Certifiable algorithms for learning and control in multiagent systems

Benjamin Chasnov[1], Tanner Fiez[1], Eric Mazumdar[2], Samuel A. Burden[1], Samuel Coogan[3], Lillian J. Ratliff[1]
[1]University of Washington; [2]University of California, Berkeley; [3]Georgia Institute of Technology

CNS-1836819

Overview

We study learning and control in multi-agent systems in which agents repeatedly interact in a simultaneous or hierarchical order of play, and update their strategies via myopic learning rules. We seek rigorous convergence guarantees on the limiting outcomes of such multi-agent interactions. Key challenges include:

- *Dynamic Environment:* The environment is non-stationary, evolving as a function of the agents' strategies which themselves are being adapted and learned in response to observations from the environment.
- *Misaligned Objectives:* Equilibrium solution concepts are studied as agents are individually optimizing their objective which may not be aligned with others.

We provide theoretical convergence guarantees for deterministic and stochastic gradient update rules and support the analysis with illustrative numerical examples. Many multi-agent learning algorithms (gradient play, policy gradient, individual Q-learning, etc.) fit in this framework.

Hierarchical Play Games

Setting: agents select actions in a sequential order; the *leader* selects an action with the knowledge that the *follower* subsequently selects a best response—i.e., the leader aims to solve $\min_{x_1}\{f_1(x_1,x_2)|\, x_2 \in \arg\min_{x'} f_2(x_1,x')\}$. Myopic gradient-based update rules for this setting are given by

$$\begin{aligned} \textit{leader}: &\quad x_{1,k+1} = x_{1,k} - \gamma_{1,k} g_1(x_k) \\ \textit{follower}: &\quad x_{2,k+1} = x_{2,k} - \gamma_{2,k} g_2(x_k) \end{aligned}$$

where $g_1 \equiv Df_1 \equiv D_1 f_1 + D_2 f_1 D\xi$ and $D\xi \equiv -D_2^2 f_2^{-1} \circ D_{12} f_2$ is defined implicitly—i.e., via the implicit mapping theorem applied to $D_2 f_2 \equiv 0$—and $g_2 \equiv D_2 f_2$. Moreover, the learning rates are such that $\gamma_{1,k} = o(\gamma_{2,k})$.

Definition: A strategy x^* is a *differential Stackelberg equilibrium* if $Df_1(x^*) = 0$, $D_2 f_2(x^*) = 0$ and $D^2 f_1(x^*) > 0$, $D_2^2 f_2(x^*) > 0$.

Asymptotic convergence guarantee

Theorem 1 With $x_0 \in B_r(x^*)$, under suitable assumptions on the noise, functions being Lipschitz, and the implicit map ξ being a globally asymptotically stable attractor, $x_{2,k}\xi(x_{1,k})$ and $x_k \to x^*$ almost surely. If $D^2 f_1(x^*) > 0$, then x^* is a stable differential Stackelberg equilibrium.

Extensions:

–Concentration bounds: Analogous bounds as those for simultaneous play hold.

–Relaxed assumption: results hold assuming ξ is a *local attractor* on $B_r(x^*)$.

–Conjectures: viewing ξ as the leader's *conjecture* about the follower, we are exploring generalizations (beyond best response) using conjectural variations.

Generative adversarial nets: a Stackelberg game

The hierarchical play update can be applied to robust ML applications such as GANs: in a zero-sum setting, the generator is the leader and the discriminator is the follower. In a convex-concave regime, e.g., the Stackelberg policy results in a lower cost relative to Nash; while generally not the case, the approach may lead to insights for obtaining more robust GAN policies.

