A Unified Approach to Reinforcement Learning, Quantal Response Equilibria, and Two-Player Zero-Sum Games

Samuel Sokota*
Carnegie Mellon University
ssokota@andrew.cmu.edu

Ryan D'Orazio*
Mila, Université de Montréal
ryan.dorazio@mila.quebec

J. Zico Kolter
Carnegie Mellon University
zkolter@cs.cmu.edu

Nicolas Loizou
Johns Hopkins University
nloizou@jhu.edu

Marc Lanctot
DeepMind
lanctot@deepmind.com

Ioannis Mitliagkas
Mila, Université de Montréal
ioannis@mila.quebec

Noam Brown
Meta AI
noambrown@fb.com

Christian Kroer
Columbia University
christian.kroer@columbia.edu

Abstract

Algorithms designed for single-agent reinforcement learning (RL) generally fail to converge to equilibria in two-player zero-sum (2p0s) games. Conversely, game-theoretic algorithms for approximating Nash and quantal response equilibria (QREs) in 2p0s games are not typically competitive for RL and can be difficult to scale. As a result, algorithms for these two cases are generally developed and evaluated separately. In this work, we show that a single algorithm—a simple extension to mirror descent with proximal regularization that we call magnetic mirror descent (MMD)—can produce strong results in both settings, despite their fundamental differences. From a theoretical standpoint, we prove that MMD converges linearly to QREs in extensive-form games—this is the first time linear convergence has been proven for a first order solver. Moreover, applied as a tabular Nash equilibrium solver via self-play, we show empirically that MMD produces results competitive with CFR in both normal-form and extensive-form games with full feedback (this is the first time that a standard RL algorithm has done so) and also that MMD empirically converges in black-box feedback settings. Furthermore, for single-agent deep RL, on a small collection of Atari and Mujoco games, we show that MMD can produce results competitive with those of PPO. Lastly, for multi-agent deep RL, we show MMD can outperform NFSP in 3x3 Abrupt Dark Hex.

1 Introduction

This work concerns approximating (regularized) optimal policies in single-agent settings and approximating Nash and logit quantal response equilibria (which are Nash equilibria of a regularized game) [41, 42] in two-player zero-sum (2p0s) games. While these problem settings are similar in certain intuitive senses—both involve control policies converging to well-defined notions of optimality—they also possess fundamental differences. For example, because of cyclical dynamics in 2p0s games (e.g., rock-paper-scissors), standard reinforcement learning (RL) algorithms generally

*Equal contribution

Preprint. Under review.

fail to converge to equilibria when applied in self-play. Conversely, convergent algorithms for 2p0s games are not generally competitive as single-agent algorithms and often involve nuances that make them difficult to scale.

Our contribution is a single algorithm that yields strong performance in both problem settings, despite their substantial differences. This algorithm, which we call magnetic mirror descent (MMD), is an extension of mirror descent [9, 46] with proximal regularization. We analyze MMD as a quantal response equilibrium (QRE) solver for extensive-form games (EFGs). We show that QRE solving in EFGs can be formulated as a variational inequality problem [16] with composite structure. By leveraging this structure we show that MMD over the sequence form [53, 66, 27] enjoys linear convergence to the reduced normal-form QRE. Previously known first order linear convergence results only exist for normal-form games (NFGs) [13]. While such algorithms can be applied to EFGs, they require first converting EFGs to NFGs, which is an exponential reduction in the size of the game. On the other hand, while there exist scalable algorithms that operate directly over the extensive-form [18, 38], they possess neither last-iterate nor linear convergence guarantees.

Our empirical contribution investigates MMD as a single-agent RL algorithm and as a 2p0s QRE and last-iterate equilibrium approximation algorithm across a variety of benchmarks. We begin by confirming our theory—showing that MMD converges exponentially fast to QREs in both NFGs and EFGs. We also find that, empirically, MMD converges to agent QREs (AQREs) when applied as a self-play RL algorithm. These results lead us to examine MMD as a Nash equilibrium solver. On this front, we show competitive performance with counterfactual regret minimization (CFR) [69] in the full feedback setting for both NFGs and EFGs. This is the first instance of a standard RL algorithm[2] yielding empirically competitive performance with CFR in tabular benchmarks when applied in self-play. We also show that MMD exhibits convergent behavior in black box settings for both NFGs and EFGs. Next, for single-agent deep RL, we show that MMD yields results that are competitive with those of PPO [55] for a small collection of Atari [10] and Mujoco [59] games. Lastly, we present multi-agent deep RL results for 3x3 Abrupt Dark Hex, where we show that MMD can outperform NFSP [23]. Combined, these results suggest that MMD is effective both as a single-agent RL algorithm and as an equilibrium approximation algorithm for 2p0s games.

2 Background

Sections 2.1 and 3.2 provide a casual treatment of our problem settings and solution concepts, and a summary of our algorithm and some of our theoretical results. Sections 2.2 through 3.1 give a more formal and detailed treatment of the same material—these sections are self-contained and safe-to-skip for readers less interested in our theoretical results.

2.1 Problem Settings and Solution Concepts

This work is concerned with both single-agent settings, such as Markov decision processes [58] and partially observable Markov decision processes [26], and 2p0s games—settings with two players in which the reward for one player is the negation of the reward for the other player. This latter setting is often formalized as an NFG, a partially observable stochastic game [21] or a perfect-recall EFG [65]. An important idea is that it is possible to convert any EFG into an equivalent NFG. The actions of the equivalent NFG correspond to the deterministic policies of the EFG. The payoffs for a joint action are dictated by the expected returns of the corresponding joint policy in the EFG.

The solution concepts studied in this work carry differing names in RL literature and game theory literature. In single-agent settings they are called optimal policies and soft-optimal policies. We say a policy is optimal if there does not exist another policy achieving a greater expected return [58]. We say a policy is soft optimal if it maximizes a weighted combination of its expected action value and its entropy: $\forall s, \pi(s) \in \arg\max_{\tilde{\pi}(s)} \mathbb{E}_{a\sim\tilde{\pi}(s)} q_\pi(s,a) + \alpha\mathcal{H}(\tilde{\pi}(s))$, where s is a state, π is a policy, q_π is the state-action value function for policy π, a is an action, α is the regularization temperature, and $\mathcal{H}$ is Shannon entropy. First, we remark that, in settings with discrete action spaces, soft optimal corresponds to playing a softmax proportional to Q-values. Second, in the limit as α goes to zero, the α-soft optimal policy (which is unique for $\alpha > 0$) approaches an optimal policy.

[2]We use "standard RL algorithm" to mean algorithms that would look ordinary to single-agent RL practitioners—excluding, e.g., algorithms that converge in the average iterate or operate over sequence form.

Both optimality and soft optimality possess analogues in 2p0s games. In 2p0s games, a joint policy is considered optimal if it is a Nash equilibrium. A joint policy is a Nash equilibrium if each player's policy is optimal, conditioned on the other player not changing its policy. Similarly, a joint policy is considered soft optimal if it is a logit QRE [41]. (As we are primarily focused on logit QREs in this work, we refer to them simply as QREs.) A joint policy is a QRE if each player's policy is soft optimal, conditioned on the other player not changing its policy. In the limit as α goes to zero, the α-QRE (which is unique for $\alpha > 0$) approaches a Nash equilibrium. Importantly, the QREs of an EFG and the QREs of the normal-form equivalent of an EFG generally do not correspond. To distinguish between them, the former is generally called an agent QRE (AQRE) [42], leaving QRE to refer to the latter.

2.2 Notation

We use superscript to denote a particular coordinate of $x = (x^1, \cdots, x^n) \in \mathbb{R}^n$ and subscript to denote time x_t. We use the standard inner product denoted as $\langle x, y \rangle = \sum_{i=1}^n x^i y^i$. For a given norm $\|\cdot\|$ on $\mathbb{R}^n$ we define its dual norm $\|y\|_* = \sup_{\|x\|=1} \langle y, x \rangle$. For example, the dual norm to $\|x\|_1 = \sum_{i=1}^n |x^i|$ is $\|x\|_\infty = \max_i |x^i|$. We assume all functions $f : \mathbb{R}^n \to (-\infty, +\infty]$ to be closed, with domain of f as $\operatorname{dom} f = \{x : f(x) < +\infty\}$, and corresponding interior $\operatorname{int} \operatorname{dom} f$. If f is convex and differentiable then its minimum x_* over a closed convex set C, $x_* \in \arg\min_{x \in C} f(x)$, is equivalent to finding x_* such that $\langle \nabla f(x_*), x - x_* \rangle \geqslant 0$ for any $x \in C$.

We also use the Bregman divergence of ψ to generalize the notion of distance. Let ψ be a convex function differentiable over $\operatorname{int} \operatorname{dom} \psi$, then the Bregman divergence with respect to ψ is $B_\psi : \operatorname{dom} \psi \times \operatorname{int} \operatorname{dom} \psi \to \mathbb{R}$, defined as $B_\psi(x; y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$. We say that f is μ-strongly convex over C with respect to $\|\cdot\|$ if $B_f(x; y) \geqslant \frac{\mu}{2} \|x - y\|^2$ for any $x \in C, y \in C \cap \operatorname{int} \operatorname{dom} \psi$. Similarly we define relative strong convexity [39]. We say g is μ-strongly convex relative to ψ over C if $\langle \nabla g(x) - \nabla g(y), x - y \rangle \geqslant \mu \langle \nabla \psi(x) - \nabla \psi(y), x - y \rangle \quad \forall x, y \in \operatorname{int} \operatorname{dom} \psi \cap C$. Note both ψ and $B_\psi(\cdot; y)$ are 1-strongly convex relative to ψ and that relative strong convexity can be equivalently stated as $B_g(x; y) \geqslant \mu B_f(x; y)$ [39].

2.3 Zero-Sum Games and QREs

In 2p0s games the solution of a QRE can be written as the solution to a negative entropy regularized saddle point problem. To model QREs (and more) we consider the regularized min max problem

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \quad \alpha g_1(x) + f(x, y) - \alpha g_2(y), \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{Y} \subset \mathbb{R}^m$ are closed, convex, and possibly unbounded, $g_1 : \mathbb{R}^n \to \mathbb{R}$ and $g_2 : \mathbb{R}^m \to \mathbb{R}$ and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. Moreover, g_1 and $f(\cdot, y)$ are differentiable and convex for every y, and similarly $-g_2$, $f(x, \cdot)$ differentiable and concave for every x. A solution (x_*, y_*) to (1) is a Nash equilibrium in the regularized game with the following best response conditions along with their equivalent first order optimality conditions

$$x_* \in \arg\min_{x \in \mathcal{X}} g_1(x) + f(x, y_*) \Leftrightarrow \langle \nabla g_1(x_*) + \nabla_x f(x_*, y_*), x - x_* \rangle \geqslant 0 \, \forall x \in \mathcal{X}, \tag{2}$$

$$y_* \in \arg\min_{y \in \mathcal{Y}} g_2(y) - f(x_*, y) \Leftrightarrow \langle \nabla g_2(y_*) - \nabla_y f(x_*, y_*), y - y_* \rangle \geqslant 0 \, \forall y \in \mathcal{Y}. \tag{3}$$

In the context of QREs we have that $\mathcal{X} = \Delta^n, \mathcal{Y} = \Delta^m$ with $f(x, y) = x^\top A y$ for some payoff matrix A, and g_1, g_2 are negative entropy. The corresponding best response conditions (2-3) can be written in closed form as $x_* \propto \exp(-A y_* / \alpha)$, $y_* \propto \exp(A^\top x_* / \alpha)$. Similarly, for EFGs, normal-form QREs take the form of (1) [37] with g_1, g_2 being dilated entropy [24], and $f(x, y) = x^\top A y$ (A being the sequence-from payoff matrix), and $\mathcal{X}, \mathcal{Y}$ the sequence form strategy spaces of both players.

2.4 Connection between zero-sum games and variational inequalities

More generally, solutions to (1) (including QREs) can be written as solutions to variational inequalities (VIs) with specific structure. The equivalent VI formulation stacks both first-order best response conditions (2-3) into one inequality.

Definition 2.1 (Variational Inequality Problem (VI))**.** Given $\mathcal{Z} \subseteq \mathbb{R}^n$ and mapping $G : \mathcal{Z} \to \mathbb{R}^n$, the variational inequality problem $\mathrm{VI}(\mathcal{Z}, G)$ is to find $z_* \in \mathcal{Z}$ such that

$$\langle G(z_*), z - z_* \rangle \geqslant 0 \quad \forall z \in \mathcal{Z}. \tag{4}$$

In particular, the optimality conditions (2-3) are equivalent to $\mathrm{VI}(\mathcal{Z}, G)$ where $G = F + \alpha\nabla g$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $g : \mathcal{Z} \to \mathbb{R}, (x, y) \mapsto g_1(x) + g_2(y)$, with corresponding operators $F(z) = [\nabla_x f(x, y), -\nabla_y f(x, y)]^\top$, and $\nabla g = [\nabla_x g_1(x), \nabla_y g_2(y)]^\top$. For more details see [16][Section 1.4.2]. Note that VIs are more general than min-max problems; they also include fixed-point problems and Nash equilibria in n-player general-sum games [16]. However, in the case of convex-concave zero-sum games and convex optimization, the problem admits efficient algorithms since the corresponding operator G is *monotone* [52].

Definition 2.2. G is said to be strongly monotone if, for $\mu > 0$, and any z, z' where G is defined $\langle G(z) - G(z'), z - z' \rangle \geqslant \mu \|z - z'\|^2$. G is monotone if this is true for $\mu = 0$.

Definition 2.3. G is said to be L-smooth with respect to $\|\cdot\|$ if for any z, z' where G is defined $\|G(z) - G(z')\|_* \leqslant L\|z - z'\|$.

For EFGs, Ling et al. [37] show that the QRE is the solution of a min max problem of the form (1) where f is bilinear and each g_i is not smooth. Therefore, we can write the problem as a VI with strongly monotone operator G having composite structure, a smooth part coming from f and non-smooth part from the regularization g_1, g_2.

Proposition 2.4. *The solution to a normal-form reduced logit QRE in a two-player zero-sum EFG is equivalent to solving* $\mathrm{VI}(\mathcal{Z}, F + \alpha\nabla\psi)$ *where* $\mathcal{Z}$ *is the cross-product of the sequence form strategy spaces, and* ψ *is the sum of the dilated entropy functions for each player. The function* ψ *is strongly convex with respect to* $\|\cdot\|$*. Furthermore,* F *is monotone and* $\max_{ij} |A_{ij}|$*-smooth (*A *being the sequence-form payoff matrix) with respect to* $\|\cdot\|$*, and* $F + \alpha\nabla g$ *is strongly monotone.*

3 Magnetic Mirror Descent

Motivated by the connection between QREs and VIs (Proposition 2.4), we now present our main algorithm, a non-Euclidean proximal gradient method to solve $\mathrm{VI}(\mathcal{Z}, F + \alpha\nabla g)$. Since ∇g is possibly not smooth we incorporate g as a proximal regularization. Doing so allows us to leverage the curvature of g to get fast convergence without assuming ∇g to be smooth.

Algorithm 3.1. *Starting with* $z_1 \in \operatorname{int}\operatorname{dom}\psi \cap \mathcal{Z}$ *at each iteration* t *do*

$$z_{t+1} = \arg\min_{z \in \mathcal{Z}} \eta \left(\langle F(z_t), z \rangle + \alpha g(z) \right) + B_\psi(z; z_t).$$

Note that we require $z_{t+1} \in \operatorname{int}\operatorname{dom}\psi$, otherwise the Bregman divergence term will be undefined at the next iteration, therefore we make the following standard assumption.

Assumption 3.2 (Well-defined)**.** Assume ψ is 1-strongly convex with respect to $\|\cdot\|$ over $\mathcal{Z}$, and for any ℓ, stepsize $\eta > 0$, and $\alpha > 0$, $z_{t+1} = \arg\min_{z \in \mathcal{Z}} \eta \left(\langle \ell, z \rangle + \alpha g(z) \right) + B_\psi(z; z_t) \in \operatorname{int}\operatorname{dom}\psi$.

