Uncoupled Learning of Differential Stackelberg Equilibria with Commitments

Robert Loftin
Delft University of Technology
Delft, The Netherlands
r.t.loftin@tudelft.nl

Mustafa Mert Çelikok
Aalto University
Espoo, Finland
mustafamert.celikok@aalto.fi

Herke van Hoof
University of Amsterdam
Amsterdam, The Netherlands
h.c.vanhoof@uva.nl

Samuel Kaski
Aalto University, Espoo, Finland
University of Manchester, Manchester
United Kingdom
samuel.kaski@aalto.fi

Frans A. Oliehoek
Delft University of Technology
Delft, The Netherlands
f.a.oliehoek@tudelft.nl

ABSTRACT

A natural solution concept for many multiagent settings is the Stackelberg equilibrium, under which a "leader" agent selects a strategy that maximizes its own payoff assuming the "follower" agent chooses their best response to this strategy. Recent work has presented asymmetric learning updates that can be shown to converge to the *differential* Stackelberg equilibria of two-player differentiable games. These updates are "coupled" in the sense that the leader requires some information about the follower's payoff function. Such coupled learning rules cannot be applied to *ad hoc* interactive learning settings, and can be computationally impractical even in centralized training settings where the follower's payoffs are known. In this work, we present an "uncoupled" learning process under which each player's learning update only depends on their observations of the other's behavior. We prove that this process converges to a local Stackelberg equilibrium under similar conditions as previous coupled methods. We conclude with a discussion of the potential applications of our approach to human–AI cooperation and multi-agent reinforcement learning.

KEYWORDS

Multiagent Learning, Ad Hoc Cooperation, Game Theory

1 INTRODUCTION

A central goal of multiagent systems research has been to understand the long-term behavior of independent learning agents that optimize their individual strategies through repeated interaction. In the context of human–AI or AI–AI interaction, theoretical results on learning dynamics allow us to determine if and when the independent learners will converge to a fixed joint strategy, and characterize the strategies they are likely to converge to. Such results also inform the design of new learning algorithms that have desirable convergence properties. Recent years have seen a surge of interest in the dynamics of multiagent learning, driven by the recognition that many machine learning problems can be formulated as games with continuous, high-dimensional strategy spaces and differentiable payoff functions. Results on multiagent learning in such *differentiable* games have found application to reinforcement learning [33] and the training of generative adversarial networks [2].

Proc. of the Adaptive and Learning Agents Workshop (ALA 2023), Cruz, Hayes, Wang, Yates (eds.), May 29-30, 2023, Online, https://ala2023.github.io/. 2023.

In this work, we consider the problem of finding *hierarchical* solutions in two-player, general-sum differentiable games. Under the hierarchical model of play, one player (the "leader") selects their strategy first, after which the other player (the "follower") selects their best-response to this strategy. The natural solution concept for the hierarchical model of play is the *Stackelberg equilibrium*, in which the leader's strategy is optimal under the assumption that the follower will play their best response to any strategy the leader might choose. The hierarchical model is well suited to cooperative settings, where the leader can play their half of a jointly optimal strategy knowing that the follower will respond appropriately. It has also been argued [13] that the Stackelberg equilibrium is a more useful solution concept for differentiable games than the Nash equilibrium, as the Stackelberg equilibrium exists in games where the Nash equilibrium does not.

This fact has motivated the development of "hierarchical" gradient ascent methods for finding Stackelberg equilibria of differentiable games. As gradient ascent methods can only hope to find local optima of non-concave functions, these methods seek *local* Stackelberg equilbria (LSE). In particular, Fiez et al. [9] have presented a hierarchical gradient update that is shown to converge to LSE in certain differentiable games. Unfortunately, this *coupled* learning update requires complete knowledge of the follower's payoff function, and therefore cannot be applied to *ad hoc* learning settings, where the other agent's payoff function is unknown. The hierarchical update also requires the Hessian of the follower's payoff function, and so may be computationally intractable in settings where second-order derivatives are expensive to estimate (such as reinforcement learning).

The main contribution of this work is a novel *uncoupled* learning rule that estimates the leader's gradient update by sampling strategies close to the leader's current strategy, and then committing to these "perturbed" strategies long enough that the follower has time to adapt to them. This learning update, which we refer to as *Hierarchical learning with Commitments* (Hi-C), does not require that the leader has access to the follower's payoff function, or detailed knowledge of their learning process. As such, the Hi-C update is applicable to learning in the ad hoc setting, and to problems where estimating the higher-order derivatives of the payoff functions is impractical. Our main theoretical results show that Hi-C converges to a local Stackelberg solution for the leader under a generic assumption that the follower's own strategy converges to its best response sufficiently fast. We also provide specific convergence guarantees for the case where the follower's payoff function is strongly concave, and the follower updates their strategy using gradient ascent. We note that the conditions under which Hi-C converges with a gradient ascent partner are only slightly more restrictive than those required for the convergence of the exact hierarchical gradient update.

We describe the hierarchical gradient update and the local Stackelberg equilibrium in Section 3. We present the Hi-C learning update and our main convergence results in Section 4, while in Section 4.2 we provide concrete convergence guarantees for the special case where the follower's payoff function is strongly concave. We conclude with a discussion of the implications of our results for the problem of *ad hoc* human–AI cooperation.

