in \mathbb{R}^{|A_j|}$, whose entries are defined as

$$\boldsymbol{\ell}_j^t[a] \coloneqq \boldsymbol{\ell}^t[(j,a)] + \sum_{j'\in\mathcal{C}_{ja}} V_{j'}^t \tag{27}$$

for each $a \in A_j$, where

$$V_j^t \coloneqq \sum_{a\in A_j} \boldsymbol{x}_j^t[a] \left(\boldsymbol{\ell}^t[(j,a)] + \sum_{j'\in\mathcal{C}_{ja}} V_{j'}^t \right) \qquad \forall j \in \mathcal{J}. \tag{28}$$

Theorem 4 (Laminar regret decomposition, (Farina, Kroer, and Sandholm 2018))**.** *At all times T, the regret R^T cumulated by the CFR algorithm can be bounded as*

$$\max_{\hat{\boldsymbol{x}}\in Q} R^T(\hat{\boldsymbol{x}}) \le \max_{\hat{\boldsymbol{x}}\in Q} \sum_{j\in\mathcal{J}} \hat{\boldsymbol{x}}[\sigma(j)] \cdot R_j^T(\hat{\boldsymbol{x}}_j)$$

where R_j^T denotes the regret cumulated by the local regret minimizer $\mathcal{R}_j$ at decision point j.

Theorem 4 in particular implies that if all local regret minimizers $\mathcal{R}_j$ ($j \in \mathcal{J}$) guarantee $O(T^{1/2})$ regret, then so does the overall algorithm, that is $R^T(\hat{\boldsymbol{x}}) = O(T^{1/2})$ for all $\hat{\boldsymbol{x}} \in Q$.

Counterfactual Loss Predictions

We now describe the construction of the counterfactual loss predictions, starting from a generic prediction $\boldsymbol{m}^t$ for $\boldsymbol{\ell}^t$ relative to the whole sequence-form strategy space Q of the player. In order to maintain symmetry with Equation (27) and Equation (28), for each decision point $j \in \mathcal{J}$, the regret minimizer $\mathcal{R}_j$ local at j is fed the loss prediction vector $\boldsymbol{m}_j^t \in \mathbb{R}^{|A_j|}$, whose entries are defined as

$$\boldsymbol{m}_j^t[a] \coloneqq \boldsymbol{m}^t[(j,a)] + \sum_{j' \in \mathcal{C}_{ja}} W_{j'}^t$$

for each $a \in A_j$, where

$$W_j^t \coloneqq \sum_{a \in A_j} \boldsymbol{x}_j^t[a] \left(\boldsymbol{m}^t[(j,a)] + \sum_{j' \in \mathcal{C}_{ja}} W_{j'}^t \right) \qquad \forall j \in \mathcal{J}.$$

It important to observe that the counterfactual loss prediction $\boldsymbol{m}_j^t$ depends on the decisions produced at time t in the subtree rooted at j. In other words, in order to construct the prediction for what loss $\mathcal{R}_j$ will observe after producing the decision $\boldsymbol{x}_j^t$, we use the "future" decisions from the subtrees under j.

In our experiments, we always set $\boldsymbol{m}^t = \boldsymbol{\ell}^{t-1}$. This is a common choice, that in other algorithms (not ours) is known to lead to asymptotically lower regret than $O(T^{1/2})$ (Syrgkanis et al. 2015; Farina, Kroer, and Sandholm 2019,?).

G Description of the Game Instances

Kuhn poker (Games [**H**] and [**I**]) is a standard benchmark in the EFG-solving community (Kuhn 1950). In Kuhn poker, each player puts an ante worth 1 into the pot. Each player is then privately dealt one card from a deck that contains R unique cards. Then, a single round of betting then occurs, with the following dynamics. First, Player 1 decides to either check or bet 1. Then,

- If Player 1 checks Player 2 can check or raise 1.
  - If Player 2 checks a showdown occurs; if Player 2 raises Player 1 can fold or call.
    - If Player 1 folds Player 2 takes the pot; if Player 1 calls a showdown occurs.
- If Player 1 raises Player 2 can fold or call.
  - If Player 2 folds Player 1 takes the pot; if Player 2 calls a showdown occurs.

When a showdown occurs, the player with the higher card wins the pot and the game immediately ends.

We used $R = 3$ in Game [**H**] (this corresponds to the original game as introduced by Kuhn (1950)), while in Game [**I**] we used $R = 13$.

Leduc poker (Games [**G**] and [**O**] to [**Q**]) is another standard benchmark in the EFG-solving community Southey et al. (2005). The game is played with a deck of R unique cards, each of which appears exactly twice in the deck. The game is composed of two rounds. In the first round, each player places an ante of 1 in the pot and is dealt a single private card. A round of betting then takes place, with Player 1 acting first. At most two bets are allowed per player. Then, a card is is revealed face up and another round of betting takes place, with the same dynamics described above. After the two betting round, if one of the players has a pair with the public card, that player wins the pot. Otherwise, the player with the higher card wins the pot. All bets in the first round are worth 1, while all bets in the second round are 2.

We set $R = 3$ in Game [**O**], $R = 5$ in Game [**P**], $R = 9$ in Game [**Q**], and $R = 13$ in Game [**G**].

Small matrix (Game [**F**]) is a small 2×2 matrix game. Given a mixed strategy $\boldsymbol{x} = (x_1, x_2) \in \Delta^2$ for Player 1 and a mixed strategy $\boldsymbol{y} = (y_1, y_2) \in \Delta^2$ for Player 2, the payoff function for player 1 is defined as

$$u(\boldsymbol{x}, \boldsymbol{y}) \coloneqq 5x_1y_1 - x_1y_2 + x_2y_2.$$

This game was found by Farina, Kroer, and Sandholm (2019) to be a hard instance for the CFR$^+$ game.