Assumption 3.3. Let F be monotone and L-smooth with respect to $\|\cdot\|$ and g be 1-strongly convex relative to ψ over $\mathcal{Z}$ with g differentiable over $\operatorname{int}\operatorname{dom}\psi$.

As a consequence of the above assumptions, we have that $F + \alpha\nabla g$ is strongly monotone[3] with a unique solution z_* [6]. Assuming that $z_* \in \operatorname{int}\operatorname{dom}\psi$[4], Algorithm 3.1 converges linearly to z_*.

Theorem 3.4. *Let Assumptions 3.2 and 3.3 hold, and assume the unique solution* z_* *to* $\mathrm{VI}(\mathcal{Z}, F + \alpha\nabla g)$ *satisfies* $z_* \in \operatorname{int}\operatorname{dom}\psi$*. Then Algorithm 3.1 converges if* $\eta \leqslant \frac{\alpha}{L^2}$ *and guarantees*

$$B_\psi(z_*; z_{t+1}) \leqslant \left(\frac{1}{1 + \eta\alpha} \right)^t B_\psi(z_*; z_1).$$

Note $\alpha > 0$ is necessary. If $\alpha = 0$ in the context of solving (1), Algorithm 3.1 with $\psi(z) = \frac{1}{2}\|z\|^2$ becomes projected gradient descent ascent, which is known to diverge or cycle for any positive stepsize. However, choosing the strong convexity constants of g and ψ to be 1 is for convenience, the theorem still holds with arbitrary constants, the stepsize condition becomes proportional to the the relative strong convexity constant of g (see Appendix for details).

Due to the generality of VIs, we have the following convex optimization result as a direct consequence.

Corollary 3.5. *Consider the composite optimization problem* $\min_{z \in \mathcal{Z}} f(z) + \alpha g(z)$*. Then under the same assumptions as Theorem 3.4 with* $F = \nabla f$*, Algorithm 3.1 converges linearly to the solution.*

[3]Notice that Assumptions (3.2-3.3) imply g is strongly convex and hence ∇g is strongly monotone.

[4]This assumption is guaranteed in the QRE setting where g is the sum of dilated entropy.

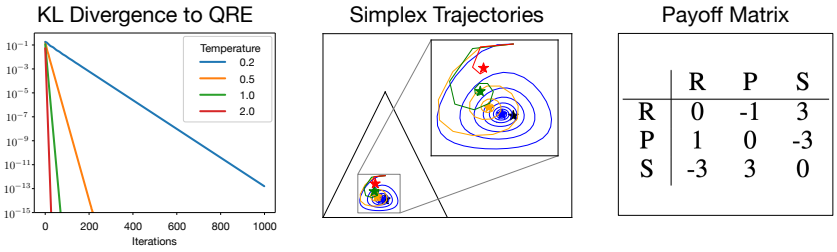

Figure 1: MMD applied to a perturbed rock paper scissors game. Left plot shows linear convergence; middle shows trajectories. Stars show the QRE for that temperature; the black star is the Nash.

3.1 Magnetic Mirror Descent and Applications to Zero-Sum Games

We define MMD to be Algorithm 3.1 with g taken to be either ψ or $B_\psi(\cdot; z')$ for some z'; in both cases the 1-relative strongly convex assumption is satisfied, and z_{t+1} is attracted to either $\min_{z\in\mathcal{Z}} \psi(z)$ or z', which we call the magnet.

Algorithm 3.6 (Magnetic Mirror Descent (MMD))**.**

$$z_{t+1} = \arg\min_{z\in\mathcal{Z}} \eta\left(\langle F(z_t), z\rangle + \alpha\psi(z)\right) + B_\psi(z; z_t) \tag{5}$$

or

$$z_{t+1} = \arg\min_{z\in\mathcal{Z}} \eta\left(\langle F(z_t), z\rangle + \alpha B_\psi(z; z')\right) + B_\psi(z; z_t). \tag{6}$$

Remark 3.7. MMD has the same computational cost as mirror descent, since the updates can be equivalently written as $z_{t+1} = \arg\min_{z\in\mathcal{Z}} \langle \ell, z\rangle + \psi(z)$ (e.g. $\ell = {}^{(\eta F(z_t) - \nabla\psi(x_t))}/_{(1+\eta\alpha)}$ for (5)).

MMD and more generally Algorithm 3.1 can be used to derive a descent-ascent method to solve the zero-sum game (1). If $g_1 = \psi_1$ and $g_2 = \psi_2$ are strongly convex over $\mathcal{X}$ and $\mathcal{Y}$, then we can let $\psi(z) = \psi_1(x) + \psi_2(y)$, and thus ψ is strongly convex over $\mathcal{Z}$. Then the MMD update rule (5) converges to the solution of (1) and conveniently splits into simultaneous descent-ascent updates,

$$x_{t+1} = \arg\min_{x\in\mathcal{X}} \eta\left(\langle \nabla_x f(x_t, y_t), x\rangle + \alpha\psi_1(x)\right) + B_{\psi_1}(x; x_t), \tag{7}$$

$$y_{t+1} = \arg\max_{y\in\mathcal{Y}} \eta\left(\langle \nabla_y f(x_t, y_t), y\rangle - \alpha\psi_2(y)\right) - B_{\psi_2}(y; y_t). \tag{8}$$

Additionally, according to Remark 3.7, the updates are easy to implement whenever mirror descent is (see Appendix for examples).

3.2 Algorithm and Theory Summary

If we take ψ to be negative entropy, then, in reinforcement learning language, MMD takes the form

$$\pi_{t+1}(s) = \arg\max_{\pi(s)} \mathbb{E}_{a\sim\pi(s)} q_{\pi_t}(s, a) - \alpha\mathrm{KL}(\pi(s), \rho(s)) - \frac{1}{\eta}\mathrm{KL}(\pi(s), \pi_t(s)), \tag{9}$$

where π_t is the current policy and ρ is a reference policy (i.e., the magnet). In settings with discrete action spaces, this update rule has a closed form

$$\pi_{t+1}(s) \propto \left[\pi_t(s)\rho(s)^{\eta\alpha} e^{\eta q_{\pi_t}(s)}\right]^{1/(1+\eta\alpha)}. \tag{10}$$

If we set $\rho(s)$ to be uniform, the magnet term $\mathrm{KL}(\pi(s), \rho(s))$ is equivalent to an entropy bonus $\mathcal{H}(\pi(s))$ and the update rule simplifies to $\pi_{t+1}(s) \propto \left[\pi_t(s) e^{\eta q_{\pi_t}(s)}\right]^{1/(1+\eta\alpha)}$.

Our main result, Theorem 3.4, and Proposition 2.4 imply that if both players simultaneously update their policies using equation (9) with a uniform magnet in 2p0s NFGs, then their joint policy converges to the α-QRE exponentially fast. Similarly, in EFGs, if both players use a type of policy called sequence form with ψ taken to be dilated entropy, then their joint policy converges to the α-QRE exponentially fast. MMD can also be a considered as a behavioral-form algorithm in which update rule (9) is applied at each state. If ρ is uniform, the unique fixed point of this instantiation in self-play

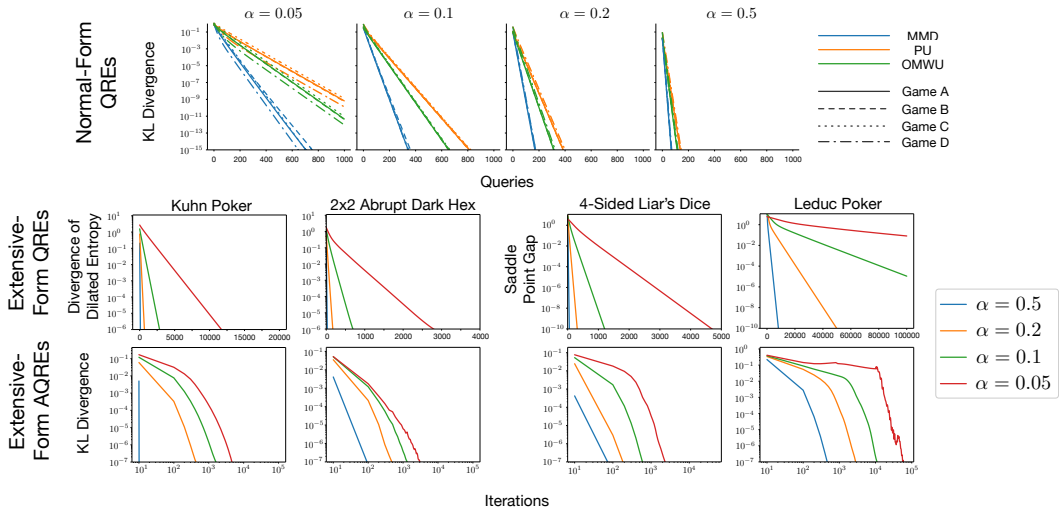

Figure 2: Solving for QREs in various settings.

is the α-AQRE. Note that all of these ideas transfer from 2p0s game settings to single agent settings, using the analogous notions of α-soft optimality.

A toy example demonstrating convergence speed and trajectories is shown in Figure 1. Further details about the theory and algorithm are included in the appendix.

4 Experiments

Our main body focuses on highlighting the high level takeaways of our main experiments. Additional discussion of each experiment, as well as additional experiments, are included in the appendix.

Convergence to Quantal Response Equilibria First, we examine MMD's performance as a QRE solver. We show the results in Figure 2. For normal-form settings, we used a 2p0s variant Diplomacy—an AI benchmark [48] in which each turn can be viewed as a normal-form game because all players act simultaneously and the board state is fully observed. These games have payoff matrices of shape $(50, 50)$, $(35, 43)$, $(50, 50)$, and $(4, 4)$, respectively, and were constructed using an open-source value function [3]. We compared against algorithms introduced by Cen et al. [13], and show results in the top row of the figure. All three algorithms converge exponentially fast, as is guaranteed by theory. The middle row shows results for QREs on EFG benchmarks from OpenSpiel [33]. These games have 54, 471, 8176, and 9300 non-terminal histories, respectively. For Kuhn Poker and 2x2 Abrupt Dark Hex, we observe that MMD's divergence converges exponentially fast, as is guaranteed by theory. For 4-Sided Liar's Dice and Leduc Poker, we found that Gambit [43] was unable to approximate the QREs, due to the size of the games. Thus, we instead report the saddle point gap (the sum of best response values in the regularized game), for which we observe linear convergence, as is guaranteed by Proposition D.3.1). The bottom row shows results for AQREs using behavioral form MMD on the same benchmarks, where we also observe convergence (despite a lack of guarantees).

Convergence to Nash Equilibria Second, we investigate whether MMD can be made to converge in terms of exploitability. There are two design choices involved in using MMD as a Nash equilibrium solver. First, whether to implement it in sequence form or behavioral form; and second, whether to induce convergence by annealing the temperature or by moving the magnet. We show experiments in behavioral form (as this corresponds to the "RL case") with an annealing temperature (though preliminary experiments suggest moving the magnet also works—see Appendix for example).

The results are in Figure 3. The top row shows the full feedback case for Diplomacy stage games. MMD is outperformed by CFR for small iteration numbers, but achieves comparable (or sometimes superior) performance as the number of iterations increases. The second row shows the black box sampling case for Diplomacy stage games. We compare MMD against OS-MCCFR, but note that it is somewhat of an unfair comparison because OS-MCCFR uses off-policy learning to bound its

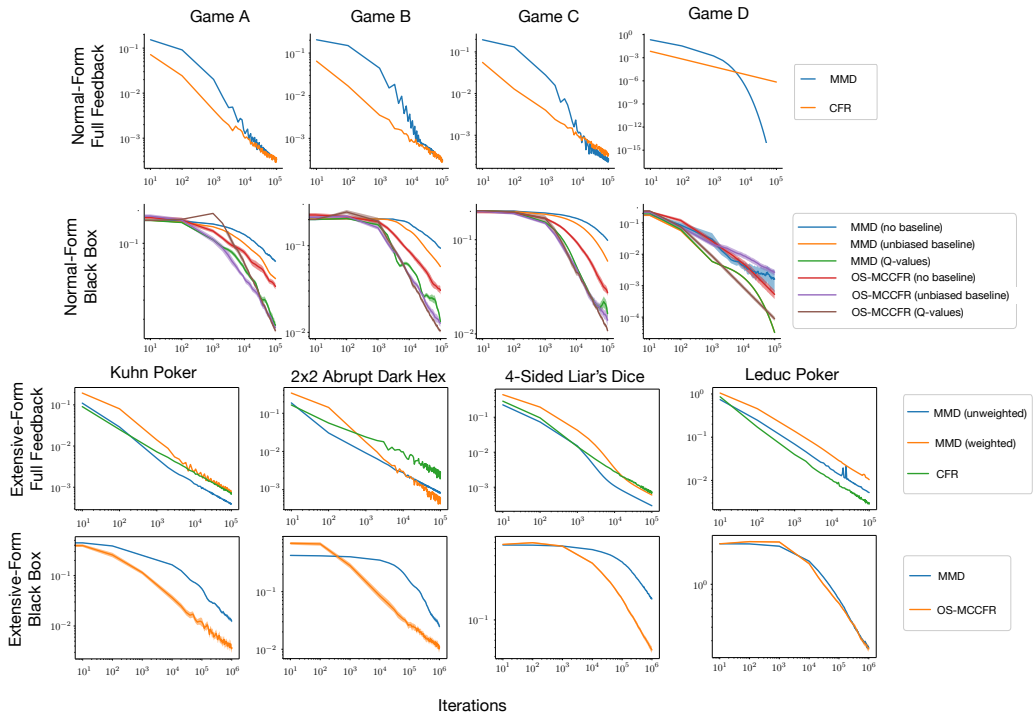

Figure 3: Solving for Nash in various settings.

Table 1: Atari and Mujoco results averaged over 3 runs, with standard errors.

	Breakout	Pong	BeamRider	Hopper-v2	Walker2d-v2	HalfCheetah-v2
PPO	409 ± 31	20.59 ± 0.40	2628 ± 626	2448 ± 596	3142 ± 982	2149 ± 1166
MMD	414 ± 6	21.0 ± 0.00	2549 ± 524	2898 ± 544	2215 ± 840	3638 ± 782

gradients, whereas we present on-policy results for MMD (which leads to unbounded gradients). Off-policy learning is equally applicable to MMD and may yield stronger performance; in contrast, on-policy learning would prohibit OS-MCCFR from converging. With that context in mind, we observe that MMD exhibits slower convergence than OS-MCCFR with unbiased gradient estimates, but matches the performance of the best instance of OS-MCCFR with biased gradient estimates.

The third row shows the full feedback case for EFGs. Weighted MMD weights update sizes by their reach probabilities, whereas unweighted MMD weights all updates equally. We observe that both converge at rates competitive with CFR. *This is the first instance of a standard RL algorithm yielding results competitive with tabular CFR in classical 2p0s benchmark games.* The last row shows the black box sampling case for EFGs. Note that the caveats discussed above for the black box normal-form setting also apply here. That said, similarly to the normal-form case, we observe that while MMD with unbiased Q-value estimates empirically converges, it is generally slower than OS-MCCFR. It may be possible to make up this gap using baselines or biased Q-value estimates, as was done in the normal-form case.

Deep Single-Agent Reinforcement Learning Next, we examine whether MMD is competitive as a single-agent RL algorithm in Atari [10] and Mujoco [59] compared to PPO [55]. We implement MMD as an adaptation of Huang et al.'s PPO implementation. We show results, compared to those reported for PPO by Huang et al. [25] in Table 1. The Atari results are for 10M time steps; the Mujoco results are for 2M time steps. While the exact numbers should be taken lightly, as they are over only three seeds, the results nevertheless suggest that MMD performs roughly comparably to PPO in these settings. This should not be seen as a surprising result, but rather a confirmation of the reality that MMD can be implemented in a fashion that shares many similarities with PPO.

Table 2: Approximate exploitability for 3x3 Abrupt Dark Hex in units of 10^{-2}.