2 RELATED WORK

In recent years, a large body of work has emerged on the problem of solving differentiable games using gradient ascent methods. It has been shown that simultaneous gradient ascent on individual payoff functions can fail to converge in such games [17, 18]. Convergence issues have led to the development of alternative solution concepts to the Nash equilibrium that are potentially better suited to differentiable games. These include chain recurrent sets [23] and local Stackelberg equilibria [13]. Other work has proposed modified gradient ascent approaches designed to achieve at least local convergence to fixed-points in certain classes of games [2, 18, 28]. Similar to our approach are methods for two-player games that update the individual strategies on two different timescales [16, 19, 22]. As with our approach, Nouiehed et al. [22] implement timescale separation by having the follower execute multiple gradient steps for every leader update, though unlike our work, their leader update does not directly attempt to *shape* the behavior of the follower.

Most closely related to our work is the two-timescale hierarchical gradient update [8, 33], discussed in more detail in Section 3, which has been shown to converge to local Stackelberg equilibria in zero-sum games. Unlike our method, the hierarchical gradient update requires that the leader have access to the follower's payoffs. Also closely related to our approach are methods that find Stackelberg equilibria by having a leader agent *commit* to a fixed strategy, and then observe the follower's response. Assuming the follower plays an immediate best-response, previous work has provided lower bounds on the sample complexity of identifying Stackelberg equilibria in Stackelberg security games [24], bandit games [1] and Markov games [26]. The challenge in our setting is that the we must assume the follower implements an incremental learning update, which may only converge to a best-response asymptotically.

Our work is also related to *opponent shaping* approaches [11, 32], where one or both learners explicitly account for their partner's learning behavior, and update their strategies accordingly. Of these methods, the model-free opponent shaping (M-FOS) framework of Lu et al. [15] is closest to our approach. The key differences from our method are that M-FOS only allows the follower to adapt for a fixed number of stages, and assumes that it can be "reset" after each learning interval. In contrast, we explicitly account for the fact that the follower's strategy depends on the entire history of interaction, and allow it to adapt over increasing time horizons, which enables asymptotic convergence. Finally, Hi-C is conceptually similar to no-regret learning methods for non-stationary tasks [7] and adaptive partners [25], in which the leader commits to candidate "expert" strategies for increasingly long time intervals.

3 PRELIMINARIES

We consider the class of two-player, general-sum differentiable games. Let $\mathcal{X} \subseteq \Re^{d_1}$ and $\mathcal{Y} \subseteq \Re^{d_2}$ denote the strategy spaces for players 1 and 2 respectively. When discussing hierarchical play, we will always let player 1 be the leader and player 2 the follower. Let $f_i : \mathcal{X} \times \mathcal{Y} \mapsto \Re$ denote the payoff function of the player i, with $f_i \in C^2(\mathcal{X} \times \mathcal{Y}, \mathbb{R})$ for all $i \in \{1, 2\}$ (f_i is twice continuously differentiable). Let $\nabla_x f_i(x, y)$ and $\nabla_y f_i(x, y)$ denote the gradients of f_i w.r.t. player 1 and player 2's strategies respectively. We denote by $\nabla_{xy} f_i(x, y) = \nabla_y [\nabla_x f_i(x, y)]$ the Jacobian of the gradient $\nabla_x f_i(x, y)$ w.r.t. y, and define $\nabla_{yx} f_i(x, y)$, $\nabla_{xx} f_i(x, y)$, and $\nabla_{yy} f_i(x, y)$ similarly. Finally, we let $\|\cdot\|$ denote the Euclidean norm.

3.1 Simultaneous Gradient Ascent

The most straightforward approach to solving differentiable games is *simultaneous gradient ascent* (SGA), where each player i performs gradient ascent on their own payoff function f_i, treating the other player's strategies as fixed. The two-player SGA updates are

$$x_{t+1} = x_t + \alpha_{1,t} \nabla_x f_1(x_t, y_t) \tag{1}$$

$$y_{t+1} = y_t + \alpha_{2,t} \nabla_y f_2(x_t, y_t) \tag{2}$$

where the sequences $\{\alpha_{1,t}\}$ and $\{\alpha_{2,t}\}$ are learning rate schedules, which may differ between the players. SGA is often the default approach for problems described by two-player games (such as training GANs [12]). We can also view SGA as a model of *ad hoc* learning between independent agents. In the ad hoc setting, players are only aware of their own payoff functions, and the strategies other players follow at each *stage t* of the game.

For non-concave payoff functions, however, we cannot expect SGA to find global optima in the strategy space of either player. This motivates the development of *local* alternatives, in particular the differential Nash equilibrium (DNE).

Definition 3.1 (Differential Nash Equilibrium [27]). Let $\omega(x, y) = (\nabla_x f_1(x, y), \nabla_y f_2(x, y))$ be the individual gradients of the players' payoff functions at (x, y). A strategy profile $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is a differential Nash equilibrium if and only if:

- **(I)** $\omega(x^*, y^*) = 0$, and
- **(II)** $\nabla_{xx} f_1(x^*, y^*)$, and $\nabla_{yy} f_2(x^*, y^*)$ are negative definite.

DNEs are a local version of the Nash equilibrium in the sense that any unilateral deviation within a small neighbourhood of (x^*, y^*) will not improve the payoff of the deviating player. We can also interpret DNEs as the fixed-points of SGA on each player's individual payoff function that have a game-theoretic meaning (i.e. Local Nash). Previous work has shown that gradient-based dynamics such as SGA can converge to DNE in specific classes of games [14, 27].

The main issue with DNEs, however, is that they fail to exist in certain games, which constrains the class of games for which they are applicable as a solution concept. For example, Nash equilibria exist for convex costs (i.e. concave payoffs) on compact and convex strategy spaces, while a DNE exists if these conditions, as described in Başar and Olsder [3, Theorem 4.3 & Chapter 4.9], are met locally within the some neighborhood around a fixed point of SGA [8]. As we will discuss below, an alternative local solution concept based on Stackelberg equilibria exists in more relaxed conditions, and is therefore applicable to a wider class of games.

