Dynamics of Learning Summary - Hanchun Wang

Tuesday, July 19, 2022 1:39 PM

Equilibrium	Action domain	$\Delta_n = \{x \in \mathbb{R}^n; 0 \le x_i \le 1, x_1 + \cdots + x_n = 1\}$ be the $(n-1)$ dimensional simplex
	Linear Payoff	$x \cdot Ay, e_i \cdot Ax = (Ax)_i$
	Indifference set	$Z_{ij} = \{x; (Ax)_i = (Ax)_j\}$
	Best Response	$\mathcal{BR}(x) = \underset{y \in \Delta}{\operatorname{argmax}}(y \cdot Ax) \overset{\text{def}}{=} \{y' \in \Delta; y \cdot Ax \le y' \cdot Ax, \forall y \in \Delta\}$
	(strict) Nash Equilibrium	$\hat{x} \in \Delta \; is \; NE \overset{\text{def}}{=} x \cdot A\hat{x} \le \hat{x} \cdot A\hat{x}, \forall x \in \Delta$, strict inequality for strict NE $\boldsymbol{\hat{x} \in \Delta \; is \; NE \leftrightarrow \hat{x} \in \mathcal{BR}(\hat{x})}$ $\hat{x} \in \Delta \; is \; NE \leftrightarrow \exists c \in \mathbb{R}, \hat{x}_i > 0, (A\hat{x})_i = c \rightarrow \forall ij, \hat{x} \in Z_{ij}$
	Evolutionary stable equilibrium (ESS)	$\hat{x} \in \Delta \; is \; ESS \overset{\text{def}}{=} \exists \epsilon > 0, x \cdot A(\epsilon x + (1-\epsilon)\hat{x}) < \hat{x} \cdot A(\epsilon x + (1-\epsilon)\hat{x})$ $\hat{x} \in \Delta \; is \; ESS \overset{\text{def}}{=} \exists \epsilon > 0, \forall y \in B_\epsilon(x), y \cdot Ay < \hat{x} \cdot Ay$ $\hat{x} \in int(\Delta) \; is \; an \; ESS \rightarrow \hat{x} \; is \; the \; unique \; NE$
	Coarse Correlated Equilibrium (CCE, no-regret set)	$(p_{ij}) \in CCE \leftrightarrow \sum_{i,j} a_{i'j}\, p_{ij} \le \sum_{i,j} a_{ij}\, p_{ij}$ and $\sum_{i,j} b_{ij'}\, p_{ij} \le \sum_{i,j} b_{ij}\, p_{ij}$
	Correlated Equilibrium (CE)	$(p_{ij}) \in CE \leftrightarrow \sum_{k} a_{i'k}\, p_{ik} \le \sum_{k} a_{ik}\, p_{ik}$ and $\sum_{l} b_{lj'}\, p_{lj} \le \sum_{l} b_{lj}\, p_{lj}$
		$\{strict \; NE\} \subset \{ESS\} \subset \{NE\}$ $\{NE\} \subset \{CE\} \subset \{CCE\}$
Replicator Dynamics (one population)	Dynamics	$\dot{x}_i = x_i\big((Ax)_i - x \cdot Ax\big), i = 1, \ldots, n$ $\frac{d}{dt}\frac{x_i}{x_j} = \frac{x_i}{x_j}\big((Ax)_i - (Ax)_j\big)$
	Convergence	**Every $n \times n$ replicator game has a Nash Equilibrium** NE: NE is an equilibrium of Replicator dynamics $\hat{x}$ is omega-limit of an orbit $x(t)$, $\hat{x} \in int(\Delta)$. Then $\hat{x}$ is NE. $\hat{x}$ is lyapunov stable. Then $\hat{x}$ is NE. ESS: ESS is asymptotically stable equilibrium $\hat{x} \in int(\Delta)$ is ESS, then $\hat{x}$ globally attracts all initial points $x \in int(\Delta)$
Replicator Dynamics two player	Dynamics	$P_A(x, y) = x \cdot Ay, \quad \mathcal{BR}_A(y) = \text{argmax}_{x \in \Delta_A} x \cdot Ay$ $P_B(x, y) = x \cdot By, \quad \mathcal{BR}_B(x) = \text{argmax}_{y \in \Delta_B} x \cdot By$ $\dot{x}_i = x_i\left(\left(Ay\right)_i - x \cdot Ay\right)$ $\dot{y}_j = y_j\left(\left(x^{tr}B\right)_j - x \cdot By\right)$
	Convergence	Each bimatrix game (A;B) has a Nash equilibrium
Best response Dynamics	Dynamics	$\dot{x} \in \mathcal{BR}(x) - x$; $\dot{x} = \mathcal{BR}(x) - x$, at differentiable x $x \rightarrow \mathcal{BR}(x)$ is upper semi-continuous Characteristic curve is continuous, almost everywhere differentiable
	Convergence	Assume that (A,B) is a **zero-sum game**. Best Response dynamics converge to the set of **Nash equilibria.**
Fictional play dynamics	Dynamics	$p(s) = \frac{1}{s}\int_0^s x(u)du; q(s) = \frac{1}{s}\int_0^s y(u)du; s = e^t$ $\dot{p}(s) = \frac{1}{s}x(s) - \frac{1}{s}p(s)$ and $\dot{q}(s) = \frac{1}{s}y(s) - \frac{1}{s}q(s)$ $x(s) \in \mathcal{BR}_A(q(s))$ and $y(s) \in \mathcal{BR}_B(p(s))$ for $s \ge 1$. $\dot{p}(s) \in \frac{1}{s}\big(\mathcal{BR}_A(q(s)) - p(s)\big)$; $\dot{q}(s) \in \frac{1}{s}\big(\mathcal{BR}_B(p(s)) - q(s)\big)$ $\dot{p}(t) = \big(\mathcal{BR}_A(q(t)) - p(t)\big); \dot{q}(t) = \big(\mathcal{BR}_B(p(t)) - q(t)\big).$