	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5
First Legal Action Taker	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Uniform Random	74 ± 1	76 ± 1	75 ± 1	73 ± 2	74 ± 2
NFSP(1M steps)	97 ± 1	90 ± 1	97 ± 1	96 ± 1	75 ± 1
NFSP(10M steps)	61 ± 2	61 ± 2	60 ± 2	58 ± 2	57 ± 2
MMD(1M steps)	39 ± 2	36 ± 2	38 ± 2	36 ± 2	61 ± 2
MMD(10M steps)	23 ± 2	24 ± 2	24 ± 2	23 ± 2	20 ± 2

Table 3: Head-to-head expected return for row player in 3x3 Abrupt Dark Hex in units of 10^{-2}.

	First Taker	Uniform	NFSP1	NFSP2	NFSP3	NFSP4	NFSP5
First Taker	0	0 ± 2	-81 ± 1	-36 ± 2	-38 ± 2	-40 ± 2	-40 ± 2
Uniform	0 ± 2	0	-38 ± 2	-46 ± 2	-46 ± 2	-39 ± 2	-45 ± 2
MMD1	62 ± 2	66 ± 2	26 ± 2	20 ± 2	33 ± 2	38 ± 2	11 ± 2
MMD2	64 ± 2	64 ± 2	25 ± 2	22 ± 2	34 ± 2	28 ± 2	14 ± 2
MMD3	64 ± 2	70 ± 2	27 ± 2	25 ± 2	37 ± 2	37 ± 2	8 ± 2
MMD4	61 ± 2	63 ± 2	23 ± 2	16 ± 2	37 ± 2	26 ± 2	9 ± 2
MMD5	61 ± 2	63 ± 2	24 ± 2	9 ± 2	24 ± 2	18 ± 2	8 ± 2

Deep Multi-Agent Reinforcement Learning Our last experiment examines MMD as a deep multi-agent RL algorithm using self play. We benchmarked against NFSP [23] in 3x3 Abrupt Dark Hex. We found that 2x2 and 3x2 have 471 and $245{,}243$ non-terminal histories, respectively, but were unable to compute the number for 3x3 due to its size. We implemented MMD by modifying RLlib's [35] PPO implementation, and used OpenSpiel's [33] NFSP implementation. This setup gives an advantage to NFSP in that RLlib's OpenSpiel interface does not inform the agent about which actions are legal (so MMD is playing a harder version of the game), but an advantage to MMD in that it is likely that RLlib's PPO implementation is more optimized than OpenSpiel's NFSP implementation.

As the game is too large to compute exact exploitability, we approximate exploitability using a DQN best response, trained for 10 million time steps. The results of this experiment are shown in Table 2. The results include checkpoints after both 1 million and 10 million time steps, as well as bots that select the first legal action and that select actions uniformly at random. As expected, both NFSP and MMD both yield lower approximate exploitability after 10M steps than they do after 1M steps; that said, among the two, MMD produces substantially better results at both checkpoints. We also show results of head-to-head match-ups in Table 3 for the 10M time step checkpoints. We find the MMD tends to outperform NFSP, both in direct head-to-head match-ups and in terms of exploiting the first action taking and uniformly random bots.

5 Related Work

Convex Optimization and Variational Inequalities Like MMD and Algorithm 3.1, the extragradient method [28, 20] and the optimistic method [49] have also been studied in the context of zero-sum games and variational inequalities more generally. However, in contrast to MMD, neither of these methods are guaranteed to converge without smoothness. Outside the context of variational inequalities, analogues of MMD and Algorithm 3.1 have been studied in convex optimization under the non-Euclidean proximal gradient method [8] originally proposed by Tseng [62]. But, in contrast to Theorem 3.4, existing convex optimization results [8, 62, 22, 7] are without linear rates because they do not assume the proximal regularization to be relatively strongly convex. In addition to convex optimization, the non-Euclidean proximal gradient algorithm has also been studied in online optimization under the name composite mirror descent [15]. Duchi et al. [15] show a $O(\sqrt{t})$ regret bound without strong convexity assumptions on the proximal term. In the case where the proximal term is relatively strongly convex, Duchi et al. [15] give an improved rate of $O(\log t)$—implying that MMD has average iterate convergence with a rate of $O(\log t/t)$ for bounded problems, like QRE solving.

Quantal Response Equilibria Among QRE solvers for NFGs, the PU and OMWPU algorithms from Cen et al. [13], which possess comparable linear convergence rates for NFGs, are most similar

to MMD. However, both PU and OMWPU require two steps per iteration (because of their similarities to mirror-prox [45] and optimistic mirror descent [51]), and PU requires an extra gradient evaluation. In contrast, our algorithm needs only one simple step per iteration (with the same computation cost as mirror descent) and our analysis applies to various choices of mirror map, meaning our algorithm can be used to compute a larger class of regularized equilibria, rather than only logit QREs. Among QRE solvers for EFGs, existing algorithms differ from MMD in that they either require second order information [37] or are first order methods with average iterate convergence [18, 38]. In contrast to these methods, MMD attains linear last-iterate convergence.

Single-Agent Deep Reinforcement Learning Considered as a reinforcement learning algorithm, MMD bears close resemblance to both KL-PPO (a variant of PPO that served as motivation for the widely adopted gradient clipping variant) [55] and MDPO [60]. In short, MMD corresponds with KL-PPO with a flipped KL term and with MDPO with additional proximal regularization. We describe these relationships using symbolic expressions in the appendix.

Average Policy Deep Reinforcement Learning for Two-Player Zero-Sum Games One class of deep reinforcement learning methods for two-player zero-sum games, which includes NFSP and PSRO, works by using single-agent reinforcement learning as a subroutine to compute best responses [23, 32, 44]. While this class of methods is very scalable, it can require computing exponentially many best responses in the number of information states in the game [40], making it very slow in some cases. MMD differs from this class in that it does not use a best response subroutine and in that it does not use averages over historical policies. Another class of methods, which includes deep CFR and double neural CFR, is motivated by scaling CFR [69]—the dominant paradigm in tabular settings—to function approximation [11, 34]. Unfortunately, the sampling variant of CFR [31] requires importance sampling across trajectories, making it difficult to apply members of this class to games with long trajectories. MMD differs from this class both in that it converges in its current policy, rather than the average policy, and in that it does not require importance sampling over trajectories.

Regularized Deep Reinforcement Learning for Two-Player Zero-Sum Games The class of methods under which MMD itself falls uses regularization to encourage last-iterate convergence in self-play. Another method belonging to this class is friction follow-the-regularized-leader (FFoReL) [50]. In terms of convergence guarantees, we prove discrete-time linear convergence for NFGs, while Pérolat et al. [50] state continuous-time (which does not imply discrete time) linear convergence for EFGs. In terms of ease-of-use, MMD offers three advantages over FFoReL: 1) it is decentralized (FFoReL is not), 2) it does not require modifying the reward function (FFoReL does), 3) MMD only requires approximating bounded quantities (FFoReL requires estimating an arbitrarily accumulating sum, making it tedious to scale with function approximation). Lastly, in terms of empirical performance, the tabular results presented in this work for MMD are substantially better than those presented for FFoReL. For example, FFoReL's best result in Leduc is an exploitability of about 0.08 after 200000 iterations—it takes MMD fewer than 1000 iterations to achieve the same value.

6 Conclusion and Future Work

In this work, we introduced MMD—an algorithm for reinforcement learning in single-agent settings and 2p0s games, and QRE solving. We presented a proof that MMD converges exponentially fast to QREs in EFGs—the first algorithm of its kind to do so. We showed empirically that MMD exhibits desirable properties as a tabular equilibrium solver, as a single-agent deep RL algorithm, and as a multi-agent deep RL algorithm. This is the first instance of an algorithm exhibiting such strong performance across all of these settings simultaneously. We hope that, due to its simplicity, MMD will help open the door to 2p0s games research for RL researchers without game theoretic backgrounds.

Our work opens up a multitude of important directions for future work. On the theoretical side, these directions include pursuing results for black box sampling, behavioral-form convergence, convergence with annealing regularization, convergence with moving magnets [36, 1], and relaxing the smoothness assumption to relative smoothness [39, 7] while removing strong convexity assumptions. On the empirical side, these directions include constructing an adaptive mechanism for adapting the stepsize, temperature, and magnet [2, 17], investigating the performance of unbiased baselines and biased Q-values estimates in extensive-form, and pushing the limits of MMD as a deep RL algorithm for large scale 2p0s games.

7 Acknowledgements

We thank Jeremy Cohen, Chun Kai Ling, Brandon Amos, Paul Muller, Gauthier Gidel, Kilian Fatras, and Julien Perolat for helpful discussions and feedback. This research was supported by the Bosch Center for Artificial Intelligence, NSERC Discovery grant RGPIN-2019-06512, Samsung, a Canada CIFAR AI Chair, and the Office of Naval Research Young Investigator Program grant N00014-22-1-2530.

References

[1] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 1200–1205, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345286. doi: 10.1145/3055399.3055448. URL `https://doi.org/10.1145/3055399.3055448`.

[2] A. P. Badia, B. Piot, S. Kapturowski, P. Sprechmann, A. Vitvitskyi, D. Guo, and C. Blundell. Agent57: Outperforming the atari human benchmark. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[3] A. Bakhtin, D. Wu, A. Lerer, and N. Brown. No-press diplomacy from scratch. *Advances in Neural Information Processing Systems*, 34, 2021.

[4] A. Bakst and M. Gardner. The second scientific american book of mathematical puzzles and diversions. *American Mathematical Monthly*, 69:455, 1962.

[5] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on control and optimization*, 42(2):596–636, 2003.

[6] H. H. Bauschke, P. L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[7] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[8] A. Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[9] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. ISSN 0167-6377.

[10] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *J. Artif. Int. Res.*, 47(1):253–279, may 2013. ISSN 1076-9757.

[11] N. Brown, A. Lerer, S. Gross, and T. Sandholm. Deep counterfactual regret minimization. *ArXiv*, abs/1811.00164, 2019.

[12] S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[13] S. Cen, Y. Wei, and Y. Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34, 2021.

[14] T. Davis, M. Schmid, and M. Bowling. Low-variance and zero-variance baselines for extensive-form games. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[15] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, volume 10, pages 14–26. Citeseer, 2010.

[16] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

[17] J. Fan, C. Xiao, and Y. Huang. GDI: rethinking what makes reinforcement learning different from supervised learning. *CoRR*, abs/2106.06232, 2021. URL https://arxiv.org/abs/2106.06232.

[18] G. Farina, C. Kroer, and T. Sandholm. Online convex optimization for sequential decision processes and extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1917–1925, 2019.

[19] C. P. Ferguson and T. S. Ferguson. *Models for the Game of Liar's Dice*, pages 15–28. Springer Netherlands, Dordrecht, 1991. ISBN 978-94-011-3760-7. doi: 10.1007/978-94-011-3760-7_3. URL https://doi.org/10.1007/978-94-011-3760-7_3.

[20] E. Gorbunov, N. Loizou, and G. Gidel. Extragradient method: O(1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 366–402. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/gorbunov22a.html.

[21] E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, page 709–715. AAAI Press, 2004. ISBN 0262511835.

[22] F. Hanzely, P. Richtarik, and L. Xiao. Accelerated bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79(2): 405–440, 2021.

[23] J. Heinrich and D. Silver. Deep reinforcement learning from self-play in imperfect-information games, 2016. URL https://arxiv.org/abs/1603.01121.

[24] S. Hoda, A. Gilpin, J. Pena, and T. Sandholm. Smoothing techniques for computing nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2):494–512, 2010.

[25] S. Huang, R. F. J. Dossa, A. Raffin, A. Kanervisto, and W. Wang. The 37 implementation details of proximal policy optimization. In *ICLR Blog Track*, 2022. URL https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/. https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.

[26] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(98)00023-X. URL https://www.sciencedirect.com/science/article/pii/S000437029800023X.

[27] D. Koller, N. Megiddo, and B. Von Stengel. Efficient computation of equilibria for extensive two-person games. *Games and economic behavior*, 14(2):247–259, 1996.

[28] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. URL https://ci.nii.ac.jp/naid/10017556617/.

[29] C. Kroer, K. Waugh, F. Kılınç-Karzan, and T. Sandholm. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, 179(1):385–417, 2020.

[30] H. Kuhn. 9. a simplified two-person poker. 1951.

[31] M. Lanctot, K. Waugh, M. Zinkevich, and M. Bowling. Monte carlo sampling for regret minimization in extensive games. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/00411460f7c92d2124a67ea0f4cb5f85-Paper.pdf.

[32] M. Lanctot, V. F. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T. Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *NIPS*, 2017.

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

[34] H. Li, K. Hu, S. Zhang, Y. Qi, and L. Song. Double neural counterfactual regret minimization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByedzkrKvH.

[35] E. Liang, R. Liaw, R. Nishihara, P. Moritz, R. Fox, K. Goldberg, J. Gonzalez, M. Jordan, and I. Stoica. RLlib: Abstractions for distributed reinforcement learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3053–3062. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/liang18b.html.

[36] H. Lin, J. Mairal, and Z. Harchaoui. Catalyst acceleration for first-order convex optimization: From theory to practice. *J. Mach. Learn. Res.*, 18(1):7854–7907, jan 2017. ISSN 1532-4435.

[37] C. K. Ling, F. Fang, and J. Z. Kolter. What game are we playing? end-to-end learning in normal and extensive form games. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 396–402, 2018.

[38] C. K. Ling, F. Fang, and J. Z. Kolter. Large scale learning of agent rationality in two-player zero-sum games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6104–6111, 2019.

[39] H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[40] S. M. McAleer, J. B. Lanier, K. Wang, P. Baldi, and R. Fox. XDO: A double oracle algorithm for extensive-form games. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=WDLf8cTq_V8.

[41] R. McKelvey and T. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995. URL https://EconPapers.repec.org/RePEc:eee:gamebe:v:10:y:1995:i:1:p:6-38.

[42] R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for extensive form games. *Experimental Economics*, 1:9–41, 1998.

[43] McKelvey, Richard D., McLennan, Andrew M., and Turocy, Theodore L. Gambit: Software tools for game theory. URL http://www.gambit-project.org.

[44] H. B. McMahan, G. J. Gordon, and A. Blum. Planning in the presence of cost functions controlled by an adversary. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 536–543. AAAI Press, 2003. ISBN 1577351894.

[45] A. Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[46] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983. ISBN 9780471103455.

[47] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[48] P. Paquette, Y. Lu, S. S. Bocco, M. Smith, S. O-G, J. K. Kummerfeld, J. Pineau, S. Singh, and A. C. Courville. No-press diplomacy: Modeling multi-agent gameplay. *Advances in Neural Information Processing Systems*, 32, 2019.

[49] L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980. Publisher: Springer.

[50] J. Pérolat, R. Munos, J.-B. Lespiau, S. Omidshafiei, M. Rowland, P. A. Ortega, N. Burch, T. W. Anthony, D. Balduzzi, B. D. Vylder, G. Piliouras, M. Lanctot, and K. Tuyls. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In M. Meila and T. Z. 0001, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8525–8535. PMLR, 2021. URL http://proceedings.mlr.press/v139/perolat21a.html.

[51] A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

[52] R. T. Rockafellar. Monotone operators associated with saddle.functions and minimax problems, 1970.

[53] I. Romanovskii. Reduction of a game with full memory to a matrix game. *Doklady Akademii Nauk SSSR*, 144(1):62–+, 1962.

[54] M. Schmid, N. Burch, M. Lanctot, M. Moravcik, R. Kadlec, and M. Bowling. Variance reduction in monte carlo counterfactual regret minimization (vr-mccfr) for extensive form games using baselines. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33012157. URL https://doi.org/10.1609/aaai.v33i01.33012157.

[55] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

[56] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, USA, 2008. ISBN 0521899435.

[57] F. Southey, M. Bowling, B. Larson, C. Piccione, N. Burch, D. Billings, and C. Rayner. Bayes' bluff: Opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, page 550–558, Arlington, Virginia, USA, 2005. AUAI Press. ISBN 0974903914.

[58] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[59] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[60] M. Tomar, L. Shani, Y. Efroni, and M. Ghavamzadeh. Mirror descent policy optimization, 2020. URL https://arxiv.org/abs/2005.09814.

[61] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

[62] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.