Goofspiel (Games [**A**] and [**L**]) This is another popular benchmark game, originally proposed by Ross (1971). It is a two-player card game, employing three identical decks of k cards each whose values range from 1 to k. At the beginning of the game, each player gets dealt a full deck as their hand, and the third deck (the "prize" deck) is shuffled and put face down on the board. In each turn, the topmost card from the prize deck is revealed. Then, each player privately picks a card from their hand. This card acts as a bid to win the card that was just revealed from the prize deck. The selected cards are simultaneously revealed, and the highest one wins the prize card. If the players' played cards are equal, the prize card is split. The players' score are computed as the sum of the values of the prize cards they have won. In Game [**L**] the value of k is $k = 4$, while in Game [**A**] $k = 5$.

Limited-information Goofspiel (Games [**M**] and [**N**]) This is a variant of the Goofspiel game used by Lanctot et al. (2009). In this variant, in each turn the players do not reveal their cards. Rather, they show their cards to a fair umpire, which determines which player has played the highest card and should therefore received the prize card. In case of tie, the umpire directs the players to discard the prize card just like in the Goofspiel game. In Game [**M**] the number of cards in each deck is $k = 4$, while in Game [**N**] $k = 5$.

Pursuit-evasion (Games [**E**], [**J**], and [**K**]) is a security-inspired pursuit-evasion game played on the graph shown in Figure 4. It is a zero-sum variant of the one used by Kroer, Farina, and Sandholm (2018), and a similar search game has been considered by Bošanskỳ et al. (2014) and Bošanskỳ and Čermák (2015).

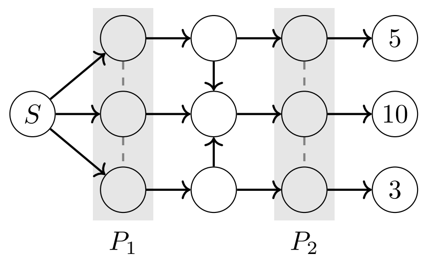


Figure 4: The graph on which the search game is played.

In each turn, the attacker and the defender act simultaneously. The defender controls two patrols, one per each respective patrol areas labeled P_1 and P_2. Each patrol can move by one step along the grey dashed lines, or stay in place. The attacker starts from the leftmost node (labeled S) and at each turn can move to any node adjacent to its current position by following the black directed edges. The attacker can also choose to wait in place for a time step in order to hide all their traces. If a patrol visits a node that was previously visited by the attacker, and the attacker did not wait to clean up their traces, they will see that the attacker was there. The goal of the attacker is to reach any of the rightmost nodes, whose corresponding payoffs are 5, 10, or 3, respectively, as indicated in Figure 4. If at any time the attacker and any patrol meet at the same node, the attacker is loses the game, which leads to a payoff of -1 for the attacker and of 1 for the defender. The game times out after m simultaneous moves, in which case both players defender receive payoffs 0. In Game [**J**] we set $m = 4$, in Game [**K**] we set $m = 5$ and in Game [**E**] we set $m = 6$.

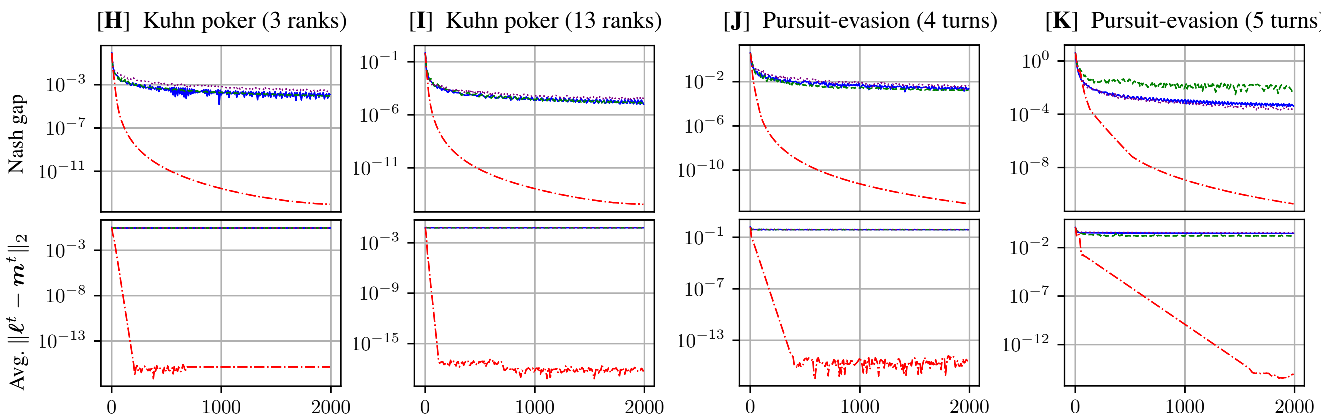
Battleship (Games [**C**] and [**R**]) is a parametric version of a classic board game, where two competing fleets take turns shooting at each other (Farina et al. 2019). At the beginning of the game, the players take turns at secretly placing a set of ships on separate grids (one for each player) of size 3×2. Each ship has size 2 (measured in terms of contiguous grid cells) and a value of 4, and must be placed so that all the cells that make up the ship are fully contained within each player's grids and do not overlap with any other ship that the player has already positioned on the grid. After all ships have been placed. the players take turns at firing at their opponent. Ships that have been hit at all their cells are considered sunk. The game continues until either one player has sunk all of the opponent's ships, or each player has completed R shots. At the end of the game, each player's payoff is calculated as the sum of the values of the opponent's ships that were sunk, minus the sum of the values of ships which that player has lost.

