Convergence analysis of gradient-based learning in continuous games

Benjamin Chasnov[1], Lillian J. Ratliff[1], Eric Mazumdar[2], Samuel A. Burden[1]; [1]University of Washington; [2]University of California, Berkeley;

CNS-1836819

Overview

Leveraging techniques from dynamical systems theory, we provide theoretical convergence guarantees for deterministic and stochastic gradient-based learning in competitive multi-agent settings 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.

Simultaneous play

Setting: Each agent $i \in [n]$ aims to select an action $x_i \in \arg\min_z f_i(z, x_{-i})$. Agents simultaneously update their actions using a gradient-based updates of the form

$$x_{i,k+1} = x_{i,k} - \gamma_i g_i(x_k)$$

in two possible settings:

– deterministic. agents have oracle access to individual gradient: $g_i(x_k) = D_i f_i(x_k)$

– stochastic. agents have an unbiased estimator: $g_i(x_k) = \widehat{D_i f_i}(x_k)$

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 $i \in [n]$.

The multiagent cost landscape

Nash equilibria fixed points are the intersection of the agents' best-response curves that have positive semi-definite Hessian for each agent. As an example, we illustrate the cost landscape of two agents with scalar actions and quadratic costs.

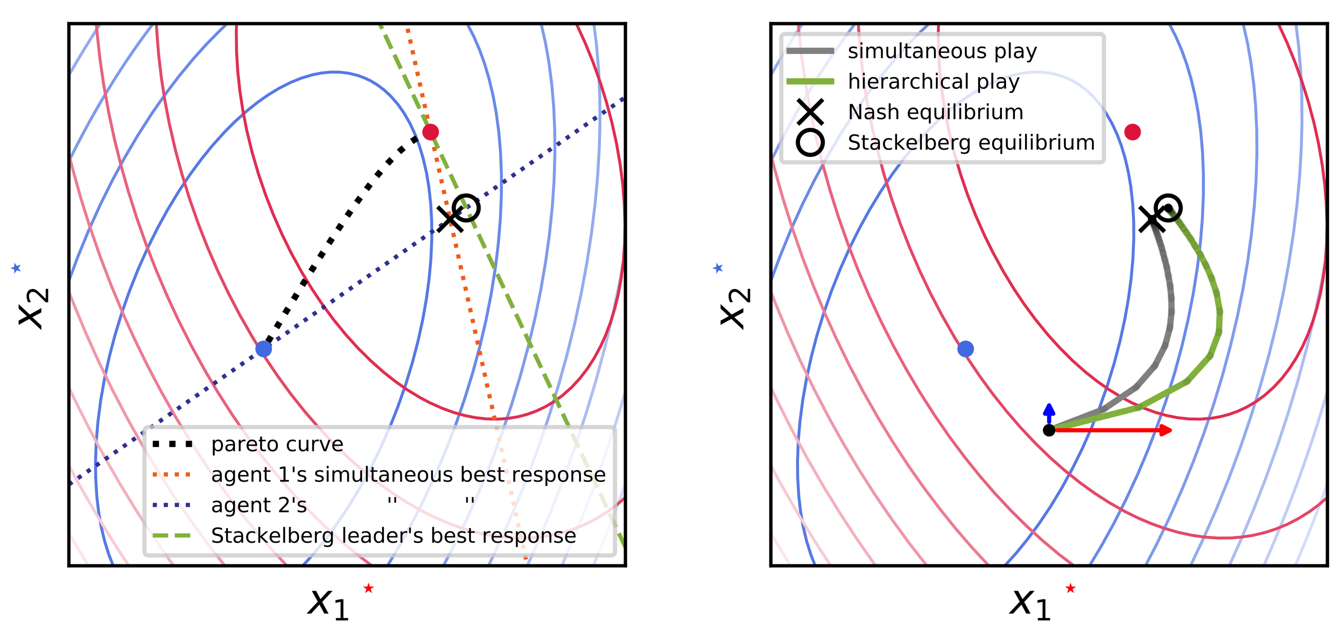


Figure 1: Two agents with costs $f_i(x) = x^T Q_i x + q_i^T x$ where $x = (x_1, x_2)^\top$. Comparison of the non-cooperative (Nash, ×), leader-follower (Stackelberg, ○) and cooperative (Pareto, ···) equilibria.