[63] T. L. Turocy. A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. *Games and Economic Behavior*, 51(2):243–263, 2005. ISSN 0899-8256. doi: https://doi.org/10.1016/j.geb.2004.04.003. URL https://www.sciencedirect.com/science/article/pii/S0899825604000739. Special Issue in Honor of Richard D. McKelvey.

[64] T. L. Turocy. Computing sequential equilibria using agent quantal response equilibria. *Economic Theory*, 42(1):255–269, 2010. ISSN 09382259, 14320479. URL http://www.jstor.org/stable/25619985.

[65] J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1947.

[66] B. Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

[67] H. Zhang, A. Lerer, and N. Brown. Equilibrium finding in normal-form games via greedy regret minimization, 2022. URL https://arxiv.org/abs/2204.04826.

[68] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1433–1438. AAAI Press, 2008. ISBN 9781577353683.

[69] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/file/08d98638c6fcd194a4b1e6992063e944-Paper.pdf.

A Problem Setting

In our notation, we use

- $s \in \mathbb{S}$ to notate Markov states,
- $a_i \in \mathbb{A}_i$ to notate actions,
- $o_i \in \mathbb{O}_i$ to notate observations,
- $h_i \in \mathbb{H}_i = \bigcup_t (\mathbb{O}_i \times \mathbb{A}_i)^t \times \mathbb{O}_i$ to denote information states (i.e., decision points).

We use

- $\mathcal{T}\colon \mathbb{S} \times \mathbb{A} \to \Delta(\mathbb{S} \cup \{\perp\})$ to notate the transition function, where $\perp$ notates termination,
- $\mathcal{R}_i\colon \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ to notate a reward function,
- $\mathcal{O}_i\colon \mathbb{S} \times \mathbb{A} \to \mathbb{O}_i$ to notate an observation function.
- $\mathcal{A}_i\colon \mathbb{H}_i \to \mathbb{A}_i$ to notate a legal action function.

We are interested in 2p0s games, in which $i \in \{1, 2\}$ and $\forall s, a, \mathcal{R}_1(s, a) = -\mathcal{R}_2(s, a)$. For convenience, we use $-i$ to notate the player "not i". Single-agent settings are captured as a special case in which the second player has a trivial action set $|\mathbb{A}_2| = 1$. Normal-form games are captured as a special case in which there is only one state s and the transition function only supports termination: $\forall a, \text{supp}(\mathcal{T}(s, a)) = \{\perp\}$. (Here, we use $\text{supp}(\mathcal{X})$ to denote the support of a distribution $\mathcal{X}$—i.e., the subset of the domain of $\mathcal{X}$ that is mapped to a value greater than zero: $\{x : \mathcal{X}(x) > 0\}$.)

Each agent's goal is to maximize its expected return

$$\mathbb{E}\left[\sum_t \mathcal{R}_i(S^t, A^t) \mid \pi\right]$$

using its policy π_i, which dictates s distribution over actions for each information state

$$\pi_i\colon \mathbb{H}_i \to \Delta(\mathbb{A}_i).$$

In game theory literature, these policies are called behavioral form and assume perfect recall.

We notate the expected value for an agent's return a_i at an information state h_i at time t under joint policy π as

$$q_\pi(h_i, a_i) = \mathbb{E}\left[\mathcal{R}_i(S, A_{-i}, a_i) + \sum_{t' > t} \mathcal{R}_i(S^{t'}, A^{t'}) \mid \pi, h_i, a_i\right].$$

Here, the first expectation samples the current Markov state S, the current joint history H, and the current opponent action A_{-i} from the posterior induced by player i reaching information state h_i, when each player uses its part of joint policy π to determine its actions. The second expectation is over trajectories under the same conditions, with the additional condition that a_i is the agent's action at the current time step.

A.1 Reduction to Normal Form

Given any game of the above form, we can reduce the game to normal form as follows. Let $\bar{\Pi}_i$ denote the set of deterministic policies—i.e., the set of policies that support exactly one action at a time:

$$\bar{\Pi}_i = \{\pi_i : \forall h_i \, |\text{supp}(\pi_i(h_i))| = 1\}.$$

The action space of the normal-form game is the space of deterministic policies: $\tilde{\mathbb{A}}_i = \bar{\Pi}_i$.[5] The reward function of the normal-form game is dictated by the expected return of the deterministic joint policy:

$$\tilde{\mathcal{R}}_i(\cdot, \bar{\pi}) = \mathbb{E}\left[\sum_t \mathcal{R}(S^t, A^t) \mid \bar{\pi}\right].$$

[5] Although the actions $\tilde{\mathbb{A}}_i$ give an equivalent normal-form representation, many of the actions are redundant because actions taken at certain decision points may make other decision points unreachable. The *reduced normal-form* (a.k.a. reduced strategic form) removes duplicate actions by identifying redundant choices at future decision points that are unreachable [47]. Hereinafter we consider the reduced normal-form.

Remark A.1. Any policy π_i can be expressed as a finite mixture over policies in $\bar{\Pi}_i$ that induces the same distribution over trajectories (against arbitrary, but fixed, opponents). Conversely, any finite mixture over policies in $\bar{\Pi}_i$ can be expressed as a policy π_i that induces the same distribution over trajectories (against arbitrary, but fixed, opponents).

By the remark above, joint policies in the original game possess counterparts in the normal-form game (and vice versa) achieving identical expected returns. It is in the sense that the normal-form game is equivalent to the original game.

A more detailed exposition on this equivalence can be found in Shoham and Leyton-Brown [56].

B Solution Concepts

Nash equilibria are perhaps the most commonly sought-after solution concept in 2p0s games. A joint policy π_1, π_2 is a Nash equilibrium if neither player can improve its expected return by changing its policy (assuming the other player does not change its policy):

$$\forall i, \pi_i \in \arg\max_{\pi_i'} \mathbb{E}\left[\sum_t \mathcal{R}_i(S^t, A^t) \mid \pi_i', \pi_{-i}\right].$$

Note that, in single-agent settings, this corresponds with the notion of an optimal policy in reinforcement learning.

Another solution concept is a logit quantal response equilibrium [41, 42]. As we only deal with logit quantal response equilibria, we generally drop logit and refer to them simply as quantal reponse equilibria. In normal-form games, there are multiple equivalent ways to define a quantal response equilibrium. One way is using entropy regularization. We say a joint policy is an α-QRE in a normal-form game if each player maximizes a weighted combination of expected return and policy entropy

$$\forall i, \pi_i \in \arg\max_{\pi_i'} \mathbb{E}\left[\mathcal{R}_i(\cdot, A) + \alpha\mathcal{H}(\pi_i') \mid \pi_i', \pi_{-i}\right].$$

In a temporally-extended game, we say a joint policy is an α-QRE if the equivalent mixture over deterministic joint policies is an α-QRE of the equivalent normal-form game.

An alternative way to extend QREs to temporally extended settings is to ask that they satisfy the normal-form QRE condition at each information state:

$$\forall i, \forall h_i, \pi_i(h_i) \in \arg\max_{\pi_i'(h_i)} \mathbb{E}_{A\sim\pi_i'(h_i)}\left[q_\pi(h_i, A) + \alpha\mathcal{H}(\pi_i'(h_i))\right].$$

When a joint policy satisfies this condition, it is called an agent QRE (as it is as if there is a separate agent playing a part of a normal-form QRE at each information state). In single-agent settings, α-AQREs correspond with the fixed point of the instantiation of expected SARSA [58] in which the policy is a softmax distribution over Q-values with temperature α.

C Reduced Normal-Form Logit-QREs and MMD

C.1 Sequence-Form Background

A Nash-equilibrium in a 2p0s extensive-form game can be formulated as a bilinear saddle point problem over the sequence form [47]

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

where $\mathcal{X}$ and $\mathcal{Y}$ are the sequence form polytopes, which equivalently can be viewed as treeplexes [24, 29]. We provide some background on the sequence form in the context of the min player (player 1), the max player follows similarly. Recall all the decision points for player 1 are denoted as $\mathbb{H}_1$ (also known as information states) and the actions available at decision point $h \in \mathbb{H}_1$ are $\mathcal{A}_1(h)$. Recall that a pollicy (a.k.a behavioral-form strategy) is denoted as π_1 with $\pi_1(h) \in \Delta(\mathcal{A}_1(h))$ being the policy at decision point h. For convenience, let $\pi_1(h, a)$ denotes the probability of taking action $a \in \mathcal{A}_1(h)$ at decision point h. Next we denote $p(h)$ as the parent sequence to reach decision point h, that is the unique previous decision point and action taken by the player before reaching h. Note

that this parent is unique due to perfect recall, and it is possible for many decision points to share the same parent. Then we can construct the sequence form from the top down, where the (h, a) sequence of $x \in \mathcal{X}$ is given by

$$x^{(h,a)} = x^{p(h)}\pi(h, a).$$

For convenience, the root sequence $\varnothing$ is defined to be the parent of all initial decision points of the game and is set to the constant $x^{\varnothing} = 1$. We denote x^h as the slice of $x = (x^{(h,a)})_{h\in\mathbb{H}_1, a\in\mathcal{A}(h)}$ corresponding to decision point h. Note we have the following relationship

$$\pi(h) = x^h / x^{p(h)}.$$

Because $x^{(h,a)}$ corresponds to the probability of player 1 choosing *all* actions along the sequence until reaching (h, a), we get that $x^\top Ay$ is the expected payoff for player 2 given a pair of sequence-form strategies x, y. Thus the sequence form allows us to get a bilinear objective.

Given the bilinear structure of the sequence-form problem, we convert the problem into a VI using first-order optimality conditions. Then, in order to apply MMD (or other first-order methods), we need a good choice of a mirror map for $\mathcal{X}$ and $\mathcal{Y}$. A useful choice of mirror map that leads to efficient proximal steps is the class of *dilated distance generating functions* [24]:

$$\psi(x) = \sum_{h\in\mathbb{H}_1} \beta_h x^{p(h)} \psi_h\left(\frac{x^h}{x^{p(h)}}\right) \tag{11}$$

$$= \sum_{h\in\mathbb{H}_1} \beta_h x^{p(h)} \psi_h(\pi(h)), \tag{12}$$

where $(\beta_h)_{h\in\mathbb{H}_1} > 0$ are per-decision-point weights and ψ_h is a distance-generating function for the simplex associated to h. If ψ_h is taken to be the negative entropy then we say ψ is the dilated entropy function. In the normal-form setting the dilated entropy is simply the standard negative entropy.

Recently it was shown that a α-QRE (for the reduced normal form) is the solution to the following saddle point problem over the sequence form [37],

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \alpha\psi_1(x) + x^\top Ay - \alpha\psi_2(y), \tag{13}$$

where ψ_1, ψ_2 are dilated entropy functions with weights $\beta_h = 1$. Note that we have the normal form α-QRE as a special case of (13).

C.2 Proof for Proposition 2.4

Proposition 2.4. *The solution to a normal-form reduced logit QRE in a two-player zero-sum EFG is equivalent to solving* $\mathrm{VI}(\mathcal{Z}, F + \alpha\nabla\psi)$ *where* $\mathcal{Z}$ *is the cross-product of the sequence form strategy spaces, and* ψ *is the sum of the dilated entropy functions for each player. The function* ψ *is strongly convex with respect to* $\|\cdot\|$*. Furthermore,* F *is monotone and* $\max_{ij}|A_{ij}|$*-smooth (*A *being the sequence-form payoff matrix) with respect to* $\|\cdot\|$*, and* $F + \alpha\nabla g$ *is strongly monotone.*

Proof. The problem of finding a reduced normal-form logit QRE is equivalent to solving the saddle-point problem stated in (13) [37]. Therefore, due to the convexity of ψ_1 and ψ_2 and the discussion from Section 2.3, we have that the solution to (13) is equivalent to the solution of $\mathrm{VI}(\mathcal{Z}, F + \nabla\psi)$ where,

$$F(z) = \begin{bmatrix} Ay \\ -A^\top x \end{bmatrix}, \quad \nabla\psi(z) = \begin{bmatrix} \nabla_x\psi_1(x) \\ \nabla_y\psi_2(y) \end{bmatrix}. \tag{14}$$

From Hoda et al. [24] we know there exists constants μ_1, μ_2 such that ψ_1 is μ_1-strongly convex over $\mathcal{X}$ with respect to $\|\cdot\|_1$ and ψ_2 is μ_2-strongly convex over $\mathcal{Y}$ with respect to $\|\cdot\|_1$ (Hoda et al. [24] do not show bounds on these constants, but we only need them to exist). Therefore, ψ is also strongly convex over $\mathcal{Z}$ with constant $\min\{\mu_1, \mu_2\}$ with respect to $\|z\| = \sqrt{\|x\|_1^2 + \|y\|_1^2}$ since, for $z = (x, y)$, and $z' = (x', y')$ we have

$$\begin{aligned}
\langle\nabla\psi(z) - \nabla\psi(z'), z - z'\rangle &= \langle\nabla\psi_1(x) - \nabla\psi_1(x'), x - x'\rangle + \langle\nabla\psi_2(y) - \nabla\psi_2(y'), y - y'\rangle \\
&\geqslant \mu_1\|x - x'\|_1^2 + \mu_2\|y - y'\|_1^2 \\
&\geqslant \min\{\mu_1, \mu_2\}\left(\|x - x'\|_1^2 + \|y - y'\|_1^2\right) \\
&= \min\{\mu_1, \mu_2\}\|z - z'\|^2.
\end{aligned}$$

Following Theorem 3.4 it is useful to characterize the smoothness of F under the same norm for which ψ is strongly-convex. First notice that for any matrix A we have that $\|Ax - Ay\|_\infty \leqslant \max_{ij} |A_{ij}| \|x - y\|_1$(see for example Bubeck et al. [12][Section 5.2.4]). Therefore altogether we have,

$$\begin{aligned}\|F(z) - F(z')\|_*^2 &= \|Ay - Ay'\|_\infty^2 + \|A^\top x - A^\top x'\|_\infty^2 \\ &\leqslant \max_{ij} |A_{ij}|^2 \left(\|y - y'\|_1^2 + \|x - x'\|_1^2\right) \\ &= \max_{ij} |A_{ij}|^2 \|z - z'\|^2,\end{aligned}$$

showing that F is $L = \max_{ij} |A_{ij}|$-smooth with respect to $\|\cdot\|$. The strong-monotonicity of $F + \nabla\psi$ follows since F is monotone and $\nabla\psi$ is strongly monotone since ψ is strongly convex. □

Note that in general the Hessian can have unbounded entries [29] meaning that $\nabla\psi$ cannot be L-smooth [8]. Our MMD algorithm which handles ψ in closed form allows us to sidestep this issue. We also have that the dual norm of $\|\cdot\|$ is simply $\|z\|_* = \sqrt{\|x\|_\infty^2 + \|y\|_\infty^2}$ [12, 45].

C.3 MMD for Finding Reduced Normal-Form QREs over the Sequence-Form

From Proposition 2.4 and Corollary D.6 we have that the MMD descent-ascent updates (7-8) with ψ_1,ψ_2, taken to be dilated entropy with $\eta \leqslant {}^{\alpha}/_{\max_{ij} |A_{ij}|^2}$ converges linearly to the solution of (13). The updates, as mentioned by Remark 3.7, can be computed in closed-form as a one-line change to mirror descent with dilated-entropy [29]. Indeed, setting $g_t = Ay_t$ (the gradient for the min player), we have that the update for the min player can be written as follows

$$\begin{aligned}x_{t+1} &\arg\min_{x\in\mathcal{X}} \langle \eta g_t, x\rangle + \eta\alpha\psi(x) + B_\psi(x; x_t) \\ &= \arg\min_{x\in\mathcal{X}} \langle \eta g_t - \nabla\psi(x_t), x\rangle + (\eta\alpha + 1)\psi(x) \\ &= \arg\min_{x\in\mathcal{X}} \langle \frac{\eta g_t - \nabla\psi(x_t)}{(1+\eta\alpha)}, x\rangle + \psi(x) \\ &= \arg\min_{x\in\mathcal{X}} \sum_{h\in\mathbb{H}_1} \langle \frac{\eta g_t^h - \nabla\psi(x_t)^h}{(1+\eta\alpha)}, x^h\rangle + x^{p(h)}\psi_h(x^h/x^{p(h)}) \\ &= \arg\min_{x\in\mathcal{X}} \sum_{h\in\mathbb{H}_1} x^{p(h)} \left(\langle \frac{\eta g_t^h - \nabla\psi(x_t)^h}{(1+\eta\alpha)}, \pi(h)\rangle + \psi_h(\pi(h))\right).\end{aligned}$$

Updates can be computed in closed-form starting from decision points h without any children and progressing upwards in the game tree.