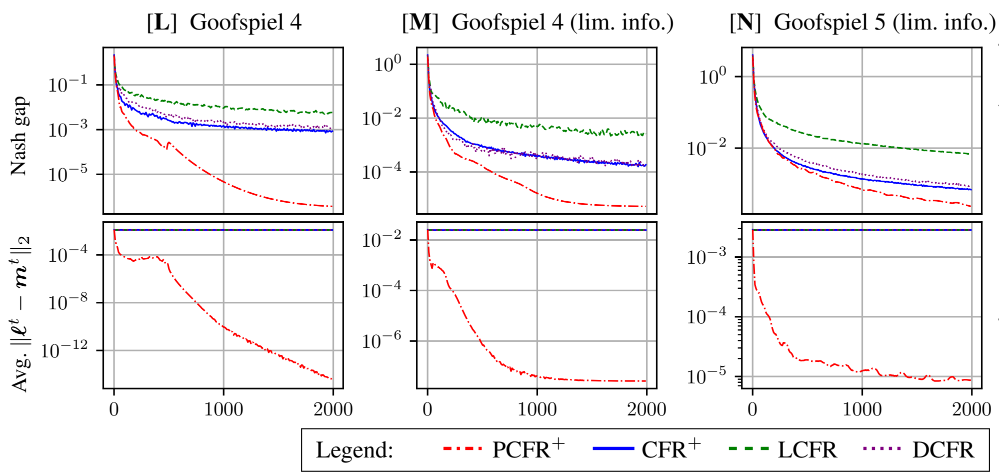
In Game [**R**] we set $R = 3$, while in Game [**C**] we set $R = 4$.

River Endgame (Game [**D**]) The river endgame is structured and parameterized as follows. The game is parameterized by the conditional distribution over hands for each player, current pot size, board state (5 cards dealt to the board), and a betting abstraction. First, Chance deals out hands to the two players according to the conditional hand distribution. Then, Libratus has the choice of folding, checking, or betting by a number of multipliers of the pot size: 0.25x, 0.5x, 1x, 2x, 4x, 8x, and all-in. If Libratus checks and the other player bets then Libratus has the choice of folding, calling (i.e. matching the bet and ending the betting), or raising by pot multipliers 0.4x, 0.7x, 1.1x, 2x, and all-in. If Libratus bets and the other player raises Libratus can fold, call, or raise by 0.4x, 0.7x, 2x, and all-in. Finally when facing subsequent raises Libratus can fold, call, or raise by 0.7x and all-in. When faced with an initial check, the opponent can fold, check, or raise by 0.5x, 0.75x, 1x, and all-in. When faced with an initial bet the opponent can fold, call, or raise by 0.7x, 1.1x, and all-in. When faced with subsequent raises the opponent can fold, call, or raise by 0.7x and all-in. The game ends whenever a player folds (the other player wins all money in the pot), calls (a showdown occurs), or both players check as their first action of the game (a showdown occurs). In a showdown the player with the better hands wins the pot. The pot is split in case of a tie. The specific endgame we use is subgame 4 from the set of open-sourced Libratus subgames at https://github.com/Sandholm-Lab/LibratusEndgames.

Liar's dice (Game [**B**]) is another standard benchmark in the EFG-solving community (Lisỳ, Lanctot, and Bowling 2015). In our instantiation, each of the two players initially privately rolls an unbiased 6-face die. The first player begins bidding, announcing any face value up to 6 and the minimum number of dice that the player believes are showing that value among the dice of both players. Then, each player has two choices during their turn: to make a higher bid, or to challenge the previous bid by declaring the previous bidder a "liar". A bid is higher than the previous one if either the face value is higher, or the number of dice is higher. If the current player challenges the previous bid, all dice are revealed. If the bid is valid, the last bidder wins and obtains a reward of $+1$ while the challenger obtains a negative payoff of -1. Otherwise, the challenger wins and gets reward $+1$, and the last bidder obtains reward of -1.

H Additional Experimental Results





Dimension of the games

	Decision points	Actions	Leaves
[H]	12	24	30
[I]	52	106	780
[J]	3.8×10^2	2.1×10^3	1.6×10^4
[K]	2.1×10^3	1.2×10^4	6.1×10^4
[L]	3.5×10^4	4.3×10^4	1.4×10^4
[M]	1.7×10^4	2.1×10^4	1.4×10^4
[N]	1.2×10^6	1.4×10^6	1.7×10^6

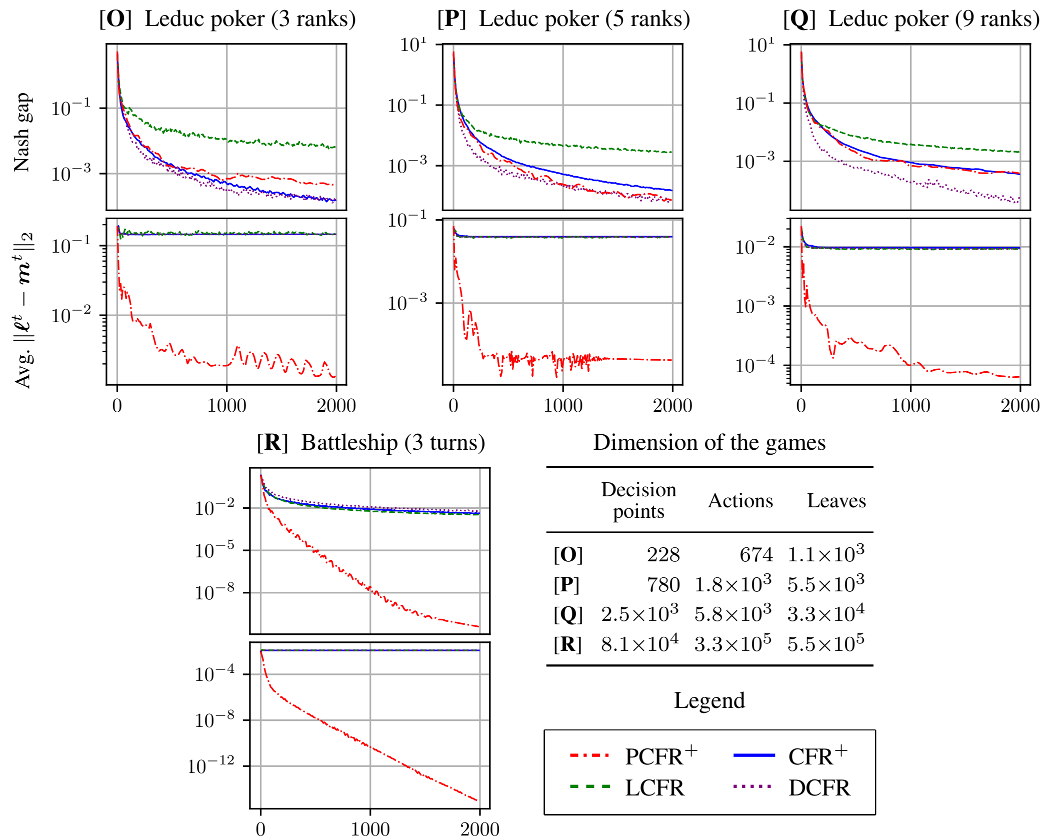
Figure 5: Performance of PCFR$^+$, CFR$^+$, DCFR, and LCFR on EFGs. In all plots, the x axis is the number of iterations of each algorithm. For each game, the top plot shows that the Nash gap on the y axis (on a log scale), the bottom plot shows and the average prediction error (on a log scale).