3.2 Hierarchical Play

In settings such as human–AI collaboration, a natural hierarchy emerges where one agent takes up the role of the leader, and the other adapts to the leader's behaviour [6, 10, 21]. The natural solution concept for hierarchical play is the Stackelberg equilibrium, in which the leader chooses a strategy that maximizes its payoff under the follower's best response.

Definition 3.2 (Stackelberg Equilibrium [29]). Let the set $\mathrm{BR}(x) = \arg\max_{y \in \mathcal{Y}} f_2(x, y)$ denote the follower's set of best-responses when the leader plays x. A joint strategy $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is a *Stackelberg equilibrium* if and only if $y^* \in \mathrm{BR}(x^*)$ and:

$$\min_{y \in \mathrm{BR}(x^*)} f_1(x^*, y) \geq \min_{y \in \mathrm{BR}(x)} f_1(x, y) \tag{3}$$

for all $x \in \mathcal{X}$. Furthermore, the individual strategy x^* satisfying Equation 3 is called a *Stackelberg solution* for the leader.

Recent work [9, 13] has shown that the hierarchical model can be applied to differentiable games as well. While a differentiable game may possess no Nash equilibria, a Stackelberg equilibrium will always exist so long as the strategy spaces $\mathcal{X}$ and $\mathcal{Y}$ are compact [3, Theorem 4.8 & Chapter 4.9]. Note that Definition 3.2 assumes that the follower breaks ties so as to minimize the leader's payoffs. Therefore, a Stackelberg solution maximizes the leader's worst-case payoff assuming the follower will act rationally, and so in zero-sum games the Stackelberg solution guarantees the leader will receive at least its security value. Procedures based on the hierarchical model have proven successful in training generative adversarial networks [8, 19] and actor–critic methods [33].

3.3 Differential Stackelberg Equilibria

Definition 3.2 assumes that both the leader and the follower have found *global* optima in their respective strategy spaces. For non-concave payoff functions, the best an individual player can hope to find with gradient ascent is a *local* optimum of its individual objective (even if the other player's strategy remained fixed). The desire to apply the hierarchical model to differentiable games has motivated the development of a local version of the SE referred to as the *differential Stackelberg equilibrium* (DSE) [8].

In describing the DSE we will make a simplifying assumption that will also be useful for analyzing the convergence of our Hi-C learning update. We assume that for each $x \in \mathcal{X}$ the follower's best-response is unique, and that there exists a continuously differentiable function $r : \mathcal{X} \mapsto \mathcal{Y}$ such that

$$r(x) = \arg\max_{y \in \mathcal{Y}} f_2(x, y), \qquad \forall x \in \mathcal{X}. \tag{4}$$

Furthermore, we assume that $\nabla_y f_2(x, r(x)) = 0$ (the best response is not on a boundary of $\mathcal{Y}$). Given such an r, the leader's objective function becomes $f_1(x, r(x))$, and so a local optimum x^* for the leader will satisfy $\nabla_x[f_1(x, r(x))] = 0$, where $\nabla_x[f_1(x, r(x))] = \nabla_x f_1(x, r(x)) + [\nabla_y f_1(x, r(x))]^\top \nabla_x r(x)$.

Definition 3.3 (Differential Stackelberg Equilibrium [8]). A strategy profile $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$, with $r(x^*) = y^*$, is a differential Stackelberg equilibrium if and only if:

- **(I)** $\nabla_x[f_1(x^*, r(x^*))] = 0$ and $\nabla_y f_2(x^*, y^*) = 0$, and
- **(II)** $\nabla_{xx}[f_1(x^*, r(x^*))]$ and $\nabla_{yy} f_2(x^*, y^*)$ are both negative definite.

Furthermore, any x^* satisfying these conditions is a differential Stackelberg solution (DSS) for the leader.

Condition **(II)** ensures that x^* and y^* are local maxima of the player's individual objectives, rather than minima or saddle points. Note that conditions **(I)** and **(II)** do not imply that $\nabla_x f_1(x^*, y^*) = 0$, and so DSE may not always be stable under gradient ascent on f_1. This reflects the nature of Stackelberg equilibria in general, as the leader's strategy is not necessarily a best-response itself.

3.4 Hierarchical Gradient Update

Assuming that under the follower's payoffs f_2 there is a unique, continuously differentiable best-response function r, a natural approach to finding DSE is to perform gradient ascent on the leader's objective function $f_1(x, r(x))$. Given only f_2, however, we may not be able to derive a closed form expression for r as a function of x. Fortunately, for a joint strategy $(x, y) \in \mathcal{X} \times \mathcal{Y}$ for which $y = r(x)$ and $\nabla_{yy} f_2(x, y)$ is nonsingular, the implicit function theorem gives us a closed-form expression for $\nabla_x r(x)$ as a function of x and y [9]. When $y = r(x)$, we have that that the Jacobian $\nabla_x r(x) = -\left(\nabla_{yy} f_2(x, y)\right)^{-1} \nabla_{xy} f_2(x, y)$. The gradient of the leader's objective then becomes

$$\nabla_x[f_1(x, r(x))] = \nabla_x f_1(x, y) \tag{5}$$

$$-\nabla_y f_2(x, y)^\top \left(\nabla_{yy} f_2(x, y)\right)^{-1} \nabla_{xy} f_2(x, y) \tag{6}$$