D Proofs

D.1 Supporting Lemmas and Propositions

Proposition D.1 ([5]Proposition 2.3)**.** *Let $\{x, y\} \subset \operatorname{dom}\psi$ and $\{u, v\} \subset \operatorname{int}\operatorname{dom}\psi$. Then*

1. $B_\psi(u; v) + B_\psi(v; u) = \langle \nabla\psi(u) - \nabla\psi(v), u - v\rangle$
2. $B_\psi(x; u) = B_\psi(x; v) + B_\psi(v; u) + \langle \nabla\psi(v) - \nabla\psi(u), x - v\rangle$
3. $B_\psi(x; v) + B_\psi(y; u) = B_\psi(x; u) + B_\psi(y; v) + \langle \nabla\psi(u) - \nabla\psi(v), x - y\rangle$.

The following result is also known as the Non-Euclidean prox theorem [8][Theorem 9.12] or the three-point property[61].

Proposition D.2. *Assuming $\mathcal{Z}$ closed convex and both f and ψ are differentiable at $\bar{z}$ (defined below). Then the following statements are equivalent*

1. $\bar{z} = \arg\min_{z\in\mathcal{Z}} \eta\langle g, z\rangle + f(z) + B_\psi(z; y)$

2. $\forall z \in \mathcal{Z} \quad \langle \eta g + \nabla f(\bar{z}), \bar{z} - z \rangle \leqslant B_\psi(z; y) - B_\psi(z; \bar{z}) - B_\psi(\bar{z}, y)$

Proof.

$$\begin{aligned}
\bar{z} = \arg\min_{z \in \mathcal{X}} \eta\langle g, z\rangle + f(z) + B_\psi(z; y) &\Leftrightarrow \langle \nabla\psi(\bar{z}) + \eta g + \nabla f(\bar{z}) - \nabla\psi(y), z - \bar{z}\rangle \geqslant 0 \quad \forall z \in \mathcal{Z} \\
&\Leftrightarrow \langle \nabla\psi(y) - \nabla\psi(\bar{z}) - \eta g - \nabla f(\bar{z}), z - \bar{z}\rangle \leqslant 0 \quad \forall z \in \mathcal{Z} \\
&\Leftrightarrow \langle \eta g + \nabla f(\bar{z}), \bar{z} - z\rangle \leqslant \langle \nabla\psi(\bar{z}) - \nabla\psi(y), z - \bar{z}\rangle \quad \forall z \in \mathcal{Z} \\
&\Leftrightarrow \langle \eta g + \nabla f(\bar{z}), \bar{z} - z\rangle \leqslant B_\psi(z; y) - B_\psi(z; \bar{z}) - B_\psi(\bar{z}; y) \, \forall z \in \mathcal{Z}.
\end{aligned}$$

The first equivalence follows by the first-order optimality condition and the last one by Proposition D.2. □

Lemma D.3. *One step of Algorithm 3.1, under the assumptions of Theorem 3.4 guarantees for all* $z \in \mathcal{Z}$

$$B_\psi(z; z_{t+1}) \leqslant \tag{15}$$
$$B_\psi(z; z_t) - B_\psi(z_{t+1}; z_t) + \langle \eta F(z_t) + \eta\alpha\nabla g(z_{t+1}), z - z_{t+1}\rangle. \tag{16}$$

Proof. Immediate from Proposition D.2. □

Lemma D.4. *Under the same assumptions as Theorem 3.4 and let* z_* *be the solution to* $\mathrm{VI}(\mathcal{Z}, F + \alpha\nabla g)$ *then for any* $z \in \mathcal{Z} \cap \operatorname{int}\operatorname{dom}\psi$ *the following inequality holds*

$$\langle \eta F(z) + \eta\alpha\nabla g(z), z_* - z\rangle \leqslant -\eta\alpha \left(B_\psi(z; z_*) + B_\psi(z_*; z)\right).$$

Proof.

$$\begin{aligned}
\langle \eta F(z) + \eta\alpha\nabla g(z), z_* - z\rangle &= \langle \eta F(z) + \eta\alpha\nabla g(z) - \eta F(z_*) - \eta\alpha\nabla g(z_*), z_* - z\rangle \\
&\quad + \langle \eta F(z_*) + \eta\alpha\nabla g(z_*), z_* - z\rangle \\
&= \underbrace{\langle \eta F(z) - \eta F(z_*), z_* - z\rangle}_{\leqslant 0} + \eta\alpha\langle \nabla g(z) - \nabla g(z_*), z_* - z\rangle \\
&\quad + \underbrace{\langle \eta F(z_*) + \eta\alpha\nabla g(z_*), z_* - z\rangle}_{\leqslant 0} \\
&\leqslant \eta\alpha\langle \nabla g(z) - \nabla g(z_*), z_* - z\rangle \\
&= -\eta\alpha\langle \nabla g(z) - \nabla g(z_*), z - z_*\rangle \\
&\leqslant -\eta\alpha\langle \nabla\psi(z) - \nabla\psi(z_*), z - z_*\rangle \\
&= -\eta\alpha \left(B_\psi(z; z_*) + B_\psi(z_*; z)\right)
\end{aligned}$$

Note that $\nabla g(z), \nabla g(z_*), \nabla\psi(z), \nabla\psi(z_*)$, are all well-defined because $z_* \in \operatorname{int}\operatorname{dom}\psi$ and Assumptions (3.2-3.3). The first inequality follows since F is monotone and $z_* \in \operatorname{sol}\mathrm{VI}(\mathcal{Z}, F + \alpha\nabla g)$. The second inequality follows since g is 1-strongly convex relative to ψ and the last equality by Proposition D.1. □

D.2 Proof of Theorem 3.4

Theorem 3.4. *Let Assumptions 3.2 and 3.3 hold, and assume the unique solution* z_* *to* $\mathrm{VI}(\mathcal{Z}, F + \alpha\nabla g)$ *satisfies* $z_* \in \operatorname{int}\operatorname{dom}\psi$. *Then Algorithm 3.1 converges if* $\eta \leqslant \frac{\alpha}{L^2}$ *and guarantees*

$$B_\psi(z_*; z_{t+1}) \leqslant \left(\frac{1}{1 + \eta\alpha}\right)^t B_\psi(z_*; z_1).$$

Proof.

$$
\begin{aligned}
&B_\psi(z_*; z_{t+1}) \leqslant B_\psi(z_*; z_t) - B_\psi(z_{t+1}; z_t) + \langle \eta F(z_t) + \eta\alpha\nabla g(z_{t+1}), z_* - z_{t+1}\rangle \\
&= B_\psi(z_*; z_t) - B_\psi(z_{t+1}; z_t) + \langle \eta F(z_t) - \eta F(z_{t+1}), z_* - z_{t+1}\rangle + \langle \eta F(z_{t+1}) + \eta\alpha\nabla g(z_{t+1}), z_* - z_{t+1}\rangle \\
&\leqslant B_\psi(z_*; z_t) - B_\psi(z_{t+1}; z_t) + \langle \eta F(z_t) - \eta F(z_{t+1}), z_* - z_{t+1}\rangle - \eta\alpha\left(B_\psi(z_{t+1}; z_*) + B_\psi(z_*; z_{t+1})\right) \\
&\leqslant B_\psi(z_*; z_t) - B_\psi(z_{t+1}; z_t) + \eta L\|z_t - z_{t+1}\|\|z_* - z_{t+1}\| - \eta\alpha\left(B_\psi(z_{t+1}; z_*) + B_\psi(z_*; z_{t+1})\right) \\
&\leqslant B_\psi(z_*; z_t) - B_\psi(z_{t+1}; z_t) + \frac{1}{2}\|z_t - z_{t+1}\|^2 + \frac{\eta^2 L^2}{2}\|z_* - z_{t+1}\|^2 - \eta\alpha\left(B_\psi(z_{t+1}; z_*) + B_\psi(z_*; z_{t+1})\right) \\
&\leqslant B_\psi(z_*; z_t) + \eta^2 L^2 B_\psi(z_{t+1}; z_*) - \eta\alpha\left(B_\psi(z_{t+1}; z_*) + B_\psi(z_*; z_{t+1})\right) \\
&\overset{\eta^2L^2\leqslant\eta\alpha}{\leqslant} B_\psi(z_*; z_t) - \eta\alpha B_\psi(z_*; z_{t+1}).
\end{aligned}
$$

The first inequality follows from Lemma D.3 and the second inequality from Lemma D.4. The third inequality by the generalized Cauchy-Schwarz inequality and the smoothness of F. The fourth inequality by elementary inequality $ab \leqslant \frac{\rho a^2}{2} + \frac{b^2}{2\rho} \quad \forall \rho > 0$, and the fifth inequality by the strong convexity of ψ since $\frac{1}{2}\|x - y\|^2 \leqslant B_\psi(x; y)$. Therefore altogether we have

$$B_\psi(z_*; z_{t+1}) \leqslant \frac{B_\psi(z_*; z_t)}{1 + \eta\alpha}.$$

Iterating the inequality yields the result. □

Corollary D.6. *Under the same assumptions as Theorem 3.4 if g is μ-strongly convex relative to ψ and ψ is μ_ψ strongly convex then if $\eta \leqslant \frac{\alpha\mu}{L^2}$, Algorithm 3.1 guarantees*

$$B_\psi(z_*; z_{t+1}) \leqslant \left(\frac{1}{1 + \eta\mu\alpha}\right)^t B_\psi(z_*; z_1).$$

Proof. Observe that that $\bar{\psi} = \frac{\psi}{\mu_\psi}$ is 1-strongly convex, and $\bar{g} = \frac{g}{\mu\mu_\psi}$ is 1-strongly convex relative to $\bar{\psi}$,

$$
\begin{aligned}
\langle \nabla\bar{g}(x) - \nabla\bar{g}(y), x - y\rangle &= \frac{1}{\mu\mu_\psi}\langle \nabla g(x) - \nabla g(y), x - y\rangle &(17)\\
&\geqslant \frac{1}{\mu_\psi}\langle \nabla\psi(x) - \nabla\psi(y), x - y\rangle &(18)\\
&= \langle \nabla\bar{\psi}(x) - \nabla\bar{\psi}(y), x - y\rangle. &(19)
\end{aligned}
$$

Rewriting the update of Algorithm 3.1 in terms of $\bar{g}$ and $\bar{\psi}$ gives

$$
\begin{aligned}
&\arg\min_{z\in\mathcal{Z}} \eta\left(\langle F(z_t), z\rangle + \alpha g(z)\right) + B_\psi(z; z_t) \\
\Leftrightarrow &\arg\min_{z\in\mathcal{Z}} \eta\left(\langle F(z_t), z\rangle + \frac{\alpha\mu\mu_\psi}{\mu\mu_\psi} g(z)\right) + \frac{\mu_\psi}{\mu_\psi} B_\psi(z; z_t) \\
\Leftrightarrow &\arg\min_{z\in\mathcal{Z}} \frac{\eta}{\mu_\psi}\left(\langle F(z_t), z\rangle + \frac{\alpha\mu\mu_\psi}{\mu\mu_\psi} g(z)\right) + \frac{1}{\mu_\psi} B_\psi(z; z_t) \\
\Leftrightarrow &\arg\min_{z\in\mathcal{Z}} \bar{\eta}\left(\langle F(z_t), z\rangle + \alpha\mu\mu_\psi\bar{g}(z)\right) + B_{\bar{\psi}}(z; z_t) \\
\Leftrightarrow &\arg\min_{z\in\mathcal{Z}} \bar{\eta}\left(\langle F(z_t), z\rangle + \bar{\alpha}\bar{g}(z)\right) + B_{\bar{\psi}}(z; z_t).
\end{aligned}
$$

The result follows from Theorem 3.4 with stepsize $\bar{\eta} = \frac{\eta}{\mu_\psi}$ and $\bar{\alpha} = \mu\mu_\psi\alpha$.

□

D.3 Euclidean MMD Example

We discuss the Euclidean case for update (5) (update 6 is similar). In the Eucldiean case were $\psi = \frac{1}{2}\|\cdot\|_2^2$ we have that update (5) reduces to

$$z_{t+1} = \Pi_{\mathcal{Z}}\left(\frac{z_t - \eta F(z_t)}{1 + \eta\alpha}\right),$$

where $\Pi_{\mathcal{Z}}$ denotes the Euclidean projection onto $\mathcal{Z}$. In the context of solving min max problems where $\psi = \psi_1 + \psi_2$, the sum of $\frac{1}{2}\|\cdot\|_2^2$ then the descent-ascent updates of (7,8) become

$$x_{t+1} = \Pi_{\mathcal{X}}\left(\frac{x_t - \eta\nabla_x f(x_t, y_t)}{1+\eta\alpha}\right),$$
$$y_{t+1} = \Pi_{\mathcal{Y}}\left(\frac{y_t + \eta\nabla_y f(x_t, y_t)}{1+\eta\alpha}\right).$$

Note that our results don't require bounded constraints, in the unconstrained setting there would be no projection step. By Theorem 3.4 the above iterations converge linearly to the solution of

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \quad \frac{\alpha}{2}\|x\|_2^2 + f(x,y) - \frac{\alpha}{2}\|y\|_2^2,$$

provided f is smooth in the sense that $F(x,y) = [\nabla_x f(x,y), -\nabla_y f(x,y)]^\top$ is smooth.

D.3.1 Bounding the Gap

Below we show how Theorem 3.4 can be used to guarantee linear convergence of the gap.

Proposition D.7. *Suppose the assumptions of Theorem 3.4 hold. Moreover, assume that g is twice continuously differentiable over $\operatorname{int}\operatorname{dom}\psi$, and $\mathcal{Z}$ is bounded. In that case, there exists a constant C and a time step t' such that for any $t \geqslant t'$*

$$\theta_{gap}(z_t) = \sup_{z\in\mathcal{Z}}\langle F(z_t) + \alpha\nabla g(z_t), z_t - z\rangle \leqslant C\left(\sqrt{\frac{1}{1+\eta\alpha}}\right)^{t-t'}\sqrt{B_\psi(z_*; z_{t'})}.$$

Proof. By Theorem 3.4 we know $z_t \to z_*$ where $\{z_t\}_{t\geqslant 1}\cup\{z_*\} \subseteq \operatorname{int}\operatorname{dom}\psi$. Therefore, we have that $\{z_t\}_{t\geqslant 1}\cup\{z_*\}$ is eventually within a closed ball centered at z_*. That is, there exists t' and a closed ball B such that $\{z_t\}_{t\geqslant t'}\cup\{z_*\} \subseteq B \subseteq \operatorname{int}\operatorname{dom}\psi$. Since B is compact and $\nabla^2 g$ is continuous over B we have that $\nabla^2 g(z)$ is bounded on B. Therefore, there exists L_B such that $\|\nabla g(z') - \nabla g(z)\|_* \leqslant L_B\|z - z'\|$ for any $z, z' \in B$. Setting $G = F + \nabla g$ we have that for any $z, z' \in B$, $\|G(z) - G(z')\|_* \leqslant \tilde{L}\|z - z'\|$ for $\tilde{L} = L + L_B$.