In all games but Leduc 13 (Game [**G**]), PCFR$^+$ significantly outperforms all other algorithms, by 2-8 orders of magnitude. In Leduc 13, PCFR$^+$ outperforms CFR$^+$ but not the DCFR algorithm. CFR$^+$ is equivalent or slightly superior to DCFR, except in Leduc 13, where it outperforms CFR$^+$ by slightly less of one order of magnitude. This is in line with the experimental results presented in the body of this paper, where we found that DCFR performs significantly better than CFR$^+$ in poker games but not other domains.

CFR$^+$, LCFR, and DCFR perform similarly in the Small matrix game (Game [**F**]), and in particular all exhibit slower than T^{-1} convergence. This is not the case for our predictive algorithm PCFR$^+$. This confirms that Small matrix is a hard instance for non-predictive methods but not for predictive methods, as already observed by Farina, Kroer, and Sandholm (2019).

In all game instances, we empirically find that the prediction error decreases quickly to extremely small values. This suggests that PCFR$^+$ might enjoy stability guarantees similar to predictive FTRL and OMD (Syrgkanis et al. 2015). Exploring such properties is an interesting future research direction.

Correlation between game structure and PCFR$^+$ performance The empirical investigation of PCFR$^+$ shows that in most classes of games PCFR$^+$ performs significantly better than CFR$^+$ and DCFR, while in other games (such as the poker games and Liar's Dice) predictivity seems to be less useful or even detrimental. It is natural to wonder what game structures can benefit from the use of predictive methods and what do not. While we do not currently have a good answer to that question, we have collected here some thoughts and observations.



	Decision points	Actions	Leaves
[O]	228	674	1.1×10^3
[P]	780	1.8×10^3	5.5×10^3
[Q]	2.5×10^3	5.8×10^3	3.3×10^4
[R]	8.1×10^4	3.3×10^5	5.5×10^5

Figure 6: Performance of PCFR$^+$, CFR$^+$, DCFR, and LCFR on EFGs. In all plots, the x axis is the number of iterations of each algorithm. For each game, the top plot shows that the Nash gap on the y axis (on a log scale), the bottom plot shows and the average prediction error (on a log scale).

- *Size*. Some predictive methods proposed in the past were found to only produce a speedup in small games, and perform worse than the state of the art in large games (Farina, Kroer, and Sandholm 2019). This is not the case for PCFR$^+$: the river endgame and Liar's Dice are not the largest games in our dataset. So, size does not seem to be a good predictor for whether predictive CFR$^+$ is beneficial over CFR$^+$ and DCFR.
- *Number of terminal states*. The river endgame and Liar's Dice both have a large ratio between the number of terminal states (leaves) and number of decision points. On the other hand, the pursuit-evasion game with 5 turns (Game [**K**]) has a significantly larger ratio than Liar's Dice but unlike in Liar's Dice, predictivity yields a speedup of more than 6 orders of magnitude on the Nash gap.
- *Private information*. Poker games and Liar's Dice have a strong private information structure: a chance node distributes independent private initial states for the two players, and each player has no information about the opponent's state. This is in contrast with, for example, the Battleship games, where each player is *not* handed a random configuration for their ships by the chance player, but rather privately picks one configuration. This shows that the "amount of private information" alone is not a good discriminator for when predictivity can be useful.
- *Private information induced by chance nodes*. From the discussion in the previous bullet, we conjecture that the way the private information arises (for example, through "dealing out cards" like in Poker games or "rolling a die" as in Liar's Dice) might affect whether predictivity helps or hurts convergence to Nash equilibrium. We leave pursuing this direction open. It is not immediately clear how one could formalize that metric.

Comparison between Linear and Quadratic Averaging in PCFR$^+$ and CFR$^+$

We also investigated the performance of CFR$^+$ with quadratic averaging in all games, as well as the performance of PCFR$^+$ with linear averaging. The experimental results are shown in Figures 7 and 9. Since only the averaging that is used when computing the (approximate) Nash equilibrium varies, but not the iterates themselves, the prediction errors are independent of the averaging variant used. Therefore, in the prediction error plots we only report one curve for each of the two algorithms.

CFR$^+$ with quadratic averaging of iterates performs similarly to CFR$^+$ with linear averaging. PCFR$^+$ with linear averaging performs similarly or slightly better than PCFR$^+$ with quadratic averaging in two games. It performs better than CFR$^+$ with either linear or quadratic averaging in 11 games, and worse than both in two games (no-limit Texas hold'em river endgame and Leduc poker). We conclude that the speedup of PCFR$^+$ is mostly due to the use of loss predictions, rather than the particular averaging of iterates.