$$= D(x, y) \tag{7}$$

Evaluating $D(x, y)$ for a given x requires the value of $y = r(x)$. One way to compute the leader's gradient update is then to optimize y via gradient ascent on f_2 while keeping x constant, and allowing y to converge to $r(x)$ before performing each gradient step for the leader's strategy. A more practical approach (see [9]) is a two-timescale algorithm in which the leader and follower strategies are updated simultaneously, with the follower using a faster learning rate than the leader. Where we only have noisy estimates of the gradients, the two-timescale hierarchical gradient updates become

$$x_{t+1} = x_t + \alpha_{1,t}(D(x_t, y_t) + w_{1,t}) \tag{8}$$

$$y_{t+1} = y_t + \alpha_{2,t}(\nabla_y f_2(x_t, y_t) + w_{2,t}). \tag{9}$$

where $\{w_{1,t}\}$ and $\{w_{2,t}\}$ independent zero-mean noise sequences, and the leader's update $D(x, y)$ is defined as in Equation 7. To achieve time-scale separation, the learning rate schedules are chosen so that $\alpha_{2,t} \gg \alpha_{1,t}$, which allows the follower's strategy to "track" its best response to the leader's current strategy. If the learning rates are chosen such that $\lim_{t \to \infty} \frac{\alpha_{1,t}}{\alpha_{2,t}} = 0$, then results on two-timescale stochastic approximation (see Borkar [5, Chapter 6.1]) can be used to analyze the convergence properties of 8 and 9. In Section 4.1, we instead apply single-timescale stochastic approximation results to analyze the convergence of our Hi-C learning update, abstracting away the follower's learning dynamics in the form of a generic "tracking error" constraint (Assumption 4.5).

3.5 Limitations of Coupled Learning

We can see that the explicit form of the leader's update in Equation 8 depends on the Hessian $\nabla_{yy}f_2$ of the follower's payoff function, which implies that the leader must know the structure of f_2. This assumption does not hold in *ad hoc* interactions, where the leader only has access to the follower's observable behavior. Even in centralized settings where the leader can estimate the gradient and Hessian of f_2 directly, this estimation can be expensive and suffer from high variance. This is particularly true in settings such as reinforcement learning, where gradients (and Hessians) must be estimated through monte-carlo simulations. Other learning updates such as LOLA [11] also depend on estimates of the follower Hessian, and suffer from the same limitations. In the next section we will describe a learning algorithm that estimates $\nabla_x[f_1(x, r(x))]$ from the follower's behavior alone, while maintaining similar convergence guarantees to the two-timescale hierarchical gradient update.

4 UNCOUPLED LEARNING WITH COMMITMENTS

Algorithm 1 The Hi-C learning algorithm, where follower strategies y_t are chosen arbitrarily, and w_t are zero-mean noise variables. $t(n) = \sum_{m=0}^{n-1} k_m$ is the stage at which interval n started.

```
Inputs: Step-size schedule {α_n}_{n≥0}, perturbation schedule
  {δ_n}_{n≥0}, commitment schedule {k_n}_{n≥0}.
Initialize: sample x_0 from X
for step n = 0, 1, ... do
    sample Δ_n uniformly from {−1, 1}^{d_1}.
    x̃_n ← x_n + δ_n Δ_n
    for stage t = t(n), ..., t(n) + k_n − 1 do
        play x̃_n.
        observe s ← f_1(x̃_n, y_t) + w_t.
    end for
    for dimension i = 1, ..., d_1 do
        x^i_{n+1} = x^i_n + α_n s/(δ_n Δ^i_n)
    end for
end for
```

From the leader's perspective, the problem of finding a differential Stackelberg equilibrium is simply that of finding a local maximum of $f_1(x, r(x))$, where $r(x)$ is the follower's (unique) best response when the leader chooses x as their strategy. The challenge in the *uncoupled* setting is that the leader cannot evaluate $\nabla_x[f_1(x, r(x))]$ directly. Specifically, it cannot evaluate the Jacobian $\nabla_x r(x)$ since it does not have access to the follower's payoff function f_2 on which $r(x)$ depends. The leader can, however, estimate the value of $r(x)$ (and therefore $f_1(x, r(x))$) by simply observing the follower's response when it plays strategy x. A natural approach then is to replace the leader's gradient ascent update with a gradient-free learning rule that only requires an unbiased estimate of $f_1(x, r(x))$, and not of $\nabla_x[f_1(x, r(x))]$.

We first consider the hypothetical case where the leader has access to an *oracle* that computes the value of $r(x)$. The leader can use this oracle evaluate $f_1(x, r(x))$ for any $x \in \mathcal{X}$. This allows us to apply a *simultaneous perturbation stochastic approximation* (SPSA) [30] method to approximate gradient ascent on $f_1(x, r(x))$. Specifically, we will derive Hi-C from the single-measurement form of SPSA [31]. For all $n \geq 0$, let Δ_n be independently and uniformly sampled from $\{-1, 1\}^{d_1}$, and let $\{\delta_n\}_{n\geq 0}$ be a decreasing *perturbation schedule*. Let $\{w_n\}_{n\geq 0}$ be a sequence of i.i.d. noise variables, with zero-mean and uniformly bounded variance. The element-wise single-measurement SPSA update is then

$$x^i_{n+1} = x^i_n + \alpha_n \frac{f_1(x_n + \delta_n\Delta_n, r(x_n + \delta_n\Delta_n)) + w_n}{\delta_n \Delta^i_n} \tag{10}$$

for all $i \in [1, d_1]$. SPSA estimates the direction of the gradient by sampling points near the current strategy x_n. Going forward, let $\tilde{x}_n = x_n + \delta_n\Delta_n$ denote the "perturbed" strategy evaluated at step n. The noise terms w_t account for settings the leader can only observe an unbiased estimator of f_1 (e.g., a single policy roll-out).