Then for any $z \in \mathcal{Z}$, $t \geqslant t'$, and denoting x_* as the solution to $\mathrm{VI}(\mathcal{Z}, G)$ we have

$$\begin{aligned}
\langle G(z_t), z_t - z\rangle &= \langle G(z_*), z_t - z\rangle + \langle G(z_t) - G(z_*), z_t - z\rangle \\
&= \underbrace{\langle G(z_*), z_* - z\rangle}_{\leqslant 0} + \langle G(z_*), z_t - z_*\rangle + \langle G(z_t) - G(z_*), z_t - z\rangle \\
&\leqslant \|G(z_*)\|_*\|z_t - z_*\| + \tilde{L}\|z_t - z_*\|\|z_t - z\| \\
&\leqslant \left(\|G(z_*)\|_* + \tilde{L}D\right)\|z_t - z_*\| \\
&\leqslant C\sqrt{B_\psi(z_*; z_t)} \\
&\leqslant C\left(\sqrt{\frac{1}{1+\eta\alpha}}\right)^{t-t'}\sqrt{B_\psi(z_*; z_{t'})}.
\end{aligned}$$

Where D is such that $\max_{z,z'\in\mathcal{Z}}\|z - z'\| \leqslant D$, and $C = \|G(z_*)\|_* + \tilde{L}D$. The first inequality is by the generalized Cauchy-Schwarz inequality and the Lipschitz property of G. The second inequality is by boundedness of $\mathcal{Z}$. The third inequality is by the fact that $B_\psi(z_*; z_t) \geqslant \frac{1}{2}\|z_* - z_t\|^2$. The fourth inequality is by applying Theorem 3.4 inductively. □

Note that we have the following well-known inequality between the saddle-point gap

$$\xi(x,y) = \max_{\bar{y}\in\mathcal{Y}} \alpha g_1(x) + f(x,\bar{y}) - \alpha g_2(\bar{y}) - \min_{\bar{x}\in\mathcal{X}} \alpha g_1(\bar{x}) + f(\bar{x}, y) - \alpha g_2(y)$$

and

$$\theta_{gap}(z) = \sup_{\bar{z}\in\mathcal{Z}}\langle F(z) + \alpha\nabla g(z), z - \bar{z}\rangle,$$

as shown below.

$$\xi(x,y) = \max_{\bar{y}\in\mathcal{Y}} \alpha g_1(x) + f(x,\bar{y}) - \alpha g_2(\bar{y}) - \min_{\bar{x}\in\mathcal{X}} \alpha g_1(\bar{x}) + f(\bar{x},y) - \alpha g_2(y)$$
$$= \alpha g_1(x) + f(x,y') - \alpha g_2(y') - (\alpha g_1(x) + f(x,y) - \alpha g_2(y))$$
$$+ (\alpha g_1(x) + f(x,y) - \alpha g_2(y)) - \big(\alpha g_1(x') + f(x',y) - \alpha g_2(y)\big) \text{ for some pair } (x',y') \in \mathcal{X}\times\mathcal{Y}$$
$$\leqslant \langle -\nabla f_y(x,y) + \alpha\nabla g_2(y), y - y'\rangle + \langle \nabla f_x(x,y) + \alpha\nabla g_1(x), x - x'\rangle$$
$$= \langle F(z) + \alpha\nabla g, z - z'\rangle \text{ for } z = (x,y) \text{ and } z' = (x',y')$$
$$\leqslant \theta_{gap}(z).$$

Therefore Proposition D.3.1 gives a guarantee on the saddle-point gap $\xi(x,y)$.

E Logit-AQREs and MMD

By Proposition D.2 the MMD update (5), restated below, has fixed points corresponding to the solutions of $\mathrm{VI}(\mathcal{Z}, F + \nabla\psi)$.

$$z_{t+1} = \arg\min_{z\in\mathcal{Z}} \eta\left(\langle F(z_t), z\rangle + \alpha\psi(z)\right) + B_\psi(z; z_t)$$

If $\mathcal{Z}$ is the cross-product of policy spaces for both players (cross product of sets of behavioral-form policies) and ψ is the sum of negative entropy over all decision points (information states), and F includes the the negative q-values for both players, then the iteration above reduces to

$$\pi_{t+1}(s) \propto \left[\pi_t(s) e^{\eta q_{\pi_t}(s)}\right]^{1/(1+\eta\alpha)}.$$

With fixed points corresponding to

$$\forall i, \forall h_i, \pi_i(h_i) \in \arg\max_{\pi'_i(h_i)} \mathbb{E}_{A\sim\pi'_i(h_i)}\left[q_\pi(h_i, A) + \alpha\mathcal{H}(\pi'_i(h_i))\right].$$

Or equivalenty, the solution to $\mathrm{VI}(\mathcal{Z}, F + \nabla\psi)$, which corresponds to a logit-AQRE.

F Experimental Domains

For our experiments with normal-form games, we used Diplomacy stage games. Diplomacy is a seven-player Markov game in which players compete to conquer Europe. Because the game is a Markov game (which means that the game is fully observable but that the players move simultaneously), each turn of the game resembles a normal-form game. We constructed the normal-form games that we used for our experiments by querying an open source value function [3] in different circumstances for a two-player, no-press (no linguistic communication) variant of the game, similarly to Zhang et al. [67]. These games have payoff matrices of shape $(50, 50)$ (game A), $(35, 43)$ (game B), $(50, 50)$ (game C), and $(4, 4)$ (game D). We normalized the payoffs of each game to $[0, 1]$.

For our extensive-form games, we used the implementations of Kuhn poker, 2x2 (and also 3x3) Abrupt Dark Hex, 4-Sided Liar's Dice, and Leduc Poker provided by OpenSpiel [33]. Kuhn poker [30] is a simplified poker game with three cards (J, Q, K). It has 54 non-terminal histories (not counting chance nodes). Abrupt Dark Hex is a variant of the classical board game Hex [4]. In Hex, two players take turns placing stones onto a board. One player's goal is to create a path of its stones connecting the east end of the board with the west end, while the other player's goal is to do the same with the north end and south end. Dark Hex is a variant in which players cannot see where their opponents are placing stones. Abrupt Dark Hex is a variant of Dark Hex in which placing a stone in an occupied position results in a loss of turn. The prefix nxn describes the size of the board. 2x2 Abrupt Dark Hex has 471 non-terminal histories. 3x3 Abrupt Dark Hex has too many non-terminal histories to enumerate on our hardware. Liar's dice [19] is a dice game in which players privately roll dice and place bids based on the observed outcomes, similarly to poker games. The prefix n-sided means that the players play with 4-sided dice. 4-Sided Liar's Dice has 8176 non-terminal histories (not counting chance nodes). Leduc Poker [57] is a small poker game with three card values (J, Q, K), each of which have two instances in the deck. It has 9300 non-terminal histories non-terminal histories (not counting chance nodes).

For our single-agent deep RL experiments, we use three Atari games [10] and three Mujoco games [59]. We selected these games because Huang et al. [25] used them to benchmark an open source implementation of PPO.

G QRE Experiments

G.1 Full Feedback QRE Convergence Diplomacy

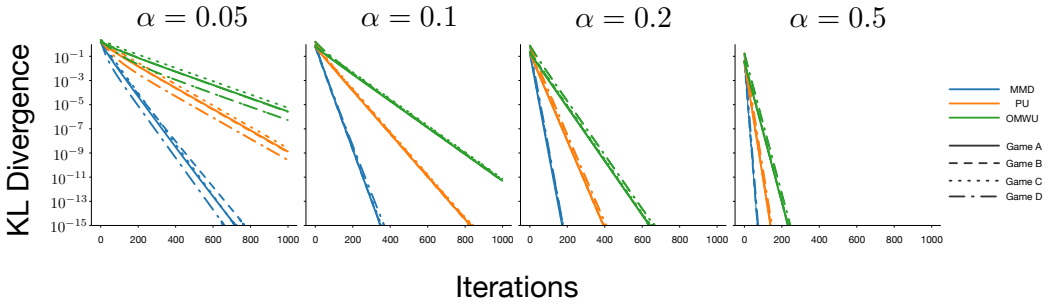

Figure 4: Solving for normal-form QREs in Diplomacy stage games with full feedback.

We perform various QRE experiments under full feedback for Diplomacy stage games. In this context, full feedback means that each player outputs a fully specified policy and receives its exact Q-values (given both players' policies) as feedback. Both players then perform the update

$$\pi_{t+1}(s) \propto [\pi_t(s) e^{\eta q_t(s)}]^{1/(1+\eta\alpha)}.$$

For our experiments, we set $\eta = \alpha$ (for each α) for MMD, which is the maximal value that retains a linear convergence guarantee for normal-form games with a max payoff magnitude of one. For PU and OMWU [13], we also used the maximal values that guarantee linear convergence. We solved for the QRE for each game using Ling et al.'s Newton's method approach. We show iterations on the x-axis and $\mathrm{KL}(\text{solution}, \text{iterate})$ on the y-axis. We count each query to the oracle as an iterate, meaning that OMWU uses two iterates for every update (contrasting MMD and PU, which only use one).

The results of the experiment, found in Figure 4, show that all three algorithms converge linearly with faster rates for larger values of alpha, as is guaranteed by theory. We find that, for our Diplomacy games, MMD converges faster than PU and OMWU (with all algorithms using their maximal theoretically allowed stepsizes). However, we note that all three algorithms also exhibited faster convergence with larger than theoretically allowed stepsizes.

G.2 Black Box QRE Convergence Diplomacy

Our second set of experiments examine convergence to QREs for our Diplomacy stage games with black box feedback. In this context, black box feedback means that each player i outputs an action A_i sampled from its current policy and that player i receives $\mathcal{R}(\cdot, A_i, A_{-i})$ (but not A_{-i}) as feedback. One way to approach such a setting is to construct an unbiased estimate of the exact Q-values. Letting r be the observed reward

$$\hat{q}_i^t(a_i) = \begin{cases} r/\pi_i^t(a_i) & \text{if } A_i = a_i \\ 0 & \text{otherwise} \end{cases}$$

is such an estimate. To see that this is true, observe

$$\begin{aligned}
\mathbb{E}[\hat{q}_i^t(a_i) \mid \pi^t] &= \mathbb{E}_{A_{-i} \sim \pi_{-i}^t} \left[\pi_i^t(a_i) \cdot \frac{\mathcal{R}(\cdot, a_i, A_{-i})}{\pi_i^t(a_i)} + \sum_{a_i' \neq a_i} \pi_i^t(a_i') \cdot 0 \right] \\
&= \mathbb{E}_{A_{-i} \sim \pi_{-i}^t} \left[\pi_i^t(a_i) \cdot \frac{\mathcal{R}(\cdot, a_i, A_{-i})}{\pi_i^t(a_i)} \right] \\
&= \mathbb{E}_{A_{-i} \sim \pi_{-i}^t} \mathcal{R}(\cdot, a_i, A_{-i}) \\
&= q_i^t(a_i).
\end{aligned}$$

In Figure 5, we show results for each of MMD, PU and OMWU, with the exact Q-values q_t replaced by the unbiased estimates $\hat{q}_t$. For each algorithm, the stepsize at iteration t was set to be equal to the maximal step size for which there exists an exponential convergence guarantee divided by $10\sqrt{t}$. In other words,

$$\eta_t = \frac{\eta}{10\sqrt{t}}.$$

Each line is an average over 30 runs. The bands depict estimates of 95% confidence intervals computed using bootstrapping. Although none of the algorithms possess existing black box convergence guarantees, we observe that they all exhibit convergent behavior empirically. In terms of convergence speed, we observe that MMD compares favorably to PU and OMWU for $\alpha \in \{0.05, 0.1, 0.2\}$; however, for $\alpha = 0.5$, OMWU performed the best, with the exception of game D. It is likely that all algorithms could achieve better performance, as we did not perform much hyperparameter tuning.

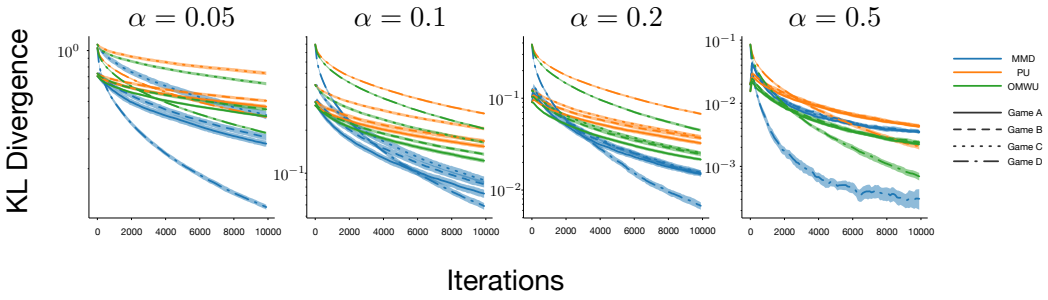

Figure 5: MMD, PU, and OMWU applied to Diplomacy stage games for QRE finding with black box sampling.

We also investigate the performance of other methods for estimating Q-values for the black box setting. One such method uses an unbiased baseline to reduce variance [54, 14]. The premise of this approach is the idea that any quantity that is zero in expectation can be subtracted from an unbiased Q-value estimate without introducing bias. As a result, if the quantity is correlated with the estimator, subtracting it from the estimate can reduce variance "for free". We call this quantity a baseline. For our baseline, we used

$$b_i^t(a_i) = \begin{cases} \tilde{q}_i^t(a_i)/\pi_i^t(a_i) - \tilde{q}_i^t(a_i) & \text{if } a_i = A_i \\ -\tilde{q}_i^t(a_i) & \text{otherwise.} \end{cases}$$

By a similar argument as above, this quantity is zero in expectation

$$\begin{aligned} \mathbb{E}[b_i^t(a_i) \mid \pi^t] &= \pi_i^t(a_i) \cdot (\tilde{q}_i^t(a_i)/\pi_i^t(a_i) - \tilde{q}_i^t(a_i)) - \sum_{a_i' \neq a_i} \pi_i^t(a_i') \cdot \tilde{q}_i^t(a_i) \\ &= \tilde{q}_i^t(a_i) - \pi_i^t(a_i)\tilde{q}_i^t(a_i) - (1 - \pi_i^t(a_i)) \cdot \tilde{q}_i^t(a_i) \\ &= \tilde{q}_i^t(a_i) - \tilde{q}_i^t(a_i) \\ &= 0. \end{aligned}$$

Also, if $\tilde{q}$ is close to q, our baseline will be correlated with $\hat{q}$. Thus, it satisfies our desired criteria. For $\tilde{q}$, we used a running estimate of the reward observed after selecting action a_i. Specifically, every time action a_i was selected, we updated

$$\tilde{q}_i^t(a_i) = (1 - \tilde{\eta})\tilde{q}_i^t(a_i) + \tilde{\eta} r.$$

We used $\tilde{\eta} = 1/2$, inspired by Schmid et al. [54].

We also investigated the use of *biased* Q-value estimates, as this is the setting that corresponds with function approximation. For this approach, we plugged in $\tilde{q}$, as computed above, instead of the exact Q-values.

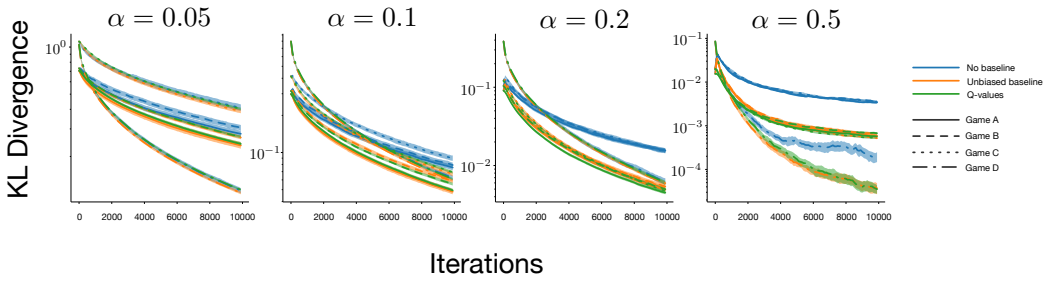

Figure 6: MMD with no baseline, an unbiased baseline, and (biased) Q-values applied to diplomacy stage games for QRE finding with black box sampling.

We show the results of the experiment if Figure 6. The column shows the temperature for the QRE. The y-axis shows the KL divergence to the corresponding logit-QRE. The x-axis shows the number of iterations. For each algorithm, the step size at iteration t was set to be equal to the maximal step size for which there exists an exponential convergence guarantee divided by $10\sqrt{t}$. Each line is an average over 30 runs. The bands depict estimates of 95% confidence intervals computed using bootstrapping. Overall, we find that both using unbiased baselines and biased Q-value estimates appears to improve convergence speed.