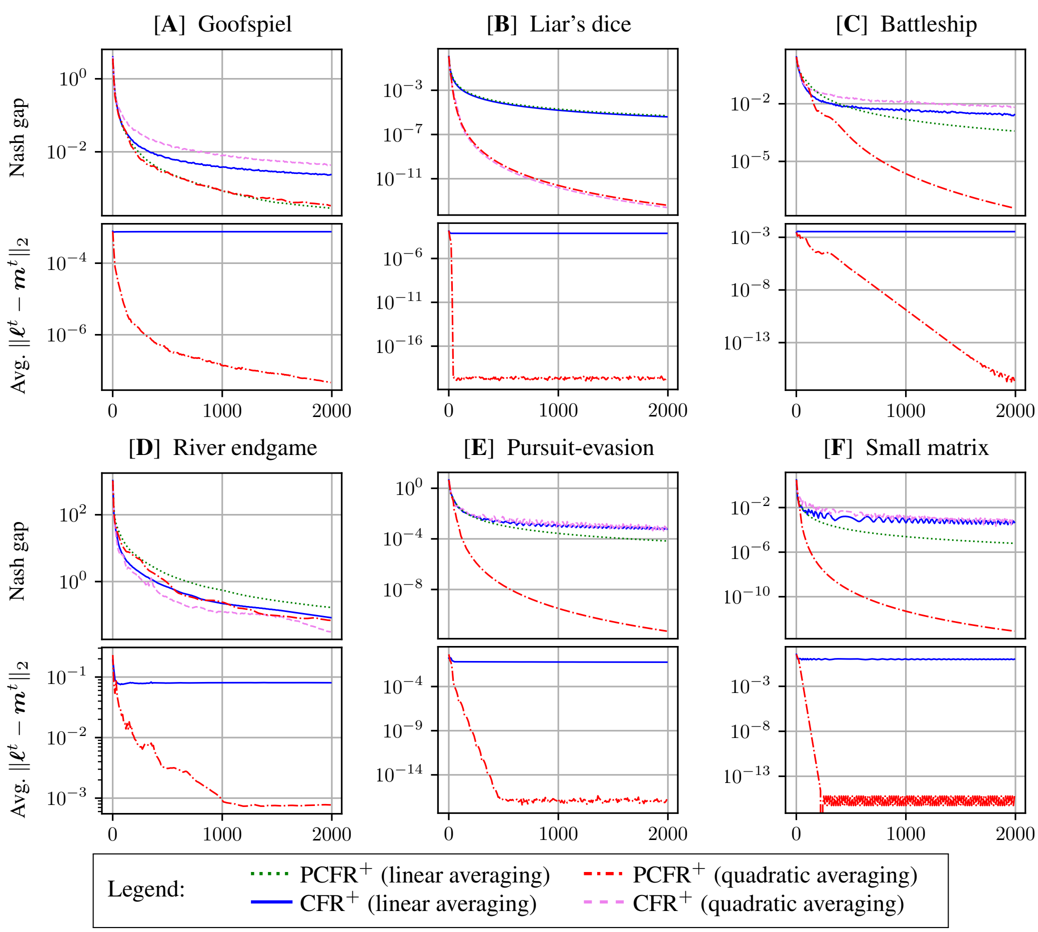


Figure 7: Performance of PCFR$^+$ and CFR$^+$ with linear and quadratic averaging on EFGs. In all plots, the x axis is the number of iterations of each algorithm. For each game, the top plot shows that the Nash gap on the y axis (on a log scale), the bottom plot shows and the average prediction error (on a log scale).