Non-asymptotic convergence guarantees

We derive finite-time convergence guarantees to locally asymptotically stable differential Nash equilibria for agents with uniform learning rates, $\gamma_i = \gamma, \forall i \in [n]$.

Oracle gradients: finite-time convergence

Theorem 1 Suppose $g = (g_1, \ldots, g_n)$ 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

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

We also derive results for agents learning with non-uniform rates.

Stochastic approximation of dynamical systems

We also derive concentration bounds for agents with unbiased estimates of their gradient.

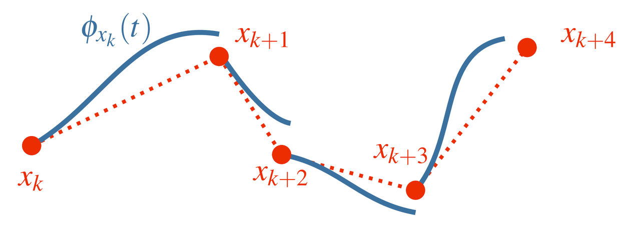
Stochastic gradients: concentration bounds

Theorem 2 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:

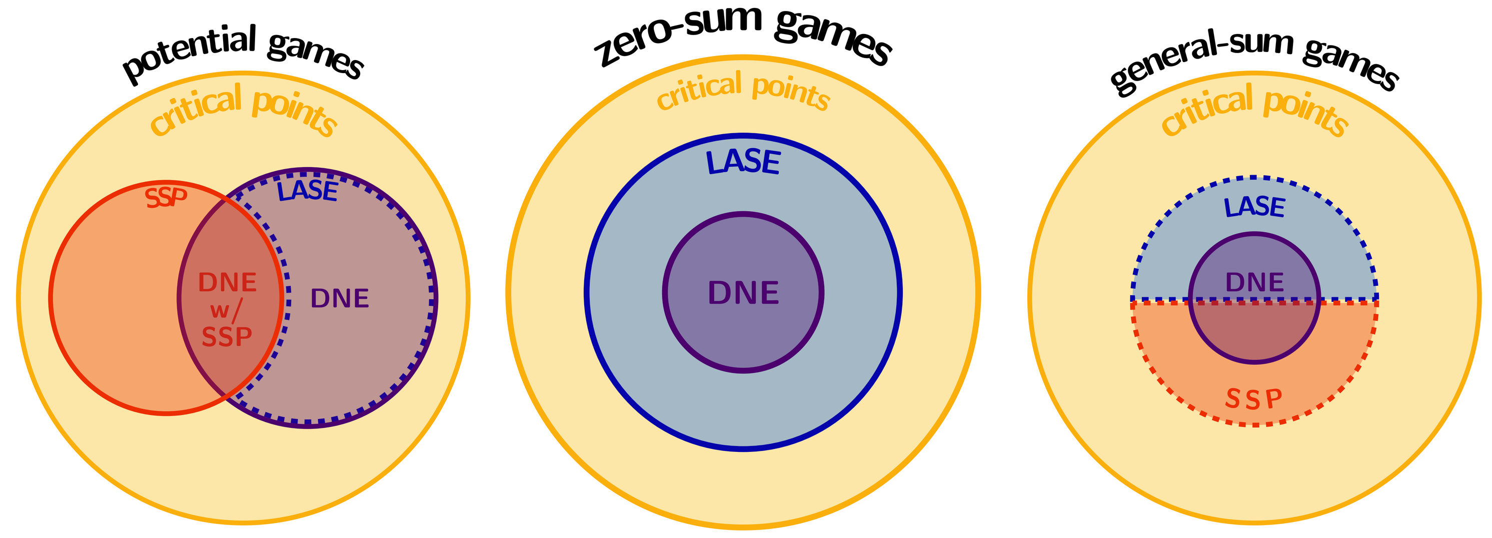
$x_n \to x^*$ almost surely conditioned on the event that $x_n \in B_r(x^*)$.