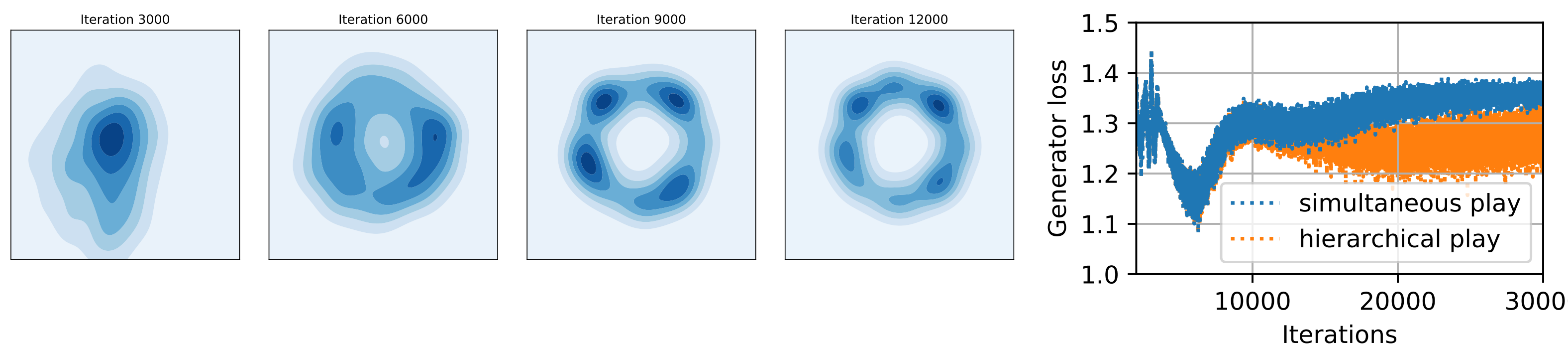

Figure 1: Hierarchical play on a GAN: the leader's cost is lower compared to simultaneous play.

Simultaneous Play Games

Setting: Each agent selects an action simultaneously in an effort to solve $\min_{x_i} f(x_i, x_{-i})$. Due to their scalability and prevalence in ML, we consider gradient-based updates for each player of the form

$$x_{i,k+1} = x_{i,k} - \gamma_i g_i(x_k) \quad \text{where} \quad g_i(x_k) = D_i f_i(x_k).$$

Definition: A strategy x^* is a *differential Nash equilibrium* if $D_i f_i(x^*) = 0$ and $D_i^2 f_i(x^*) > 0$ for all $i \in \{1, \ldots, n\}$.

Oracle gradient access: finite-time convergence

Theorem 2 Suppose g is Lipschitz and let

$$\alpha = \min_{x \in B_r(x)} \sigma^2_{\min}((Dg(x) + Dg(x)^T)/2), \beta = \max_{x \in B_r(x)} \sigma^2_{\max} Dg(x),$$

and $\gamma = \frac{\sqrt{\alpha}}{\beta}$. Then $x_0 \in B_r(x^*) \implies x_k \in B_\varepsilon(x^*), \forall k \geq T$ where

$$T = \lceil 2\tfrac{\beta}{\alpha} \log(r/\varepsilon) \rceil.$$

Concentration bounds for stochastic gradients

Theorem 3: For sufficiently large N,

$$\Pr(x_k \in B_\varepsilon(x^*), \forall k \geq N | x_N \in B_r(x^*)) \geq 1 - o(\textstyle\sum_{k \geq N} \gamma_k^2)$$

Corollary 1:
$x_n \to x^*$ almost surely conditioned on the event that $x_n \in B_r(x^*)$.

Extensions:

–Oracle gradient access: we have non-uniform learning rate convergence guarantees which parallel the study of preconditioning in optimization

–Stochastic gradient access: leveraging multi-timescale stochastic approximation analysis, we also have concentration bounds for the non-uniform learning rate setting. We are investigating how the particular choice of learning rate impacts the concentration bound.

The multiagent cost landscape

Nash and Stackelberg stationary points: $\{x|\ g(x) = 0,\ D_i g_i(x) \geq 0\}$. To contrast simultaneous and hierarchical play, consider the stationary solutions and learning paths for a quadratic two-player game.