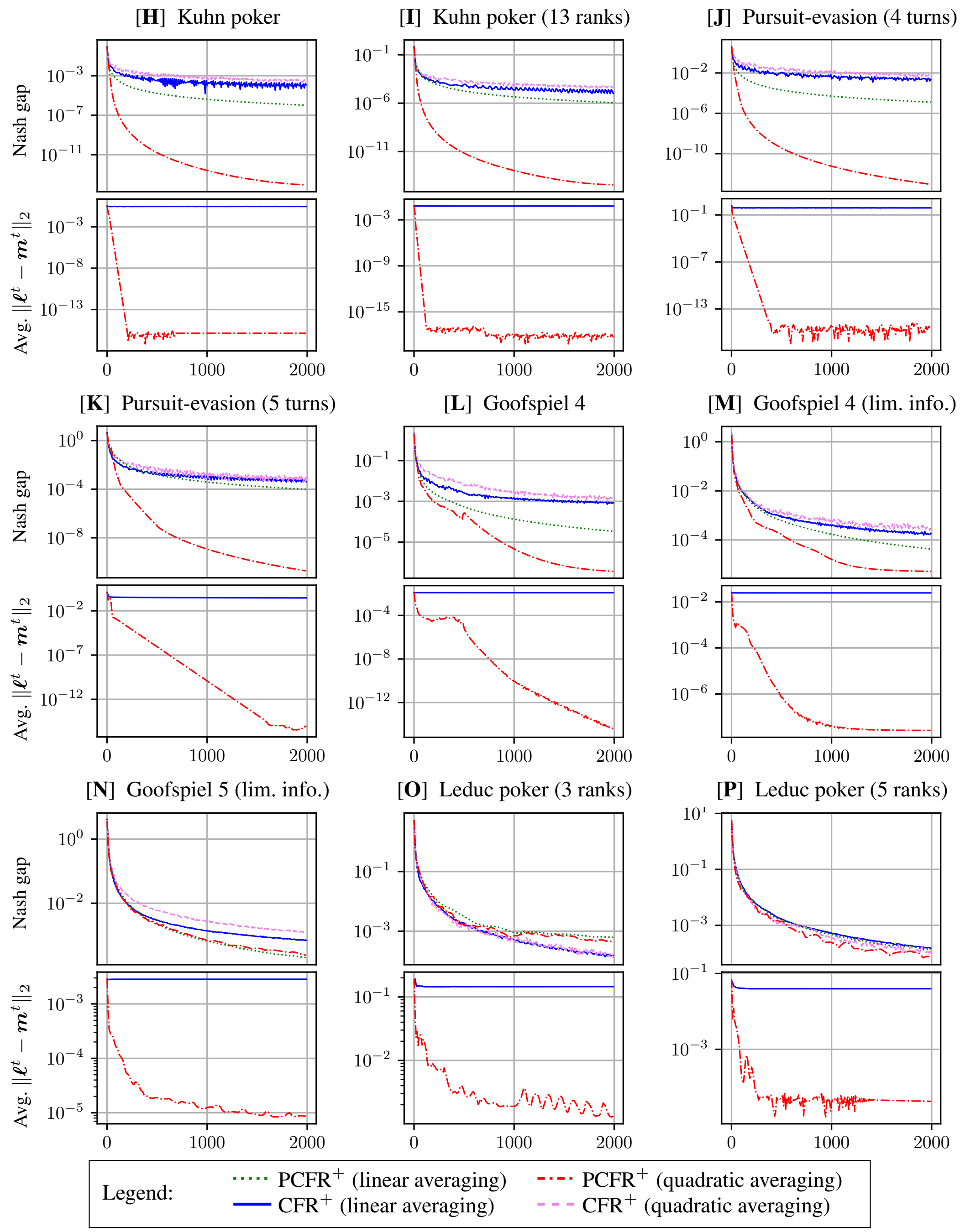


Figure 8: (continued) Performance of PCFR$^+$ and CFR$^+$ with linear and quadratic averaging on EFGs. In all plots, the x axis is the number of iterations of each algorithm. For each game, the top plot shows that the Nash gap on the y axis (on a log scale), the bottom plot shows and the average prediction error (on a log scale).

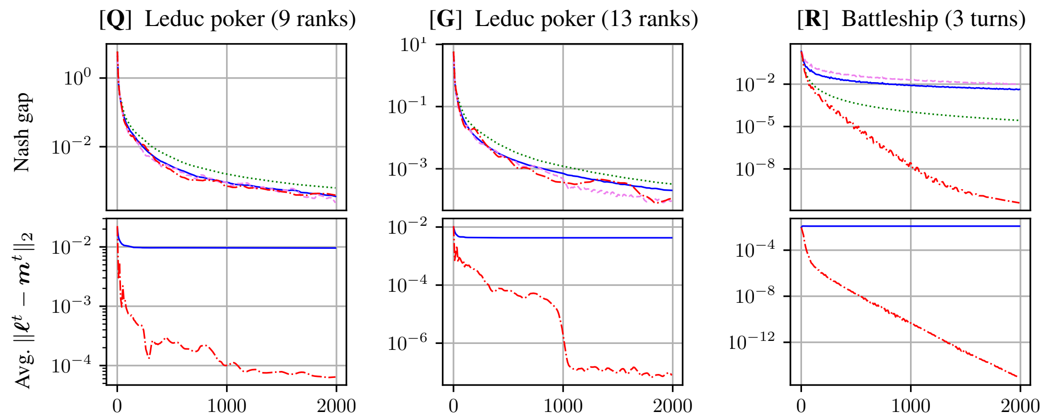


Figure 9: (continued) Performance of PCFR$^+$ and CFR$^+$ with linear and quadratic averaging on EFGs. In all plots, the x axis is the number of iterations of each algorithm. For each game, the top plot shows that the Nash gap on the y axis (on a log scale), the bottom plot shows and the average prediction error (on a log scale).

Predictive Discounted CFR and Quadratic-Average Loss Prediction

DCFR is the regret minimizer that results from applying the counterfactual regret minimization framework (Appendix F) using the *discounted regret matching* regret minimizer at each decision point. We experimentally evaluated a predictive-in-spirit[1] variant of discounted regret matching shown in Algorithm 6.

Algorithm 6: Predictive discounted regret matching

$\boldsymbol{z}^0 \leftarrow \mathbf{0} \in \mathbb{R}^n, \;\; \boldsymbol{x}^0 \leftarrow \mathbf{1}/n \in \Delta^n$

$\alpha \leftarrow 1.5, \beta \leftarrow 0$

function NEXTSTRATEGY($\boldsymbol{m}^t$)

 ▷ Set $\boldsymbol{m}^t = \mathbf{0}$ for non-predictive version

 $\boldsymbol{\theta}^t \leftarrow \frac{t^\alpha}{1+t^\alpha}[\boldsymbol{z}^{t-1}]^+ + \frac{t^\beta}{1+t^\beta}[\boldsymbol{z}^{t-1}]^- + \langle \boldsymbol{m}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{m}^t$

 if $\boldsymbol{\theta}^t \neq \mathbf{0}$ **return** $\boldsymbol{x}^t \leftarrow \boldsymbol{\theta}^t \,/\, \|\boldsymbol{\theta}^t\|_1$

 else **return** $\boldsymbol{x}^t \leftarrow$ arbitrary point in Δ^n

function OBSERVELOSS($\boldsymbol{\ell}^t$)

 $\boldsymbol{z}^t \leftarrow \frac{t^\alpha}{1+t^\alpha}[\boldsymbol{z}^{t-1}]^+ + \frac{t^\beta}{1+t^\beta}[\boldsymbol{z}^{t-1}]^- + \langle \boldsymbol{\ell}^t, \boldsymbol{x}^t \rangle \mathbf{1} - \boldsymbol{\ell}^t$

To maintain symmetry with predictive CFR and predictive CFR$^+$, we coin *predictive DCFR* the algorithm resulting from applying the counterfactual regret minimization framework (Appendix F) using the predictive discounted regret matching regret minimizer at each decision point of the game.

