Distributed Bandit Online Tracking for Nash Equilibrium under Partial-decision Information Setting

2023 东南大学校庆研究生学术报告会

Zhangcheng Feng, Wenying Xu, and Jinde Cao

School of Mathematics
Southeast University

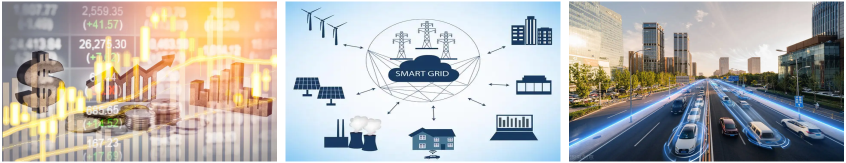
May 5, 2023

Outline

Table of Contents

Introduction-I

- **Game theory:** mathematical framework for modeling the decision-making process with selfish interest and interdependent action
- **Applications:** economy, smart grid, smart transportation...
- **Example:** multi-vehicle automated driving
  - optimize motion planning via intelligent actions *(e.g. acceleration/deceleration, lane change, overtaking...)*
  - $\begin{cases} \min_{x_i} \text{objective}_i(x_i, \boldsymbol{x}_{-i}) & \textit{(e.g. minimize travel time/fuel consumption)} \\ \text{s.t. } x_i \in \text{vehicle dynamics}_i & \\ \quad\;\; x_i \in \text{safety}_i(x_i, \boldsymbol{x}_{-i}) & \end{cases}$



(a) economy (b) smart grid (c) smart transportation

Introduction-II

- **Nash equilibrium (NE):** stable state where each player has no incentive to change decision individually
- **Classification:** Static game / online (dynamic) game
  - time-invariant / time-varying objective function
  - making decisions after / before knowing objective
- **Feedback model:** complete (explicit function form) / incomplete (bandit) (function values only)
- **Gradient estimator:** one-point / two-point / multi-point
- **Decision model:** full- / partial-decision information

Time-varying cost	Bandit feedback	Partial-decision	Comparable result with full-information feedback
✓	✓	✓	✓

Table of Contents

Online game-I

Essential elements:

- Game: $\Gamma\left(\mathcal{N},\{X_i\}_{i\in\mathcal{N}},\{f_{i,t}\}_{i\in\mathcal{N}}\right)$
- Player: $\mathcal{N}=\{1,\cdots,N\}$
- Decision: $x_i\in X_i\subset\mathbb{R}^{n_i}$, $x_{i,t}$ ——i's decision at time t
- Cost function: $f_{i,t}(x_i,\boldsymbol{x}_{-i}):X\subset\mathbb{R}^n\to\mathbb{R}, X=X_1\times\cdots\times X_N, n=\sum_{i=1}^N n_i$, where $\boldsymbol{x}_{-i}:=\mathrm{col}(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_N)$

Aim of each player: Selfishly minimizes own cost:

$$\forall i,\ \min_{x_{i,t}\in X_i} f_{i,t}(x_{i,t},x_{-i,t}).$$

Definition 1 (Nash equilibrium, NE)

An NE of the game $\Gamma\left(\mathcal{N},\{X_i\}_{i\in\mathcal{N}},\{f_{i,t}\}_{i\in\mathcal{N}}\right)$ *is a collective strategy* $\boldsymbol{x}_t^*=\mathrm{col}(x_{1,t}^*,\cdots,x_{N,t}^*)$ *such that for every* $i\in\mathcal{N}$

$$f_{i,t}(x_{i,t}^*,\boldsymbol{x}_{-i,t}^*)\le f_{i,t}(x_{i,t},\boldsymbol{x}_{-i,t}^*),\quad \forall x_{i,t}\in X_i.$$

Online game-II

Process of online game: At round $t \in \{1, \cdots, T\}$,

(i) player i makes a decision $x_{i,t}$ from X_i;

(ii) the environment selects a cost function $f_{i,t}$, and player i receives information about $f_{i,t}$.

Definition 2 (Regret)

$$Reg_i(T) := \sum_{t=1}^{T} [f_{i,t}(x_{i,t}, \boldsymbol{x}^*_{-i,t}) - f_{i,t}(x^*_{i,t}, \boldsymbol{x}^*_{-i,t})].$$

In general, an online algorithm is regarded as performing well if $Reg_i(T)$ *grows sublinearly with* T*, i.e., there exists* $\theta \in (0,1)$ *such that* $Reg_i(T) = \mathcal{O}(T^\theta)$*.* *(*$\frac{Reg_i(T)}{T} \to 0$*)*

Two-point bandit feedback

Each agent only observes the values of cost function at two points, thus the exact gradient information is **unavailable**.