- $\phi_{x_k}(t)$: flow of $\dot{x} = -g(x)$ initialized at x_k.
- $\bar{x}_k(t)$ = linear interpolation between x_k and x_{k+1}.
- union bound over events $\|\phi_{x_k}(t) - \bar{x}_k(t)\| \geq \delta$ for each $k \geq N$.

Asymptotic stability of differential Nash equilibria

Critical points under the learning dynamics can be differential Nash equilibria (DNE), strict saddle points (SSP), and/or locally asympototically stable equilibria (LASE).



Game dynamics with non-uniform learning rates

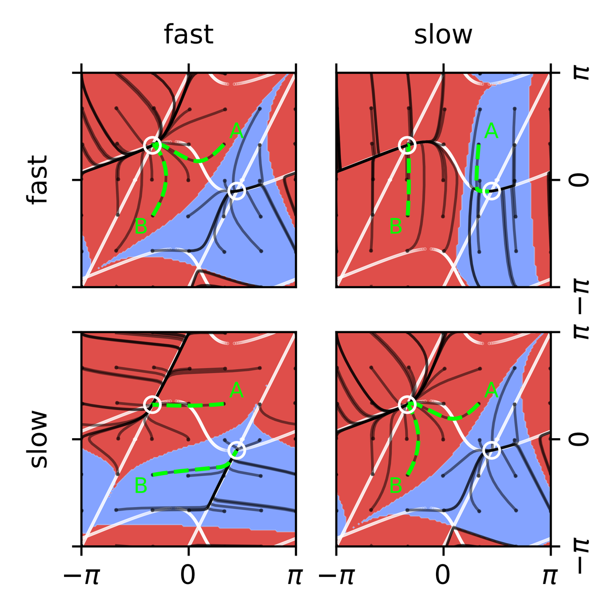


Figure 2: Region of attraction of the two NEs.

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: $f_i(\theta_i, \theta_{-i}) = \alpha_i \cos(\theta_i) + \cos(\theta_i - \theta_{-i})$.

Generative adversarial nets: a zero-sum continuous game

The simultaneous play update and its extensions can be applied to ML applications such as GANs: a zero-sum continuous game between a generator and a discriminator, where the "actions" of each agents are the parameters of a function approximator. The convergence analysis we provide can lead to more performant training algorithms.

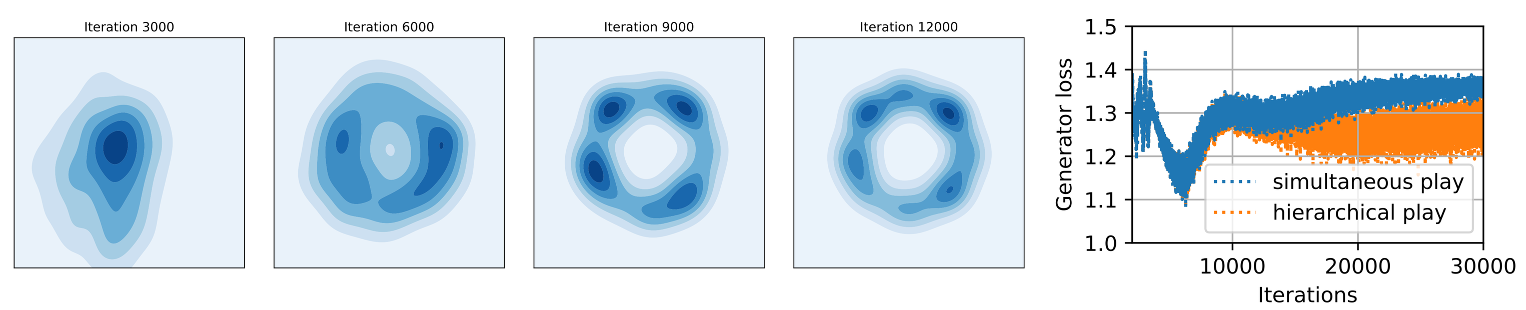


Figure 3: Simultaneous and hierarchical play on a GAN: the leader's cost is lower.