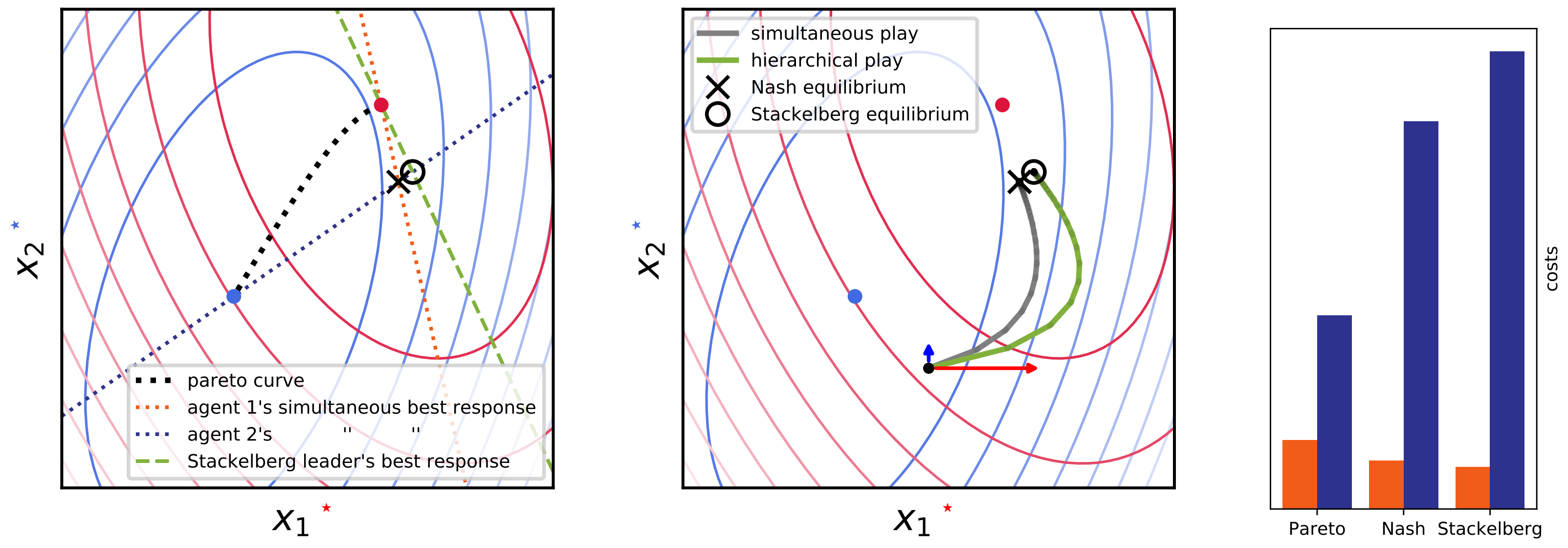

Figure 2: Two scalar agents with quadratic costs, $c_i(x) = x^T Q_i x + q_i^T x$. Convergence to the Nash $(\times)$ and Stackelberg equilibrium $(\circ)$ are shown, contrasted with the Pareto front $(-)$.

References

- B. Chasnov, L. Ratliff, E. Mazumdar, S. Burden. Convergence Analysis of Gradient-Based Learning with Non-Uniform Learning Rates in Non-Cooperative Multi-Agent Settings, UAI 2019 (available on ArXiv)
- T. Fiez, B. Chasnov, L. Ratliff. Convergence of Learning Dynamics in Stackelberg Games, available on ArXiv, 2019

Policy gradient for linear quadratic games

LQ games are a nice benchmark for understanding the effectiveness of gradient-based learning in games: under mild conditions, feedback Nash exist, are unique, and can be computed fairly efficiently via Lyapunov iterations. Agents with linear dynamics $z(t+1) = Az(t) + B_1 u_1(t) + B_2 u_2(t)$ and infinite time quadratic costs,

$$f_i(u_i, u_{-i}) = \mathbb{E}_{z_0 \sim D}\left[\sum_{t=0}^{\infty} z(t)^T Q_i z(t) + u_i(t)^T R_{i,i} u_i(t) + u_{-i}(t)^T R_{i,-i} u_{-i}(t)\right]$$

have unique Nash feedback matrices K_i^* where $u_i(t) = K_i^* z(t)$. We solve for these linear policies using a variant of policy gradient in which we perform rollouts in minibatches using sampled policies (e.g., $u_t = K_t x_t + w_{t+1}$, $w_{t+1} \sim \mathcal{N}(0, \sigma^2 I)$):

$$K_i^+ = K_i - \gamma \widehat{\nabla_{K_i} f_i}(K_i, K_{-i})$$

$$\nabla_{K_i} f_i(K_i, K_{-i}) = 2\Big(R_{i,i} K_i - B_i^T P_i \Big(A - \sum_j B_j K_j\Big)\Big) \mathbb{E}_{z_0 \sim D}\left[\sum_{t=0}^{\infty} z_t z_t^T\right]$$