Definition 3 (Gradient estimator & smoothed function)

Let $f : \mathbb{X} \subset \mathbb{R}^p \to \mathbb{R}$*, and assume that there exist positive constants* r *and* R *such that* $r\mathbb{B}^p \subseteq \mathbb{X} \subseteq R\mathbb{B}^p$*. (*$\mathbb{B}^p$*: unit ball,* $\mathbb{S}^p$*: unit sphere) Then,*

(i) *Two-point gradient estimator of* f *is defined as*

$$\hat{\nabla} f(x) := \frac{p}{\rho}[f(x+\rho u) - f(x)]u, \forall x \in (1-\eta)\mathbb{X},$$

where $u \in \mathbb{S}^p$ *is a uniformly distributed random vector,* $\rho \in (0, r\eta]$ *and* $\eta \in (0,1)$*;*

(ii) *Smoothed version of* f *is defined as*

$$\hat{f}(x) := \mathbf{E}_{v \in \mathbb{B}^p}[f(x+\rho v)], \forall x \in (1-\eta)\mathbb{X},$$

where the mathematical expectation is taken with respect to uniform distribution.

- ***Relation:*** $\mathbf{E}[\hat{\nabla} f(x)] = \nabla \hat{f}(x)$

Partial-decision

Estimation variable:

- $z_{ij,t}$: i's estimation of j's decision
- $z_{i,t} = \mathrm{col}(z_{i1,t}, \cdots, z_{iN,t})$: i's estimation of all other players
- $z_{i,-i,t} = \mathrm{col}(z_{ij,t})_{j \in \mathcal{N}/\{i\}}$: except i's decision

Communication network:

- Graph: $\mathcal{G} = (\mathcal{N}, \mathcal{E}), \mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$
- i can access to j's information $\Leftrightarrow (j, i) \in \mathcal{E}$, assume $(i, i) \in \mathcal{E}, \forall i$
- Directed path: $\{(i, e_1), (e_1, e_2), \cdots, (e_m, j)\}$
- Strongly-connected: exists a directed path from any node to all the others
- Weighted adjacency matrix: $W = [w_{ij}] \in \mathbb{R}^{N \times N}$, $w_{ij} > 0$ if $(i, j) \in \mathcal{E}$ and $w_{ij} = 0$ otherwise.

Assumption 1 (Graph & matrix)

The directed graph $\mathcal{G}$ is assumed to be strongly connected, and the corresponding adjacency matrix W is doubly stochastic, i.e., $\sum_{i=1}^{N} w_{ij} = \sum_{j=1}^{N} w_{ij} = 1, \forall i, j \in \mathcal{N}$.

Table of Contents

Algorithm

Algorithm 1 Partial-decision distributed online NE tracking with two-point bandit feedback

Initialize: $y_{i,0}, z_{ij,0} \in (1-\eta_0)X_i$, $u_{i,0} \in \mathbb{S}^{n_i}$, $x_{i,0} = y_{i,0} + \rho_0 u_{i,0}$.

Update: At each round $t \geq 0$, agent i

1. Receives $f_{i,t}(x_{i,t}, z_{i,-i,t})$ and $f_{i,t}(z_{i,t})$ after the decision $x_{i,t}$ and the estimation $z_{i,t} = (y_{i,t}, z_{i,-i,t})$ are updated;
2. Estimates the gradient by using two-point gradient estimator:

$$\hat{\nabla}_i f_{i,t}(z_{i,t}) := \frac{n_i}{\rho_t}[f_{i,t}(x_{i,t}, z_{i,-i,t}) - f_{i,t}(z_{i,t})]u_{i,t}, \tag{1}$$

where $u_{i,t} \in \mathbb{S}^{n_i}$ is a random vector of uniform distribution;

3. Updates its local variables according to the following rules:

$$y_{i,t+1} = \mathrm{proj}_{(1-\eta_{t+1})X_i}\left[y_{i,t} - \delta_{t+1}\hat{\nabla}_i f_{i,t}(z_{i,t})\right], \tag{2}$$

$$x_{i,t+1} = y_{i,t+1} + \rho_{t+1}u_{i,t+1}, \tag{3}$$

$$z_{ij,t+1} = \sum_{r=1}^{N} w_{ir} z_{rj,t} + \beta_i w_{ij}(y_{j,t} - z_{ij,t}), \tag{4}$$

where $\{\eta_t\}, \{\rho_t\}$ and $\{\delta_t\}$ are scalar sequences on $(0, 1]$ that do not increase, and $\beta_i > 0$ is a constant.