We also investigate the use of the quadratic average of past loss vectors,

$$\boldsymbol{m}^t = \frac{6}{t(t-1)(2t-1)} \sum_{\tau=1}^{t-1} \tau^2 \boldsymbol{\ell}^\tau,$$

as the prediction for the next loss $\boldsymbol{\ell}^t$. We call this loss prediction the "quadratic-average loss prediction".

We compare predictive DCFR (with and without quadratic-average loss prediction), PCFR$^+$ (with and without quadratic-average loss prediction), CFR$^+$, and DCFR in Figures 10 and 11.

[1] In fact, we do not have a proof that our variant is predictive in the formal sense described in the body of the paper. However, the variant we describe follows the natural pattern of predictive RM and predictive RM$^+$.

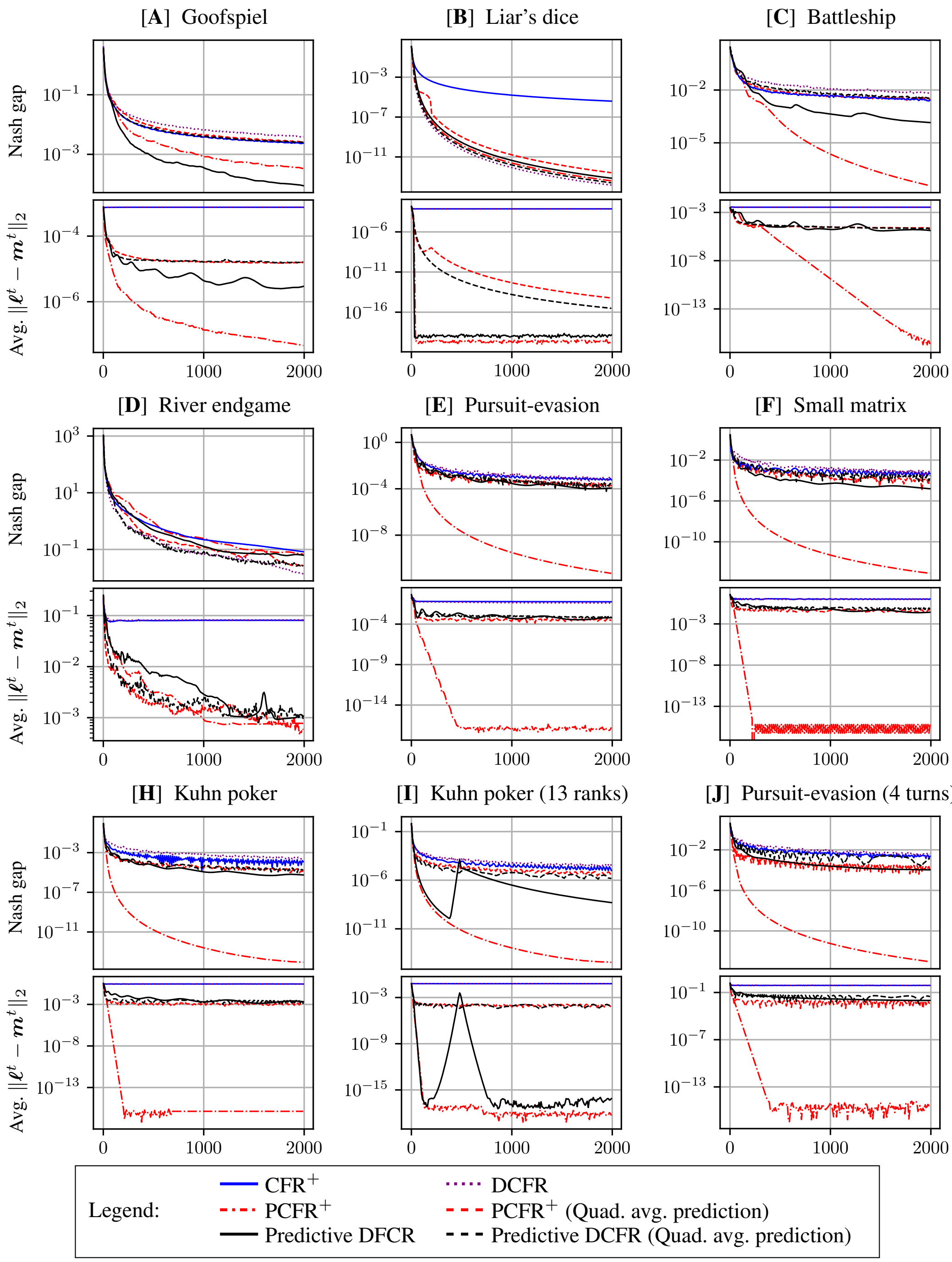


Figure 10: Comparison between of discounted CFR and CFR$^+$, with and without quadratic-average loss prediction. In all plots, the x axis is the number of iterations of each algorithm. For each game, the top plot shows that the Nash gap on the y axis (on a log scale), the bottom plot shows and the average prediction error (on a log scale).