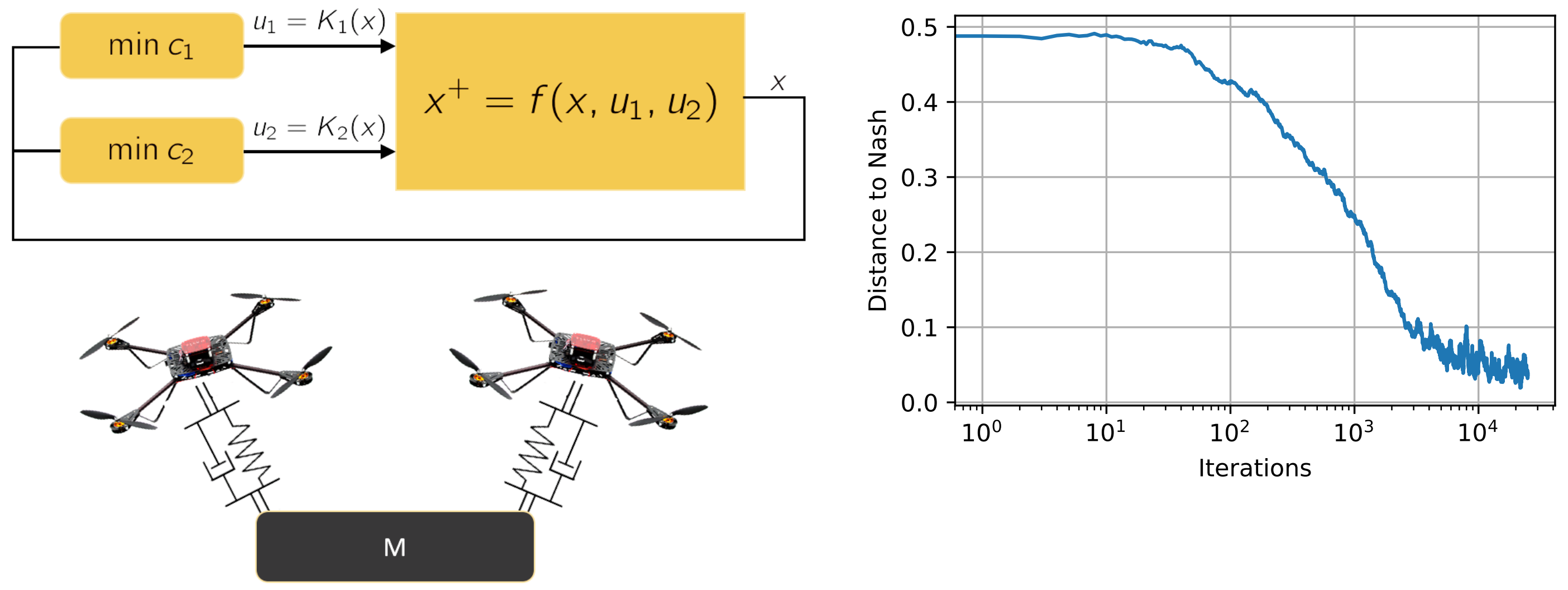

Figure 3: Agents play a dynamic game with costs dependent on a shared state $z(t)$, controls for the individual $u_i(t)$ and for others $u_{-i}(t)$. Convergence of our gradient method to the Nash equilibrium is shown for a randomly generated stable system A and stochastic updates $\widehat{g}$.