Estimating $r(\tilde{x}_n)$. In the uncoupled setting, the leader has no way of directly computing $r(x_n)$. What the leader can do is observe the strategies played by the follower, which is assumed to be updating its own strategy so as to maximize its payoff under f_2. This suggests an asynchronous, two-timescale learning process, in which the leader *commits* to playing the perturbed strategy $\tilde{x}_n$ for some k_n stages before updating x_n. For sufficiently large k_n we should hope that after k_n stages the follower's strategy will have approximately converged to its best-response $r(\tilde{x}_n)$.

Under the Hi-C learning update (Algorithm 1), at each interval $n \geq 0$ the leader samples a perturbed strategy $\tilde{x}_n$, and then plays this strategy for the next k_n stages. After k_n stages, the leader updates its strategy element-wise as

$$x^i_{n+1} = x^i_n + \alpha_n \frac{f_1(\tilde{x}_n, \tilde{y}_n) + w_n}{\delta_n \Delta^i_n} \tag{11}$$

where the follower's final strategy within the interval, which we denote by $\tilde{y}_n$, is used as an approximation of $r(\tilde{x}_n)$.

4.1 Convergence Analysis

In this section we make no assumptions about the follower's learning update, and instead prove convergence of the leader's strategy under a generic assumption about the convergence rate of the "tracking error" $\|\tilde{y}_n - r(\tilde{x}_n)\|$ between the follower's strategy and its best-response. In Section 4.2 we will show that, when the follower's payoffs are strongly concave, for an appropriate choice of *commitment schedule* $\{k_n\}_{n\leq 0}$ the tracking error will decrease quickly enough to satisfy this assumption.

The follower is assumed to update their strategy at every stage t, while the leader only performs an update after k_n stages, and so additional notation will be helpful. Let $t(n) = \sum_{m=0}^{n-1} k_m$ be the stage at which the leader begins its nth commitment interval, and let $n(t) = \max\{n : t(n) \leq t\}$ be the current interval at stage t. We let x_n ($n \geq 0$) denote the leader's *mean* strategy after n updates, or $t(n)$ stages, and let y_t denote the strategy the follower played at stage t. We then have $\tilde{y}_n = y_{(t(n)+k_n-1)}$, the last strategy the follower played during the nth commitment interval.

We will need to make several assumptions to prove the convergence of Hi-C . We first require that the follower's best-response is described by a unique function $r(x)$:

Assumption 4.1. There exists a unique function $r : \mathcal{X} \mapsto \mathcal{Y}$ defined by $r(x) = \arg\max_{y \in \mathcal{Y}} f_2(x, y), \forall x \in \mathcal{X}$. Furthermore, r is L_r-Lipschitz and K_r-smooth.

The following three assumptions are standard for the analysis of simultaneous perturbation methods [4, Chapter 5].

Assumption 4.2. x_n and y_t are bounded almost surely:

$$\sup_{n \geq 0} \|x_n\| < \infty \quad \text{and} \quad \sup_{t \geq 0} \|y_t\| < \infty \quad \text{a.s.} \tag{12}$$

This immediately implies that $\tilde{x}_n$ and $\tilde{y}_n$ are bounded a.s., and because r is Lipschitz, it implies $r(\tilde{x}_n)$ is bounded almost surely as well. In practice, the assumption that the strategies remain bounded can be enforced by choosing $\mathcal{X}$ and $\mathcal{Y}$ to be bounded, and projecting the strategies back to $\mathcal{X}$ and $\mathcal{Y}$ whenever necessary.

Assumption 4.3. The leader's payoff function $f_1(x, y)$ is L_1-Lipschitz, and K_1-smooth in both of its arguments.

Assumption 4.3 implies that $\|\nabla_y f_1(x, y)\| \leq L_1$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Combined with Assumption 4.1, it also implies that the leader's hierarchical objective function $g(x) = f_1(x, r(x))$ is also Lipschitz and smooth.

Assumption 4.4. The step-size schedule $\{\alpha_n\}_{n \geq 0}$ and perturbation schedule $\{\delta_n\}_{n \geq 0}$ satisfy:

$$\lim_{n \to \infty} \alpha_n = 0, \quad \lim_{n \to \infty} \delta_n = 0 \tag{13}$$

$$\sum_{n=0}^{\infty} \alpha_n = \infty, \quad \sum_{n=0}^{\infty} \frac{\alpha_n^2}{\delta_n^2} < \infty \tag{14}$$

The decreasing magnitude δ_n of the perturbations means that eventually even small errors in the approximation of $r(\tilde{x}_n)$ could lead to large errors in the estimate of the gradient. We therefore require a fairly strong assumption on the rate of convergence of the tracking error.