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., $u_t = K_t x_t + w_{t+1}$, $w_{t+1} \sim \mathcal{N}(0, \sigma^2 I)$):

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

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

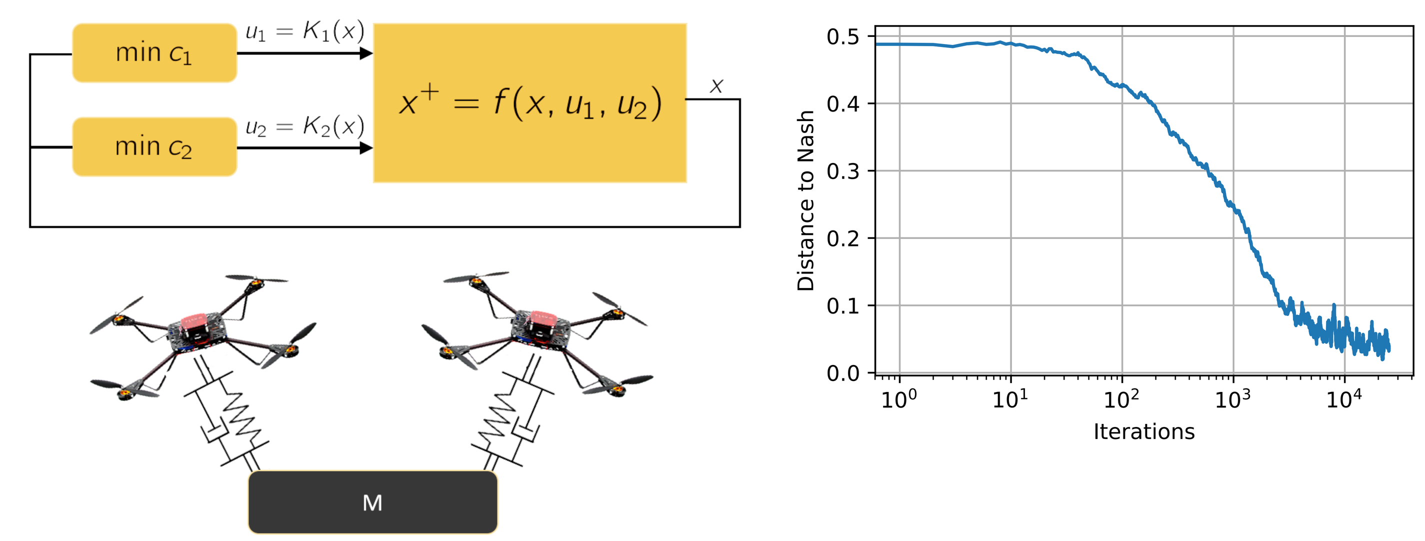


Figure 4: Agents play a dynamic game with costs dependent on a shared state $z(t)$, controls for the individual $u_i(t)$ and for others $u_{-i}(t)$. Convergence of our gradient method to the Nash equilibrium is shown for a randomly generated stable system A and stochastic updates $\widehat{g}$.

Extensions: hierarchical and conjectural play

Setting: Agents select actions while being cognizant of others' decision making process, for example, $x_{i,k+1} = x_{i,k} - \gamma_i g_i(x_{i,k}, \xi(x_{i,k}))$ where the function ξ maps how other agents would act in respond to the choice of $x_{i,k}$.

– hierarchical play:. the leader implicitly assumes that the followers play best-response and optimizes accordingly: $\min_{x_i} f_i(x_i, \xi(x_i))$, where $D\xi \equiv -D_j^2 f_j^{-1} \circ D_{ij} f_j$, $\forall i \neq j$.

– conjectural play:. agents are doubled-sided and form a *conjecture* of other agents' learning process and anticipate it: $\min_{x_i} f_i(x_i, \xi(x_i))$, where $\xi(x_i) = x_{-i} - \gamma_{-i} D_{-i} f_{-i}(x)$.

Future Work

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.

References

B. Chasnov, L. Ratliff, E. Mazumdar, S. Burden. *Convergence Analysis of Gradient-Based Learning with Continuous Games*, UAI 2019 (ArXiv: 1906.00731).

T. Fiez, B. Chasnov, L. Ratliff. *Convergence of Learning Dynamics in Stackelberg Games*, 2019 (ArXiv:1906.01217).

Acknowledgements: B. Chasnov is funded by the Computational Neuroscience Center at UW.