Assumptions

Assumption 2 (Function & sets)

(i) *Let sets* X_i ($\forall i \in \mathcal{N}$) *be nonempty and convex, and* $\exists\ r_i > 0$ *and* $R_i > 0$ *satisfy* $r_i\mathbb{B}^{n_i} \subseteq X_i \subseteq R_i\mathbb{B}^{n_i}$, *where the constants* $\{r_i\}_{i\in\mathcal{N}}$ *are known a prior.*

(ii) *For* $\forall t \in \mathbb{N}_+$ *and* $\forall i \in \mathcal{N}$, *let* $f_{i,t}(x_i, x_{-i})$ *be differentiable and convex in* $x_i \in X_i$ *for given* $x_{-i} \in \mathbb{R}^{n-n_i}$. *In addition, for any* (x_i, x_{-i}), $\exists\ M_x > 0, B_f > 0$ *and* $M_f > 0$ *satisfy*

$$\|x_i\| \leq M_x, \|f_{i,t}(x_i, x_{-i})\| \leq B_f,$$
$$\|\nabla_i f_{i,t}(x_i, x_{-i})\| \leq M_f,$$

where $\nabla_i f_{i,t}(x) := \frac{\partial f_{i,t}}{\partial x_i}(x)$.

(iii) $\nabla_i f_{i,t}(\cdot)$ ($\forall i \in \mathcal{N}$) *is Lipschitz continuous with a constant* $L > 0$, *i.e.,*$\forall x, y \in X$,

$$\|\nabla_i f_{i,t}(x) - \nabla_i f_{i,t}(y)\| \leq L\|x - y\|.$$

Assumption 3 (Game mapping)

The pseudo-gradient mapping of the online game $\Gamma(\mathcal{N}, \{X_i\}_{i\in\mathcal{N}}, \{f_{i,t}\}_{i\in\mathcal{N}})$, *defined as* $F_t(x) := col(\nabla_i f_{i,t}(x))_{i\in\mathcal{N}}$, *is* μ*-strongly monotone, i.e.,* $\forall x, y \in X$,

$$\langle F_t(x) - F_t(y), x - y\rangle \geq \mu\|x - y\|^2.$$

Regret bound-I

Theorem 1 (Dynamic regret bound)

Suppose Assumptions 1-3 hold. Let

$$\delta_t = \frac{1}{t^{\theta_1}}, \rho_t = \frac{r_{\min}}{(t+1)^{\theta_2}}, \eta_t = \frac{1}{(t+1)^{\theta_2}} \tag{5}$$

in Algorithm 1, where the constants $\theta_1, \theta_2 \in (0,1)$ *and* $r_{\min} := \min_{i\in\mathcal{N}}\{r_i\}$. *Then, for* $\forall i \in \mathcal{N}$,

$$\begin{aligned} \mathbf{E}[Reg_i(T)] \leq & \mathcal{O}\left(T^{\max\{1-\frac{\theta_1}{2}, \frac{1}{2}+\frac{\theta_1}{2}, 1-\frac{\theta_2}{2}, \frac{1}{2}\}}\right) \\ & + \mathcal{O}\left(\sqrt{T \log T}\right) + \mathcal{O}\left(T^{\frac{1}{2}+\frac{\theta_1}{2}}\sqrt{\Phi_T^*}\right), \end{aligned} \tag{6}$$

where $\Phi_T^* := \sum_{t=1}^{T} \|x_{t+1}^* - x_t^*\|$ *is the path-length (accumulated violation) of the NE sequence.*

Corollary 1 (Sublinear growing)

Under the conditions of Theorem 1, Algorithm 1 achieves sublinear $\mathbf{E}[Reg_i(T)]$ *if* $\Phi_T^* = \mathcal{O}(T^a)$ *for some constant* $a \in (0,1)$.