Game dynamics with non-uniform learning rates

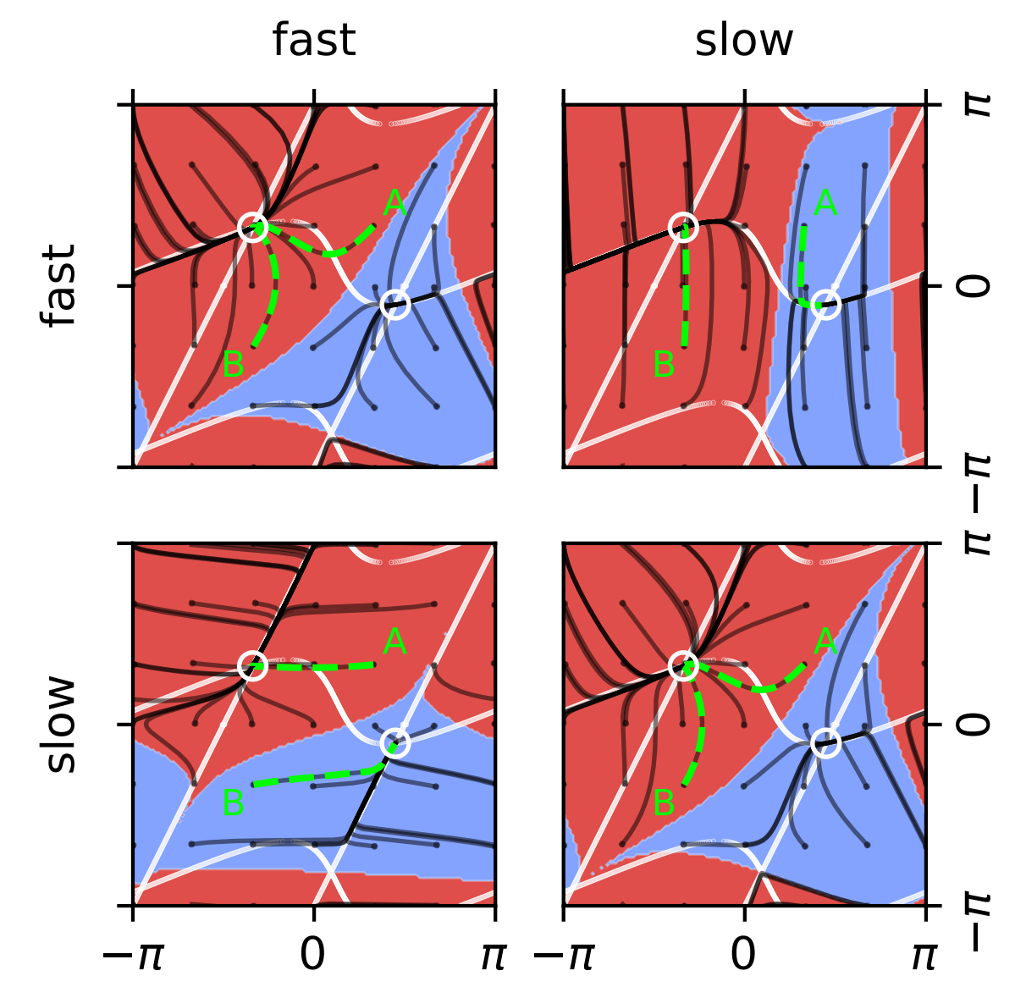

Figure 4: Region of attraction of the two NEs.

Autonomous agents may learn at different rates, which causes warping of the vector field learning dynamics and convergence to different stationary points (*caeteris paribus*). We compare the convergence of a "fast" and "slow" agent to different Nash equilibria under simultaneous play in a location game on the unit torus: $f_i(\theta_i, \theta_{-i}) = \alpha_i \cos(\theta_i) + \cos(\theta_i - \theta_{-i})$.

Future Work

- Hybrid dynamics: Generalize results to settings where the action space consists of a mix of discrete and continuous inputs.
- Limited or Bandit Feedback: Explore learning in multi-agent systems under limited feedback leveraging zero–th order optimization, asynchronous stochastic approximation, and bandit models.
- Model-Free vs Model-Based: investigate adaptive control and conjecture-based learning paradigms for strategically biasing opponent.
- Incentive Design: expanding analysis to more general hierarchical decision problems with the goal of, e.g., influencing behavior.

additional funding: T. Fiez is funded by an NDSEG Fellowship; B. Chasnov is funded by the Computational Neuroscience Center at UW.