G.3 Full Feedback QRE Convergence EFGs

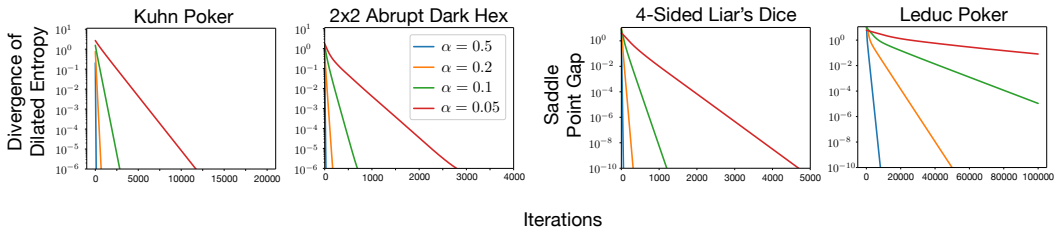

Figure 7: Solving for normal-form QREs in EFGs.

We perform several experiments for solving reduced normal-form logit QREs by using MMD over the sequence form with dilated entropy. We use the descent-ascent updates

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \eta \left(\langle \nabla_x f(x_t, y_t), x \rangle + \alpha \psi_1(x) \right) + B_{\psi_1}(x; x_t),$$
$$y_{t+1} = \arg\max_{y \in \mathcal{Y}} \eta \left(\langle \nabla_y f(x_t, y_t), y \rangle - \alpha \psi_2(y) \right) - B_{\psi_2}(y; y_t).$$

The method is full feedback since $\nabla_x f(x_t, y_t) = A y_t$ and $\nabla_y f(x_t, y_t) = A^\top x_t$, where A is the sequence form payoff matrix. Note in the normal form setting $-Ay_t$ and $A^\top x_t$ are the Q-values for both players and the algorithm is the same as described in Section G.1. We set the stepsize to be $\eta = \frac{\alpha}{U^2}$ where U is the maximal absolute payoff of the game. Note that $U \geqslant \max_{ij} |A_{ij}|$ and so the stepsize condition of Theorem 3.4 is satisfied. For more details on the sequence form algorithm see Section C.3.

For Kuhn Poker and 2x2 Abrubt Dark Hex we used Gambit [43, 63] to compute the reduced normal-form QRE. We then check the convergence of MMD by plotting the sum of Bregman divergences with respect to dilated entropy $B_\psi(z_*; z_t) = B_{\psi_1}(x_*; x_t) + B_{\psi_1}(y_*; y_t)$, with respect to the solution $z_* = (x_*, y_*)$. As predicted by Theorem 3.4 we observer linear convergence with faster convergence for larger values of alpha.

For 4-Sided Liar's Dice and Leduc Poker, the games were too large for Gambit [43, 63] to compute the reduced normal-form QRE in reasonable amount of time or memory. Therefore, we check the convergence of MMD by plotting the saddle point gap $\xi(x_t, y_t)$ of the min max problem given by

Ling et al. [37],

$$\xi(x_t, y_t) = \max_{\bar{y} \in \mathcal{Y}} \alpha\psi_1(x_t) + x_t^\top A\bar{y} - \alpha\psi_2(\bar{y}) - \min_{\bar{x} \in \mathcal{X}} \alpha\psi_1(\bar{x}) + \bar{x}^\top A y_t - \alpha\psi_2(y_t).$$

Theorem 3.4 guarantees that the gap will converge to zero. Note the gap is zero if and only if at the solution, and by Proposition D.3.1 the gap is also guaranteed to converge linearly. In both 4-Sided Liar's Dice and Leduc Poker we observe linear convergence of the gap, with faster convergence for larger values of alpha. Additionally, due to the $O(\log(t))$ regret bound from Duchi et al. [15], we have that the gap is guaranteed to converge at a rate of $O\left(\frac{\log(t)}{t}\right)$ for the average iterates of both players.

G.4 Full Feedback AQRE Convergence EFGs

Next, we investigate whether MMD can be made to converge to AQREs in extensive-form games. For these experiments we applied MMD in behavioral form, as described in Section E. Specifically, we computed $q_i^t(h_i)$ for each player i and each information state h_i. Then, we applied the update rule

$$\pi_i^{t+1}(h_i) \propto [\pi_i^t(h_i) e^{\eta_t q_i^t(h_i)}]^{1/(1+\eta_t \alpha_t)}.$$

for each player i and information state h_i. To induce convergence, we found it helpful to use a schedule of $\{(\alpha_t, \eta_t)\}_t$, rather than fixed values.

For Kuhn poker, we used

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \max(\frac{1}{\sqrt{t}}, \alpha).$$

For 2x2 Dark Hex, we used

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \max(\frac{1}{\sqrt{t}}, \alpha).$$

For 4-Sided Liar's dice, we used

$$\eta_t = \frac{2}{\sqrt{t}}, \alpha_t = \max(\frac{1}{\sqrt{t}}, \alpha).$$

For Leduc Poker, we used

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \max(\frac{5}{\sqrt{t}}, \alpha).$$

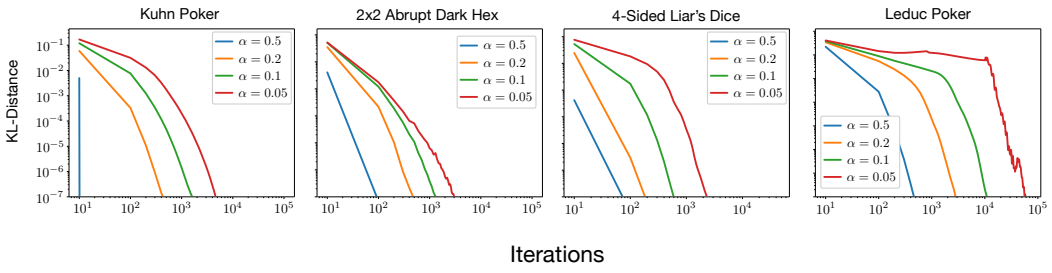

Figure 8: Solving for AQREs in EFGs.

We show the results in Figure 8. We measure convergence against solutions computed using Gambit [43, 64]. Despite a lack of proven convergence guarantees, we observe that MMD converges to the AQRE in each game, for each temperature.

H Nash Equilibria Experiments

Next, we investigate the convergence of MMD as a Nash equilibrium solver. To induce convergence, we anneal the temperature over time, as was done in the AQRE experiments.

H.1 Full Feedback Nash Convergence Diplomacy

In our full feedback Nash convergence Diplomacy experiments, we used

$$\eta = \frac{1}{10}, \alpha_t = \frac{1}{20\sqrt{t}}.$$

We show the results of the experiment in Figure 9. Over short iteration horizons, we observe that CFR tends to outperform MMD. However, for longer horizons, we find that MMD tends to catch up with CFR. In game D, the qualitatively different behavior is likely to due the fact that the Nash equilibrium is a pure strategy, unlike the Nash equilibria of the first three games, which are mixed.

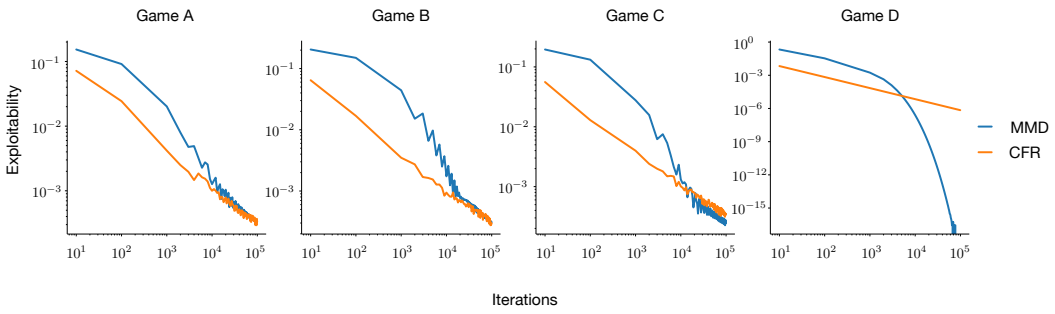

Figure 9: MMD and CFR applied to diplomacy stage games for computing Nash equilibria.

H.2 Black Box Nash Convergence Diplomacy

For our black box Nash convergence experiments, we compare against the "opponent on-policy" variant of Monte Carlo CFR [31]. In this variant, the two players alternate between an updating player and an on-policy player. The updating player plays off-policy according to a policy that provides sufficiently large support to each action (in our Diplomacy experiments we used a uniform policy). The advantage to this setup is that it guarantees that the updating player will receive bounded gradients, which is necessary for Monte Carlo CFR's convergence proof. In contrast, we show results for an on-policy Monte Carlo variant of MMD, despite the fact that this causes unbounded gradients. This is not a fair comparison in the sense that the same "opponent on-policy" setup is equally applicable to MMD and would keep the gradients bounded, whereas the "on-policy" version of Monte Carlo CFR does not converge. We made this decision because the on-policy Monte Carlo variant of MMD is simpler and corresponds more closely with the most straightforward way to scale MMD to deep multi-agent reinforcement learning. Nevertheless, we believe that the "opponent on-policy" version of MMD remains an interesting direction for future, and would very possibly yield faster convergence.

We again investigated three ways of estimating Q-values. For our unbiased estimator with no baseline we used

$$\eta_t = \frac{1}{5\sqrt{t}}, \alpha_t = \frac{20}{\sqrt{t}}$$

for games A, B, and C, and

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \frac{1}{\sqrt{t}}$$

for game D. For our unbiased estimator with baseline, we used

$$\eta_t = \frac{1}{10\sqrt{t}}, \alpha = \frac{10}{\sqrt{t}}$$

for games A, B, and C, and

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \frac{1}{\sqrt{t}}$$

for game D. For our biased estimator, we used

$$\eta_t = \frac{1}{5\sqrt{t}}, \alpha_t = \frac{2}{\sqrt{t}}$$

for games A, B, and C, and

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \frac{1}{\sqrt{t}}$$

for game D. We show results in Figure 12, with averages across 30 runs and estimates of 95% confidence intervals computed from bootstrapping.

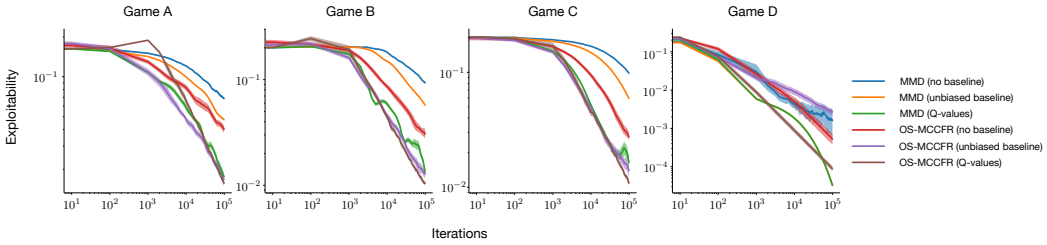

Figure 10: MMD and OS-MCCFR applied to diplomacy stage games for computing Nash equilibria with black box sampling.

For MMD, we observe that Q-values generally perform best, followed by an unbiased estimate with baseline, followed by an unbiased estimate without baseline, except on game D, where the unbiased baseline performs similarly to Q-values. We also find that CFR tends to follow this trend, though the difference between Q-values and an unbiased baseline is less pronounced, except on game D, where the unbiased baseline performs poorly. Between MMD and CFR, CFR tends to perform better on an estimator to estimator basis in games A, B and C, though MMD is relatively competitive with CFR for Q-values. For game D we observe that this comparison is more favorable for MMD than the other games.

H.3 Full Feedback Nash Convergence EFGs

For our full feedback Nash convergence EFG experiments we examined two variants of MMD. The first, which we call unweighted MMD, corresponds with the version tested in the AQRE experiments

$$\pi_i^{t+1}(h_i) \propto (\pi_i^t(h_i) e^{\eta_t q_i^t(h_i)})^{1/(1+\eta\alpha)}.$$

The second, which we call weighted MMD,

$$\pi_i^{t+1}(h_i) \propto (\pi_i^t(h_i) e^{\mathcal{P}^{\pi_t}(h_i)\eta_t q_i^t(h_i)})^{1/(1+\mathcal{P}^{\pi_t}(h_i)\eta\alpha)}.$$

In other words, it weights the stepsize of the update by the probability of reaching that information state under the current policy. We test this variant because it corresponds with a "determinized" version of black box sampling for temporally extended settings.

For unweighted MMD, we used

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \frac{1}{\sqrt{t}}$$

for Kuhn Poker,

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \frac{1}{\sqrt{t}},$$

for 2x2 Dark Hex,

$$\eta_t = \frac{2}{\sqrt{t}}, \alpha_t = \frac{1}{\sqrt{t}},$$

for 4-Sided Liar's dice, and

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \frac{5}{\sqrt{t}}$$

for Leduc Poker.

For weighted MMD, we used

$$\eta_t = \frac{2}{\sqrt{t}}, \alpha_t = \frac{2}{\sqrt{t}}$$

for Kuhn Poker,

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \frac{1}{\sqrt{t}}$$

for 2x2 Abrupt Dark Hex,

$$\eta_t = \frac{100}{\sqrt{t}}, \alpha_t = \frac{2}{\sqrt{t}}$$

for 4-Sided Liar's Dice, and

$$\eta_t = \frac{500}{\sqrt{t}}, \alpha_t = \frac{10}{\sqrt{t}}$$

for Leduc Poker. Note that larger stepsize values are required for MMD for large games to achieve competitive because, otherwise, the reach probability make updates at the bottom of the tree very small.

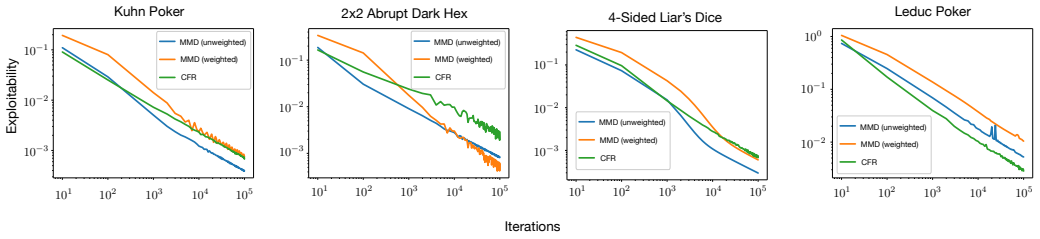

Figure 11: MMD (unweighted) and MMD (weighted by reach probability) compared against CFR across standard OpenSpiel games.

We show the results of our experiments in Figure 11. We find that both weighted MMD and unweighted MMD exhibit convergent behavior. Furthermore, they converge at rates comparable with CFR on average across the games.

H.4 Black Box Nash Convergence EFGs

For our black box Nash convergence EFG experiments, we used the Monte Carlo CFR implementation in OpenSpiel [33], which uses an update policy with a 0.4 weight on the current policy and a 0.6 weight on the uniform policy. For MMD, we used the sampling version of weighted MMD, meaning that the information states touched during the trajectory are updated with the full stepsize, while information not touched during the trajectory are not updated. For Kuhn Poker, we used

$$\eta_t = \frac{1}{10\sqrt{t}}, \alpha_t = \frac{50}{\sqrt{t}}.$$

For 2x2 Abrupt Dark Hex, we used

$$\eta_t = \frac{7}{20\sqrt{t}}, \alpha = \frac{50}{\sqrt{t}}.$$

For 4-Sided Liar's Dice, we used

$$\eta_t = \frac{2}{\sqrt{t}}, \alpha_t = \frac{200}{\sqrt{t}}.$$

For Leduc Poker, we used

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \frac{300}{\sqrt{t}}.$$

Noting again that the caveats about comparing on-policy MMD to opponent on-policy Monte Carlo CFR also apply here, we present the results in Figure 11. Results are averaged across 30 runs and shown with 95% confidence intervals estimated from bootstrapping. As in the normal-form experiments, we find that Monte Carlo CFR generally outperforms MMD for unbiased gradient estimates with no baseline.