Regret bound-II

Corollary 2 (Optimal bound)

If the conditions of Theorem 1 hold, and let $\theta_1 = \frac{1}{2}$ *and* $\theta_2 \geq \frac{1}{2}$, *then one has*

$$\mathbf{E}[Reg_i(T)] \leq \mathcal{O}\left(T^{\frac{3}{4}}\right) + \mathcal{O}\left(T^{\frac{3}{4}}\sqrt{\Phi_T^*}\right).$$

Algorithms	Feedback model	Decision model	Gradient estimator	(Expected) Regret bounds
[1][1]	Full-information feedback	Partial-decision	Not used	$Reg_i(T) = \mathcal{O}\left(T^{\frac{1}{2}+\eta}\left(\sqrt{1+\Phi_T^*} + T^{\frac{1-3\eta}{2}}\right)\right), \eta \in (0, \frac{1}{2})$
[2][2]	Full-information feedback	Partial-decision	Not used	$Reg_i(T) = \mathcal{O}\left(T^{\frac{7}{8}} + T^{\frac{5}{6}}\sqrt{\Phi_T^*}\right)$
[3][3]	Bandit feedback	Full-decision	One-point	$\mathbf{E}[Reg_i(T)] = \mathcal{O}\left(T^{\frac{13}{14}} + T^{\frac{13}{14}}\sqrt{\Phi_T^*}\right)$
Algorithm 1	Bandit feedback	Partial-decision	Two-point	$\mathbf{E}[Reg_i(T)] = \mathcal{O}\left(T^{\frac{3}{4}} + T^{\frac{3}{4}}\sqrt{\Phi_T^*}\right)$

Table 1: Comparison of this paper to some related works on distributed online game

[1] Kaihong Lu, Guangqi Li, and Long Wang. "Online Distributed Algorithms for Seeking Generalized Nash Equilibria in Dynamic Environments". In: *IEEE Transactions on Automatic Control* 66.5 (2021), pp. 2289–2296. DOI: 10.1109/TAC.2020.3002592.

[2] Min Meng et al. "Decentralized Online Learning for Noncooperative Games in Dynamic Environments". In: *arXiv preprint arXiv:2105.06200* (2021). arXiv: 2105.06200. URL: http://arxiv.org/abs/2105.06200.

[3] Min Meng, Xiuxian Li, and Jie Chen. "Decentralized Nash Equilibria Learning for Online Game with Bandit Feedback". In: *arXiv preprint arXiv:2204.09467* (2022). arXiv: 2204.09467. URL: http://arxiv.org/abs/2204.09467.

Table of Contents

Simulation-I

Nash-Cournot game:

N firms, each firm $i \in \mathcal{N}$ participates in n_i markets and determines its production quantities $x_i \in \mathbb{R}^{n_i}$. $N = 5,\ n_i = 1.$

- Producing cost: $g_{i,t}(x_{i,t}) := x_{i,t}(\sin(t/12) + 1)$
- Commodity price: $p_{i,t}(x_{i,t}) := 22 + i/9 - 0.5i\sin(t/12) - \sum_{j=1}^{N} x_{j,t}$
- Cost function: $f_{i,t}(x_{i,t}, x_{-i,t}) = g_{i,t}(x_{i,t}) - x_{i,t}p_{i,t}(x_{i,t})$
- Communication graphs:

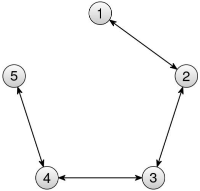


Figure 1: Fixed and strongly connected communication graph.

Simulation-II

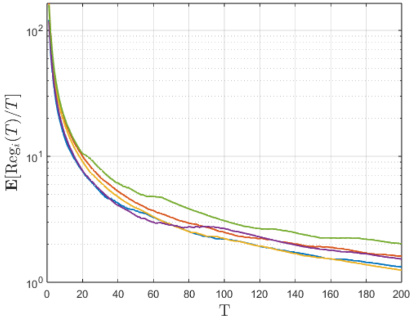


Figure 2: The trajectories of $\mathbf{E}[Reg_i(T)/T]$ generated by Algorithm 1.

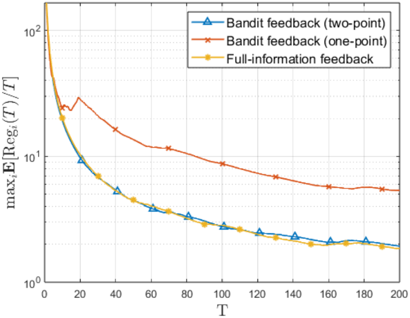


Figure 3: The trajectories of $\max_i \mathbf{E}[Reg_i(T)/T]$ generated by Algorithm 1 (two-point gradient estimator), algorithm with one-point gradient estimator and algorithm under full-information feedback.

Table of Contents

Conclusion

- **Problem setting:** NE tracking problem under bandit feedback & partial-decision information setting
- **Technique:** Employing two-point gradient estimator & leader-following consensus protocol
- **Result:** Improved theoretical results and simulation verification

Thank you for listening!

E-mail: 220211757@seu.edu.cn