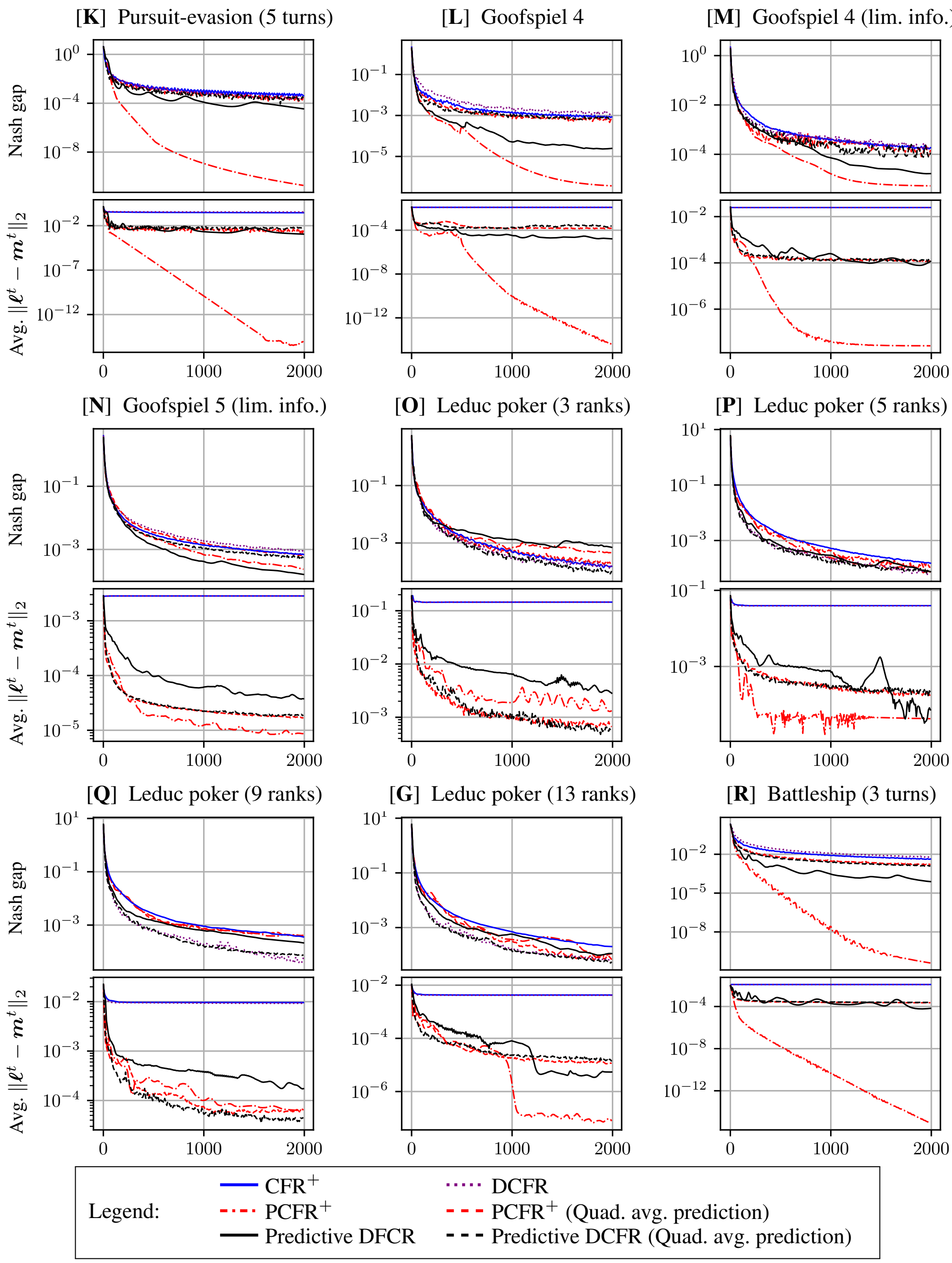


Figure 11: (continued) Comparison between of discounted CFR and CFR$^+$, with and without quadratic-average loss prediction. In all plots, the x axis is the number of iterations of each algorithm. For each game, the top plot shows that the Nash gap on the y axis (on a log scale), the bottom plot shows and the average prediction error (on a log scale).

References

Abernethy, J.; Bartlett, P. L.; and Hazan, E. 2011. Blackwell Approachability and No-Regret Learning are Equivalent. In *COLT*, 27–46.

Bošanskỳ, B.; and Čermák, J. 2015. Sequence-form algorithm for computing Stackelberg equilibria in extensive-form games. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Bošanskỳ, B.; Kiekintveld, C.; Lisý, V.; and Pěchouček, M. 2014. An Exact Double-Oracle Algorithm for Zero-Sum Extensive-Form Games with Imperfect Information. *Journal of Artificial Intelligence Research* 829–866.

Farina, G.; Kroer, C.; and Sandholm, T. 2018. Online Convex Optimization for Sequential Decision Processes and Extensive-Form Games. In *arXiv*.

Farina, G.; Kroer, C.; and Sandholm, T. 2019. Optimistic Regret Minimization for Extensive-Form Games via Dilated Distance-Generating Functions. In *Advances in Neural Information Processing Systems*, 5222–5232.

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

Kroer, C.; Farina, G.; and Sandholm, T. 2018. Robust Stackelberg Equilibria in Extensive-Form Games and Extension to Limited Lookahead. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Kuhn, H. W. 1950. A Simplified Two-Person Poker. In Kuhn, H. W.; and Tucker, A. W., eds., *Contributions to the Theory of Games*, volume 1 of *Annals of Mathematics Studies, 24*, 97–103. Princeton, New Jersey: Princeton University Press.

Lanctot, M.; Waugh, K.; Zinkevich, M.; and Bowling, M. 2009. Monte Carlo Sampling for Regret Minimization in Extensive Games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Lisỳ, V.; Lanctot, M.; and Bowling, M. 2015. Online Monte Carlo counterfactual regret minimization for search in imperfect information games. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems*, 27–36.

Rakhlin, A.; and Sridharan, K. 2013. Online Learning with Predictable Sequences. In *Conference on Learning Theory*, 993–1019.

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

Southey, F.; Bowling, M.; Larson, B.; Piccione, C.; Burch, N.; Billings, D.; and Rayner, C. 2005. Bayes' Bluff: Opponent Modelling in Poker. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI)*.

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

von Stengel, B. 1996. Efficient Computation of Behavior Strategies. *Games and Economic Behavior* 14(2): 220–246.

Zinkevich, M.; Bowling, M.; Johanson, M.; and Piccione, C. 2007. Regret Minimization in Games with Incomplete Information. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.