Assumption 4.5. Define $\varepsilon_n = \|\tilde{y}_n - r(\tilde{x}_n)\|$, for $n \geq 0$. Given the commitment schedule $\{k_n\}_{n \geq 0}$ and perturbation schedule $\{\delta_n\}_{n \geq 0}$ we have:

$$\lim_{n \to \infty} \frac{\varepsilon_n}{\delta_n} = 0 \quad \text{and} \quad \sup_{n \geq 0} \frac{\varepsilon_n}{\delta_n} < \infty \quad \text{a.s.} \tag{15}$$

Under Assumption 4.5, the additional error introduced by using $\tilde{y}_n$ rather that $r(\tilde{x}_n)$ is bounded and $o(1)$ almost surely, and so becomes negligible asymptotically. To see this, we can rewrite Equation 11 as

$$x_{n+1}^i = x_n^i + \alpha_n \left(\frac{f_1(\tilde{x}_n, r(\tilde{x}_n)) + w_t}{\delta_n \Delta_n^i} + \eta_n^i \right) \tag{16}$$

where

$$\eta_n^i = \frac{f_1(\tilde{x}, \tilde{y}_n) - f_1(\tilde{x}_n, r(\tilde{x}_n))}{\delta_n \Delta_n^i} \tag{17}$$

$$\leq \frac{L_1 \|\tilde{y}_n - r(\tilde{x}_n)\|}{\delta_n \Delta_n^i} \tag{18}$$

since f_1 is L_1-Lipschitz by Assumption 4.3. We then have

$$|\eta_n^i| \leq \frac{L_1 \|\tilde{y}_n - r(\tilde{x}_n)\|}{\delta_n} = L_1 \frac{\varepsilon_n}{\delta_n} < \infty \quad \text{a.s.} \tag{19}$$

where the final inequality comes from the fact that $\frac{\varepsilon_n}{\delta_n}$ is bounded almost surely by Assumption 4.5. We can also see that $\lim_{n \to \infty} |\eta_n^i| = 0$ almost surely, because $\frac{\varepsilon_n}{\delta_n} \to 0$ a.s.. We are now ready to state our main convergence result:

Theorem 4.6. *Let $H \subseteq \mathcal{X}$ be the set $\{x \in \mathcal{X} : \nabla_x [f_1(x, r(x))] = 0\}$. Assume $H \neq \emptyset$, and that Assumptions 4.1–4.5 are satisfied under the Hi-C update (Algorithm 1). Then the leader's strategy x_n will converge to H almost surely as $n \to \infty$.*

This results follows immediately from Bhatnagar et al. [4, Theorem 5.2] by noting that the Hi-C update in Equation 11 is equivalent to the single measurement SPSA update (Equation 10) save for the bounded, $o(1)$ error term η_n^i, which becomes negligible asymptotically (see [5, Chapter 2]). Under stronger assumptions, we can show that x_n converges to a differential Stackelberg solution.

Corollary 4.7. *Additionally, assume that H consists only of isolated, asymptotically stable equilibria of the ODE:*

$$\dot{x}(t) = \nabla_x [f_1(x(t), r(x(t)))]. \tag{20}$$

Then, under the Hi-C update, x_n will converge to a differential Stackelberg solution of the game (f_1, f_2) almost surely as $n \to \infty$.

This follows from the fact that if $x \in H$ is an asymptotically stable equilibrium of $\dot{x}(t) = \nabla_x [f_1(x(t), r(x(t)))]$, then the Hessian $\nabla_{xx} [f_1(x(t), r(x(t)))]$ must be negative definite. Combined with $\nabla_x [f_1(x(t), r(x(t)))] = 0$, this satisfies the requirements of Definition 3.3. At first it may seem contradictory that we can prove convergence to a DSS when these are not guaranteed to exist. The conditions under which Corollary 4.7 holds, however, are precisely those conditions under which a DSS does exist, that is, when $f_1(x, r(x))$ has a strict local minimum in $\mathcal{X}$.

Note that these results make no direct assumptions about the follower's payoff function or learning update. Indeed, if we relaxed Assumption 4.1 they could be satisfied for finite $\mathcal{Y}$ and discontinuous $r(x)$. We simply requires that for every $x \in \mathcal{X}$ the follower's strategy will converge to some unique fixed point $r(x)$ at a sufficiently fast rate *relative* to the leader's commitment schedule. In the next section we will consider specific scenarios in which this requirement is satisfied, and how we can select a suitable commitment schedule given some additional information from the follower.

4.2 Choosing the Commitment Schedule

To identify commitment schedules that satisfy Assumption 4.5, we will need finite-time convergence rate guarantees for the follower's strategy. In this section, we consider the well-studied case where the follower's objective function is strongly concave. Throughout this section we will make a couple of additional assumptions on the payoff function f_2, and the best-response function r:

Assumption 4.8. $\forall x \in \mathcal{X}$, $f_2(x, y)$ is K_2-smooth and μ-strongly concave w.r.t. y.

Under these assumptions, deterministic gradient ascent on f_2 with a fixed step-size schedule $\beta_t = \beta$ is sufficient for the follower's strategy to converge to its best-response.

Proposition 4.9 (Nesterov [20, Chapter 2]). *Let the follower update its strategy using deterministic gradient ascent with a fixed step-size*

$\beta \in (0, \frac{1}{K_2}]$, *such that*

$$y_{t+1} = y_t + \beta \nabla_y f_2(\tilde{x}_{n(t)}, y_t) \tag{21}$$

then for any stage $t \geq 0$, and any $k \in [1, k_{n(t)}]$, we have

$$\|y_{t+k} - r(\tilde{x}_{n(t)})\| \leq (1 - \beta\mu)^{\frac{k}{2}} \|y_t - r(\tilde{x}_{n(t)})\| \tag{22}$$

Now assume that we are given step-size and perturbation schedules $\{\alpha_n\}_{n\geq 0}$ and $\{\delta_n\}_{n\geq 0}$ satisfying Assumption 4.4. To determine a suitable commitment schedule, we first choose an arbitrary sequence $\{\xi_n\}_{n\geq 0}$ such that $\lim_{n\to\infty} \xi_n = 0$ and $\sup_n \xi_n < \infty$. We need to choose a commitment schedule $\{k_n\}_{n\geq 0}$ such that:

$$\frac{1}{\delta_n} \|\tilde{y}_n - r(\tilde{x}_n)\| \leq \xi_n \tag{23}$$

for all $n \geq 0$. To apply Proposition 4.9, we need to be able to bound $\|y_t - r(\tilde{x}_{n(t)})\|$ for all $t \geq 0$. Previously we simply required that the strategies be bounded almost surely (Assumption 4.2). We now assume that this bound is fixed, and known in advance.

