Uncoupled Learning of Differential Stackelberg Equilibria with Commitments

Robert Loftin[1], Mustafa Mert Çelikok[2], Herke van Hoof[3], Samuel Kaski[2,4] and Frans A. Oliehoek[1]

1. Delft University of Technology, 2. Aalto University, 3. University of Amsterdam, 4. University of Manchester

Introduction

We consider the problem of learning **Stackelberg equilibria** in general sum **differentiable games**. Under the Stackelberg equilibrium, the "leader" selects a strategy that maximizes their utility under the assumption that the "follower" will choose their best response to this strategy. The Stackelberg equilibrium is a natural solution concept in many settings, particularly those requiring cooperation between agents with conflicting preferences.

Previous work has presented gradient ascent algorithms for finding "local" Stackelberg equilibria in two-player differentiable games. These methods are **coupled** however, in the sense that the leader's gradient update depends on knowledge of the follower's utilities. As such, these methods cannot be applied to **ad hoc** settings, where the leader and follower are independent agents that have not previously interacted. Our work presents an **uncoupled** algorithm for learning local Stackelberg equilibria, based on zeroth-order optimization.

Stackelberg Equilbria

Let f_1 and f_2 be the leader and follower utilities. The Stackelberg objective for the leader's strategy x is then defined as

$$g(x) = f_1(x, BR(x)),$$

where $BR_2(x)$ is the follower's best response to x, that is

$$BR_2(x) = \arg\max_{y \in \mathcal{Y}} f_2(x, y).$$

We assume that $BR_2(x)$ is unique for each x. A **differential** Stackelberg equilibrium is a joint strategy $\langle x, y \rangle$ s.t.

$$\nabla_x g(x) = 0, \quad \nabla_y f_2(x, y) = 0,$$

and for which the Hessians $\nabla_{xx} g(x)$ and $\nabla_{yy} f_2(x, y)$ are negative definite [1].

Hierarchical Gradient Ascent

The challenge in optimizing $g(x)$ is that its gradient $\nabla_x g(x)$ depends on the Jacobian of the follower's best response function, $\nabla_x BR_2(x)$. In [1], the follower's Jacobian is computed as

$$\nabla_x BR_2(x) = -[\nabla_{yy} f_2(x, y)]^{-1} \nabla_{xy} f_2(x, y),$$

based on the implicit function theorem. Alternatively, [2] differentiate through a finite number of follower gradient ascent steps, specifically taking

$$\nabla_x BR_2(x) \approx \eta \nabla_{xy} f_2(x, y)$$

as an approximation of the Jacobian of the follower's strategy after a single gradient ascent step. **Both of these estimates depend on the Hessian of the follower's utility function $\nabla^2 f_2$.**

Uncoupled Learning

To optimize $g(x)$ without knowing $\nabla^2 f_2$, we use a **gradient free** optimization method, in this case the one-sample SPSA update [3] given by

$$x_{n+1} = x_n + \alpha_n \frac{g(x_n + \delta_n \Delta_n)}{\delta_n} \Delta_n,$$

where Δ_n is sampled uniformly from $\{-1, 1\}^d$. To compute $g(\tilde{x}_n)$, the leader **commits** to the perturbed strategy $\tilde{x}_n$ for k_n episodes, where $\{k_n\}_{n \geq 0}$ is a time-varying **commitment schedule**. It then uses the follower's most recent strategy after k_n episodes, which we denote as $\tilde{y}_n$, to approximate the followers true best response to $\tilde{x}_n$.

The **Hi-C** learning algorithm – follower strategies y_t are chosen by an unknown learning rule. Let $t(n) = \sum_{m=0}^{n-1} k_m$.

Inputs: Step-sizes $\{\alpha_n\}_{n \geq 0}$,
perturbation schedule $\{\delta_n\}_{n \geq 0}$,
commitment schedule $\{k_n\}_{n \geq 0}$.
Initialize: sample x_0 from $\mathcal{X}$
for step $n = 0, 1, \ldots$ **do**
  sample Δ_n from $\{-1, 1\}^{d_1}$.
  $\tilde{x}_n \leftarrow x_n + \delta_n \Delta_n$
  for $t = t(n), \ldots, t(n) + k_n - 1$ **do**
    play $\tilde{x}_n$.
    observe $\tilde{y}_n \leftarrow y_t$.
  end for
  for dimension $i = 1, \ldots, d_1$ **do**
    $x^i_{n+1} \leftarrow x^i_n + \frac{\alpha_n}{\delta_n \Delta^i_n}[f_1(\tilde{x}_n, \tilde{y}_n) + w_t]$
  end for
end for

Convergence Results

For the right choice of commitment schedule $\{k_n\}_{n \geq 0}$, existing convergence results for one-sample SPSA apply.

- The "approximation error" of the follower's best response, that is,

  $$\epsilon_n = \|\tilde{y}_n - BR_2(\tilde{x}_n)\|,$$

  must decrease sufficiently fast. We must ensure that $\lim_{n \to \infty} \frac{\epsilon_n}{\delta_n} = 0$.
- When $f_2(x, \cdot)$ is **strongly concave**, choosing a commitment schedule for which $k_n = O(\log n)$ ensures that x_n will converge to a local optimum of $g(x)$ as n goes to inifinity.

References

[1] Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. "Implicit learning dynamics in Stackelberg games: Equilibria characterization, convergence analysis, and empirical study". In: *ICML*. 2020.

[2] Jakob Foerster et al. "Learning with Opponent-Learning Awareness". In: *AAMAS*. 2018.

[3] James C Spall. "A one-measurement form of simultaneous perturbation stochastic approximation". In: *Automatica* (1997).