Convergence

Fictitious play converges to the no-regret set CCE
Assume that (A,B) is a zero-sum game. Fictional Play dynamics converge to the set of **Nash equilibria.**
Fiction Play orbits Pareto dominates Nash payoff
Time averages of Replicator Dynamics converge to pseudo-orbits of Fictitious Play

Reinforcement Learning

Dynamics

Model update

Cross Learning	$\theta^{t+1} = \left(1 - \vartheta u^t\right)\theta^t + \vartheta u^t x^t, \quad t \geq 1$
Erev-Roth Cumulative payoff matching (CPM)	$\theta^{t+1} = \theta^t + u^t x^t, \quad t \geq 1$
Arthur model	$\theta^{t+1} = \left(\theta^t + u^t x^t\right)\dfrac{C(t+1)}{Ct + u^t}, \quad t \geq 1$

Action choose

Proportional	$p(t) = Q^t / \lvert Q^t \rvert$
Greedy	$p(Q) = (1-\epsilon)\mathcal{B}R_I(Q) + \epsilon(1/n, \ldots, 1/n)$
Softmax	$\text{softmax}_T(Q) = \dfrac{1}{\sum_i \exp\left(Q_i/T\right)}\left(\exp\left(Q_1/T\right), \ldots, \exp\left(Q_n/T\right)\right)$

Q-learning

Dynamics

$$Q^{t+h}(s) = Q^t(s) + \alpha h \cdot \left(u_A^t + \gamma \max_j Q_j^t(s) - Q^t(s) \cdot a(t)\right) a(t)$$

$$\frac{dx_i}{dt} = x_i \tau \left(\frac{dQ_i}{dt} - \sum_j \frac{dQ_j}{dt} x_j\right)$$

$$\frac{dx_i}{dt} = x_i \tau \alpha \left(r_i - \sum x_j r_j + (1/\tau) \sum_j x_j \log\left(x_j / x_i\right)\right)$$

$$\frac{dx_i}{dt} = x_i \tau \alpha \left((Ay)_i - x \cdot Ay + (1/\tau)\left[-\log x_i + \sum_j x_j \log x_j\right]\right)$$

$$\frac{dy_i}{dt} = y_i \tau \alpha \left((Bx)_i - y \cdot Bx + (1/\tau)\left[-\log y_i + \sum_j y_j \log y_j\right]\right)$$

No-regret Learning

Dynamics

$$\text{SWAP}_A^i(j, k) = \begin{cases} e_k \cdot Ay^i & \text{if } x^i = e_j \\ x^i \cdot Ay^i & \text{if } x^i \neq e_j \end{cases}$$

$$\text{DIFF}_A^t(j, k) = \frac{1}{t}\left(\sum_{i=1}^{t}\left[\text{SWAP}_A^i(j, k) - x^i \cdot Ay^i\right]\right)$$

$$\text{REGRET}_A^t(j, k) = \max\left(\text{DIFF}_A^t(j, k), 0\right)$$

$$p_j^{t+1} = \frac{1}{\mu}\text{REGRET}_A^t(j^*, j) \text{for all } j \neq j^*$$

$$p_{j^*}^{t+1} = 1 - \sum_{j \neq j^*} p_j^{t+1} \text{ when } j = j^*$$

Convergence

(Hart and Mas-Colell). Provided we fix μ sufficiently large, if player A follows this algorithm then almost surely $\text{REGRET}_A^t(j, k) \to 0$
If two player play no-regret learning, the system converges to CE

Blackwall's Approachability

a convex set $C \in R^k$ is approachable for the vector payoff A if for each t and all probabilities $\{p^i, q^i\}_{i=1}^{t-1}$, there exists a choice p^t so that for each choice of q^t (which player A does not know before choosing pt), the vectors a_t converge to C as $t \to \infty$
For any closed convex set C the following are equivalent. 1. C is approachable for the vector payoff A; 2. for each q there exists p so that $A(p, q) \in C$ 3. every half space containing C is approachable.

Connections between Learning dynamics

FP vs. RD

The time average of a replicator orbit corresponds to a pseudo-orbit of fictitious play dynamics
Exists hyperbolic orbits in FP, then exists a corresponding orbits in RD

RL vs. FP

Reinforcement learning with choosing -greedy choices is very closely related to the type of dynamics one sees in Best Response dynamics and Fictitious Play.

RD vs. RL

Softmax Q-learning with $\alpha = 1/\tau \to 0$, the dynamics converges to the usual replicator dynamics
https://arxiv.org/abs/nlin/0408039