Assumption 4.10. There exists a deterministic constant $B < \infty$ such that $\sup_{t\geq 0} \|y_t\| < \frac{B}{2}$ and $\sup_{n\geq 0} \|r(\tilde{y}_n)\| < \frac{B}{2}$ almost surely.

We then have $\sup_{t\geq 0} \|y_t - r(\tilde{x}_n(t))\| \leq B$ almost surely. Then, under Assumptions 4.8 and 4.10, we then have that

$$\frac{1}{\delta_n} \|\tilde{y}_n - r(\tilde{x}_n)\| \leq \frac{1}{\delta_n} (1 - \beta\mu)^{\frac{k_n}{2}} B. \tag{24}$$

Upper-bounding this by ξ_n, we have

$$\frac{1}{\delta_n} (1 - \beta\mu)^{\frac{k_n}{2}} B \leq \xi_n \tag{25}$$

$$\frac{k_n}{2} \ln(1 - \beta\mu) \leq \ln \frac{\delta_n \xi_n}{B} \tag{26}$$

$$2 \frac{\ln \delta_n \xi_n - \ln B}{\ln(1 - \beta\mu)} \leq k_n. \tag{27}$$

Then, setting $\xi_n = \frac{1}{n^p}$ for $p > 0$, we obtain the convergence result:

Corollary 4.11. *For the leader's perturbation schedule $\{\delta_n\}_{n\geq 0}$, define the commitment times as*

$$k_n = \left\lceil 2 \frac{\ln \delta_n - \ln B - p \ln n}{\ln(1 - \beta\mu)} \right\rceil \tag{28}$$

for $p > 0$ and $\beta \in (0, \frac{1}{K_2}]$. Under Assumptions 4.1 through 4.4, and Assumptions 4.8 and 4.10, if the follower updates their strategy using gradient ascent with step-size β, then the leader strategies x_n computed by Hi-C converge to H almost surely as $n \to \infty$.

The specified commitment schedule gives us $\frac{1}{\delta_n} \|\tilde{y}_n - r(\tilde{x}_n)\| \leq \frac{1}{n^p}$, and so $\{k_n\}_{n\geq 0}$ satisfies Assumption 4.5. The result then follows immediately from Theorem 4.6. The assumption that the follower's payoffs are strongly concave is restrictive, but makes intuitive sense in this setting. The only information the leader can obtain about the follower's asymptotic best-responses is through their finite time adaptation to the leader's current strategy. If, over some subset $S \subset \mathcal{Y}$ containing $r(x)$, the curvature of f_2 is allowed to be arbitrarily small, then once the follower's strategy reaches S it may converge to $r(x)$ arbitrarily slowly. From the leader's perspective, it would appear that the follower has already settled on a best-response, such that the leader may over- or under-estimate the value of their current strategy. It is therefore reasonable to require that the leader have some information about how fast the follower's strategy should be expected to converge. In Corollary 4.11, the necessary commitment schedule depends on $\beta\mu$, which also determines the follower's convergence rate.

5 DISCUSSION

A major motivation for our interest in "uncoupled" learning in general sum games is the problem of ad hoc cooperation between humans and AI, where the AI cannot assume anything about how the human's behavior will change over time. In such settings, the human and AI will need to adapt to one another simultaneously. Previous analysis of *naive* simultaneous learning updates such as SGA has suggested that such learning processes may be highly unstable, and may fail to converge to good joint strategies. Research in differentiable games has in recent years focused on the types of centralized training settings commonly arising in deep learning. Therefore, many of training algorithms developed for this setting require that one or both learners have detailed knowledge about the other's loss function and learning update. Such methods, and corresponding theoretical analysis, are therefore not directly applicable to human–AI interaction. Our work overcomes this limitation.

Our approach also has the potential to be useful in centralized training as well. Compared to the coupled hierarchical gradient update, Hi-C will generally have much lower per-step computational costs. An open question, however, is whether there is a trade-off between per-step complexity and the number of training steps required. It is possible that in high-dimensional settings Hi-C will take much longer to converge due to noise introduces by the gradient estimation procedure. An empirical evaluation of how Hi-C compares to the coupled hierarchical gradient update, or other coupled approaches such as LOLA when scaled to high-dimensional training problems such as GANs or multi-agent reinforcement learning is an important direction for future work.

6 CONCLUSION

We have presented what is, to the best of our knowledge, the first *uncoupled* learning update that can be shown to converge to differential Stackelberg solutions for a broad class of general-sum differentiable games. The Hi-C learning update can be implemented without access to the follower's payoff function or the details of their learning update. This also means that Hi-C does not need to estimate the gradients or Hessians of the follower's payoffs. Our results provide theoretical insights into uncoupled hierarchical learning processes, where one agent must learn about the preferences of another agent through its observable behavior alone.

Immediate future directions would include expanding the class of follower learning updates and payoff functions for which we can provide concrete convergence guarantees. These could include more flexible learning strategies such as stochastic gradient descent, or no-regret learning algorithms such as online mirror descent. Another question is whether convergence to a Stackelberg solution can be guaranteed for a fixed commitment time, where at each interval the follower builds on the learning progress made in previous intervals? Finally, future work could consider convergence guarantees for non-concave follower payoffs with multiple local optima, and therefore, non-unique best-response functions.

ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI; grants 319264, 313195, 305780, 292334, 328400, 28400), the UKRI Turing AI World-Leading Researcher Fellowship EP/W002973/1, and the Finnish Science Foundation for Technology and Economics (KAUTE). This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, grant number 024.004.022.

REFERENCES

[1] Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. 2021. Sample-efficient learning of Stackelberg equilibria in general-sum games. *Advances in Neural Information Processing Systems* 34, 25799–25811.
[2] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. 2018. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*. PMLR, 354–363.
[3] Tamer Başar and Geert Jan Olsder. 1998. *Dynamic noncooperative game theory*. SIAM.
[4] S. Bhatnagar, H.L. Prasad, and L.A. Prashanth. 2012. *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods*. Springer.
[5] Vivek S Borkar. 2009. *Stochastic approximation: a dynamical systems viewpoint*. Springer.
[6] Mustafa Mert Çelikok, Frans A Oliehoek, and Samuel Kaski. 2022. Best-Response Bayesian Reinforcement Learning with Bayes-adaptive POMDPs for Centaurs. In *International Conference on Autonomous Agents and Multiagent Systems*. 235–243.
[7] Daniela Pucci de Farias and Nimrod Megiddo. 2003. How to combine expert (or novice) advice when actions impact the environment. *Advances in Neural Information Processing Systems* 17.
[8] Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. 2020. Implicit learning dynamics in Stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *International Conference on Machine Learning*. PMLR, 3133–3144.
[9] Tanner Fiez, Benjamin Chasnov, and Lillian J Ratliff. 2019. Convergence of learning dynamics in Stackelberg games. *arXiv preprint arXiv:1906.01217* (2019).
[10] Jaime F Fisac, Eli Bronstein, Elis Stefansson, Dorsa Sadigh, S Shankar Sastry, and Anca D Dragan. 2019. Hierarchical game-theoretic planning for autonomous vehicles. In *International Conference on Robotics and Automation*. IEEE, 9590–9596.
[11] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with Opponent-Learning Awareness. In *International Conference on Autonomous Agents and MultiAgent Systems*. IFAAMAS, 122–130.
[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Vol. 27. 2672–2680.
[13] Chi Jin, Praneeth Netrapalli, and Michael Jordan. 2020. What is local optimality in nonconvex-nonconcave minimax optimization?. In *International Conference on Machine Learning*. PMLR, 4880–4889.
[14] Alistair Letcher, David Balduzzi, Sébastien Racanière, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. 2019. Differentiable Game Mechanics. *Journal of Machine Learning Research* 20, 84 (2019), 1–40.
[15] Christopher Lu, Timon Willi, Christian A Schroeder De Witt, and Jakob Foerster. 2022. Model-free opponent shaping. In *International Conference on Machine Learning*. PMLR, 14398–14411.
[16] Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. 2019. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838* (2019).
[17] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. 2018. Cycles in adversarial regularized learning. In *Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2703–2717.
[18] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. 2017. The numerics of GANs. *Advances in Neural Information Processing Systems* 30.
[19] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. 2017. Unrolled Generative Adversarial Networks. In *International Conference on Learning Representations*.
[20] Yurii Nesterov. 2018. *Lectures on convex optimization*. Springer.
[21] Stefanos Nikolaidis, Swaprava Nath, Ariel D Procaccia, and Siddhartha Srinivasa. 2017. Game-theoretic modeling of human adaptation in human-robot collaboration. In *ACM/IEEE International Conference on Human-Robot Interaction*. 323–331.
[22] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. 2019. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems* 32.
[23] Christos H. Papadimitriou and Georgios Piliouras. 2018. From Nash Equilibria to chain recurrent Sets: an algorithmic solution concept for game theory. *Entropy* 20, 10 (2018), 782.
[24] Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. 2019. Learning optimal strategies to commit to. In *AAAI Conference on Artificial Intelligence*, Vol. 33. 2149–2156.
[25] Jan Poland and Marcus Hutter. 2005. Defensive universal learning with experts. In *International Conference on Algorithmic Learning Theory*. Springer, 356–370.
[26] Giorgia Ramponi and Marcello Restelli. 2022. Learning in Markov games: can we exploit a general-sum opponent?. In *Uncertainty in Artificial Intelligence*.
[27] Lillian J. Ratliff, Samuel A. Burden, and S. Shankar Sastry. 2016. On the characterization of local Nash equilibria in continuous games. *IEEE Trans. on Automatic Control* 61, 8 (2016), 2301–2307.
[28] Florian Schäfer and Anima Anandkumar. 2019. Competitive gradient descent. *Advances in Neural Information Processing Systems* 32.
[29] Marwaan Simaan and Jose B Cruz. 1973. On the Stackelberg strategy in nonzero-sum games. *Journal of Optimization Theory and Applications* 11, 5 (1973), 533–555.
[30] James C Spall. 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. on Automatic Control* 37, 3 (1992), 332–341.
[31] James C Spall. 1997. A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica* 33, 1 (1997), 109–112.
[32] Timon Willi, Alistair Letcher, Johannes Treutlein, and Jakob Foerster. 2022. COLA: consistent learning with opponent-learning awareness. In *International Conference on Machine Learning*. PMLR, 23804–23831.
[33] Liyuan Zheng, Tanner Fiez, Zane Alumbaugh, Benjamin Chasnov, and Lillian J Ratliff. 2022. Stackelberg actor-critic: Game-theoretic reinforcement learning algorithms. In *AAAI Conference on Artificial Intelligence*, Vol. 36. 9217–9224.