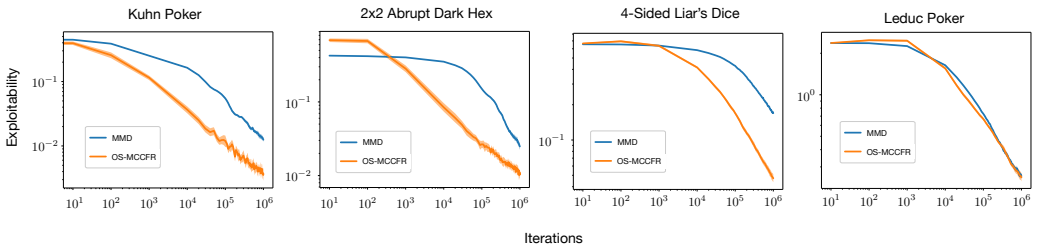

Figure 12: MMD compared against OS-MCCFR across standard OpenSpiel games with black box sampling.

H.5 Moving Magnet

Next, we investigate using a moving magnet, rather than an annealing temperature, to induce convergence to a Nash equilibrium. In the moving magnet setup, updates take the form

$$\pi_i^{t+1}(h_i) \propto [\pi_i^t(h_i)\rho_i^t(h_i)^{\eta\alpha} e^{\eta q_{\pi^t}(h_i)}]^{1/(1+\eta\alpha)},$$

where ρ^t slowly trails behind π^t. In our experiment, we used

$$\rho_i^t(h_i) = (1 - 1e{-}5)\rho_i^t(h_i) + 1e{-}5\pi_i^t(h_i)$$

and

$$\alpha = 1, \eta = 0.1.$$

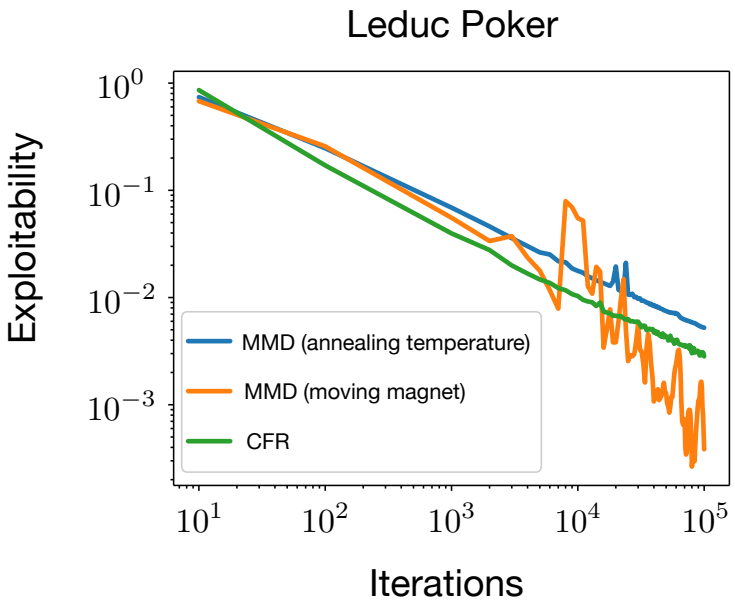

Figure 13: Comparing a moving magnet to an annealing temperature.

We show the results in Figure 13, compared against CFR and MMD with an annealing temperature (with the same hyperparameters as before). Encouragingly, we find that that moving the magnet behind the current iterate also appears to encourage convergence. However, while the convergence may even be faster than with an annealing temperature, it is also less stable in terms of exploitability. We observed similar phenomenon in other preliminary experiments with moving magnets. This suggests that moving the magnet is a promising direction for future work, though it is not clear that doing it as we do here is the "right way" to do so.

H.6 Other Regularization

Lastly, we examine the convergence properties of other types of regularization. We examine four different alternatives. One is a state entropy reward bonus, wherein $\alpha\mathcal{H}(\pi_i^t(h_i)$ is added to the reward for player i for reaching information state h_i. This corresponds with a maximum entropy objective in reinforcement learning [68]. A second is an action reward bonus, wherein $\alpha \log \pi_i^t(h_i, a_i)$ is added to the reward for agent i for playing action a_i in information state h_i. The third is a simultaneously giving a state entropy reward bonus (as in the first), while also penalizing the player with the opponent's state entropy. This third approach can be viewed as a modification of the first approach that makes the game zero-sum. The last approach is simultaneously giving an action bonus (as in the second), while also penalizing the player with the opponent's action bonuses. This fourth approach can be viewed as a modification of the third approach that makes the game zero sum. The fourth approach is the kind of regularization that was examined in Pérolat et al. [50].

For each algorithm, we used

$$\eta_t = \frac{1}{\sqrt{t}}, \alpha_t = \frac{5}{\sqrt{t}}.$$

We show the results in Figure 14. We find that all four alternative kinds of regularization exhibit convergent behavior. While MMD (no bonus) is preferable to these alternatives in that it is less burdensome (the alternatives require reward modification, and penalties prohibit decentralization), it is interesting from a scientific perspective to see that they all exhibit convergent behavior.

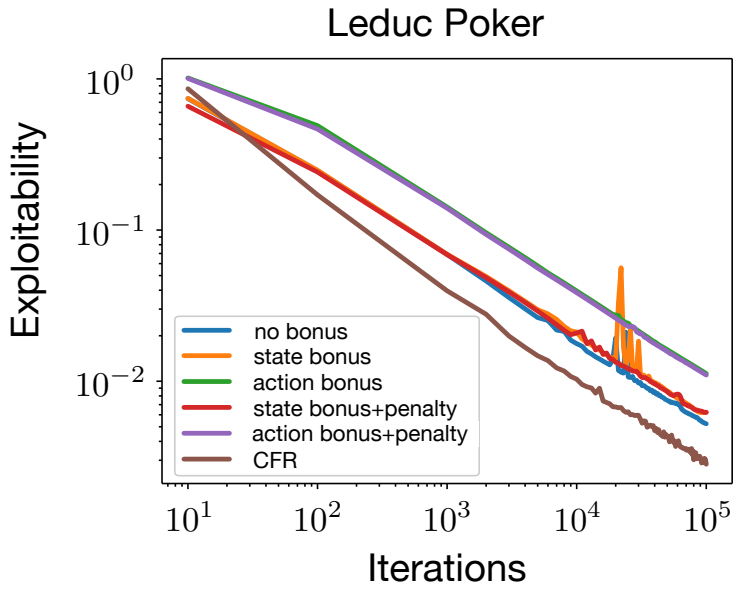

Figure 14: Comparing different kinds of regularization.

I Atari and Mujoco Experiments

For our single-agent deep RL experiments, we implemented MMD as a modification to Huang et al.'s implementation of PPO. For Atari, we added a reverse KL penalty with a coefficient $1/\eta = 0.001$; we kept the temperature as the default value set by Huang et al. [25] ($\alpha = 0.01$). For Mujoco, we added a reverse KL penalty with a coefficient $1/\eta = 0.1$; we added an entropy bonus (Huang et al. [25] do not use an entropy bonus) with a value of $\alpha = 0.0001$. Otherwise, for both Atari and Mujoco, the hyperparameters were set to those selected by Huang et al. [25]. We show the results again in Table 4 for convenience. The baseline results for PPO are copied directly from Huang et al. [25]. The exact numbers should be interpreted somewhat cautiously as they are averaged over only three runs, leaving high levels of uncertainty. That said, the results in Table 4 still provide evidence that MMD can perform comparably to PPO. But even without looking at empirical results, the idea that

a deep form of MMD can perform comparably to PPO should not be surprising, as MMD can be implemented in a way that resembles PPO in many aspects.

Table 4: Atari and Mujoco results averaged over 3 runs, with standard errors.

	Breakout	Pong	BeamRider	Hopper-v2	Walker2d-v2	HalfCheetah-v2
PPO	409 ± 31	20.59 ± 0.40	2628 ± 626	2448 ± 596	3142 ± 982	2149 ± 1166
MMD	414 ± 6	21.0 ± 0.00	2549 ± 524	2898 ± 544	2215 ± 840	3638 ± 782

J 3x3 Abrupt Dark Hex Experiments

For our 3x3 Abrupt Dark Hex experiments, we implemented MMD as a modification to PPO, as implemented by RLlib [35]. This involved modifying the loss to add entropy regularization, as well a changing the adaptive forward KL regularization to a constant reverse KL regularization. We used a schedule

$$\eta_t = \frac{50}{\sqrt{t/10}}, \alpha_t = \frac{50}{\sqrt{t/10}},$$

where t is the number of time steps (not the number of episodes!). Otherwise, we used the default hyperparameters. We ran this implementation in self-play using RLlib's OpenSpiel environment wrapper, modified to work with information states, rather than observations. For NFSP [23], we used the same hyperparameters as those found in the NFSP Leduc example in the OpenSpiel codebase. For the best response, we used the OpenSpiel's DQN best response code, without modifying any hyperparameters. We ran the best response for 10 million time steps and evaluated all match-ups over 2000 games (with each agent being the first-moving player in 1000).

There are two caveats to consider in interpreting these experiments. First, it is likely that RLlib's default PPO hyperparameters are generally stronger than the default hyperparameters for NFSP in the OpenSpiel. In this respect, the results we present may be unfair to NFSP. Second, RLlib's OpenSpiel wrapper does endow agents with knowledge about which actions are legal—instead, if an illegal action is selected, the agent is given a small penalty and a random legal action is executed. In contrast, OpenSpiel's implementation of NFSP uses information about the legal actions to perform masking. In other words, MMD faces a harder version of the game than NFSP faces. In this respect, the results we present are unfair to MMD.

Table 5: Approximate exploitability for 3x3 Abrupt Dark Hex in units of 10^{-2}.

	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5
First Legal Action Taker	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Uniform Random	74 ± 1	76 ± 1	75 ± 1	73 ± 2	74 ± 2
NFSP(1M steps)	97 ± 1	90 ± 1	97 ± 1	96 ± 1	75 ± 1
NFSP(10M steps)	61 ± 2	61 ± 2	60 ± 2	58 ± 2	57 ± 2
MMD(1M steps)	39 ± 2	36 ± 2	38 ± 2	36 ± 2	61 ± 2
MMD(10M steps)	23 ± 2	24 ± 2	24 ± 2	23 ± 2	20 ± 2

We show the results of our approximate exploitability experiments again in Table 5 for convenience. The agents we evaluated are a bot that always takes the first legal action, a bot that takes actions uniformly at random, NFSP after 1 million time steps, NFSP after 10 million time steps, MMD after 1 million time steps, and MMD and 10 million time steps. Note that there are three kinds of uncertainty involved here: 1) uncertainty over training the agent (captured in table by the seed), 2) uncertainty over training the best response (also captured in the table by the seed), and 3) uncertainty in evaluating the match-up (captured in the table by the standard error).

First, we observe that there is a substantial difference in approximate exploitability between playing totally randomly and playing deterministically—the uniform random bot's approximate exploitability ranged from 0.74 to 0.76, while the first action taking bot had full exploitability for every seed. Second, we observe that NFSP appears to be significantly slower to converge than MMD. After 1 million time steps, none of the five NFSP seeds have approximate exploitability that outperforms playing uniformly randomly in terms of approximate exploitability. In contrast, with the same budget,

all five of the MMD seeds outperform uniform random play. After 10 million time steps, NFSP convincingly outperforms uniform random play, but generally does not even match the performance of MMD(1M steps). MMD(10M steps) yields the best performance, with approximate exploitability ranging from 0.20 to 0.24. In terms of winning percentage, this means the exploiter defeated MMD(10M steps) between 60% and 62% of the time.

Table 6: Head-to-head expected return for row player in 3x3 Abrupt Dark Hex in units of 10^{-2}.

	First Taker	Uniform	NFSP1	NFSP2	NFSP3	NFSP4	NFSP5
First Taker	0	0 ± 2	-81 ± 1	-36 ± 2	-38 ± 2	-40 ± 2	-40 ± 2
Uniform	0 ± 2	0	-38 ± 2	-46 ± 2	-46 ± 2	-39 ± 2	-45 ± 2
MMD1	62 ± 2	66 ± 2	26 ± 2	20 ± 2	33 ± 2	38 ± 2	11 ± 2
MMD2	64 ± 2	64 ± 2	25 ± 2	22 ± 2	34 ± 2	28 ± 2	14 ± 2
MMD3	64 ± 2	70 ± 2	27 ± 2	25 ± 2	37 ± 2	37 ± 2	8 ± 2
MMD4	61 ± 2	63 ± 2	23 ± 2	16 ± 2	37 ± 2	26 ± 2	9 ± 2
MMD5	61 ± 2	63 ± 2	24 ± 2	9 ± 2	24 ± 2	18 ± 2	8 ± 2

We show results of head-to-head match-ups in Table 6 for convenience. The agents we evaluated are the ones that were trained for 10 million time steps. Again, note that there are three uncertainties involved: training each of the two agents and evaluating the match-up. The former is captured by the agent seed, while the latter is captured by the standard error.

First, we observe that, despite the large difference in approximate exploitability between always taking the first legal action and playing uniformly at random, the two bots are relatively evenly matched in head-to-head games. Furthermore, they generally achieve similar expected returns against the MMD and NFSP agents. This observation reflects the significant difference between worst case performance and average case performance. In general, we find that MMD exploits the first legal action taking bot and the uniform random bot more than NFSP: MMD's performance ranges from 0.62 to 0.70, while, with the exception of one seed, NFSP's performance ranges from 0.36 to 0.46. However, NFSP also has a seed with outlyingly strong performance against the first action taking bot that achieves an expected return of 0.81. In terms of head-to-head performance against each other, we found that MMD yielded stronger performance than NFSP—MMD's expected return ranged from 8 to 38. In terms of winning percentage, this means that MMD won between 54% of the games and 69% of the games against NFSP.

K Relationship to KL-PPO and MDPO

On the single-agent deep reinforcement learning side, MMD most closely resembles KL-PPO [55] and MDPO [60]. KL-PPO uses the policy loss function

$$\mathbb{E}_t \left[\frac{\pi(a_t \mid s_t)}{\pi_{\text{old}}(a_t \mid s_t)} \hat{A}_t + \alpha \mathcal{H}(\pi(s_t)) - \beta \text{KL}(\pi_{\text{old}}(s_t), \pi(s_t)) \right],$$

where $\hat{A}_t$ is an advantage function (a learned estimate of $q_{\pi_t}(s_t) - v_{\pi_t}(s_t)$). In expectation, the first term acts as $\langle \pi_t(s_t), q_{\pi_t}(s_t) \rangle$, which is the first term of MMD's loss function. The second term is the same entropy bonus as exists in MMD's loss function. However, unlike MMD, KL-PPO's KL regularization goes forward $\text{KL}(\pi_{\text{old}}(s_t), \pi(s_t))$. In contrast, MMD's KL regularization goes backward $\text{KL}(\pi(s_t), \pi_{\text{old}}(s_t))$.

MDPO uses the policy loss function

$$\mathbb{E}_t \left[\frac{\pi(a_t \mid s_t)}{\pi_{\text{old}}(a_t \mid s_t)} \hat{A}_t - \beta \text{KL}(\pi(s_t), \pi_{\text{old}}(s_t)) \right],$$

where $\hat{A}_t$ is the approximate advantage function (a learned estimate of $q_{\pi_t}(s_t) - v_{\pi_t}(s_t)$). Just as mirror descent is MMD without the additional proximal regularization, MDPO is MMD without the additional proximal regularization. In the context of a negative entropy distance generating function and a uniform magnet, MDPO differs from MMD in that it does not include an entropy regularization term $\alpha \mathcal{H}(\pi(s_t))$.

For contrast, in this notation, MMD with a negative entropy distance generating function and a uniform magnet takes the form

$$\mathbb{E}_t \left[\frac{\pi(a_t \mid s_t)}{\pi_{\text{old}}(a_t \mid s_t)} \hat{A}_t + \alpha \mathcal{H}(\pi(s_t)) - \beta \text{KL}(\pi(s_t), \pi_{\text{old}}(s_t)) \right],$$

where β acts as an inverse stepsize.