Contextual Games:
Multi-Agent Learning with Side Information

Pier Giuseppe Sessa
Ilija Bogunovic
Andreas Krause
Maryam Kamgarpour

ETH zürich

NEURAL INFORMATION PROCESSING SYSTEMS

Contextual Games

- Novel class of repeated games with side information.

e.g., weather, traffic conditions, etc..

At each round t,

- Nature reveals **context** $z_t \in \mathcal{Z}$
- Players observe z_t and pick actions $a_t^1, \dots, a_t^N$
- Each player i obtains reward $r^i(a_t^i, a_t^{-i}, z_t), \quad i = 1,\dots,N$

Contextual regret of player i :

$$R_c^i(T) := \max_{\pi \in \Pi^i} \sum_{t=1}^{T} r^i(\pi(z_t), a_t^{-i}, z_t) - \sum_{t=1}^{T} r^i(a_t^i, a_t^{-i}, z_t)$$

set of **all** policies π mapping contexts to actions

- Standard notion in contextual bandits (e.g., [5])

- No assumption on the (potentially adversarial) contexts sequence $z_1, \dots, z_T$

Equilibria and Welfare

Def. Contextual Coarse Correlated Equilibrium (**c-CCE**): policy $\rho : \mathcal{Z} \to \Delta^{|\mathcal{A}^1 \times \cdots \times \mathcal{A}^N|}$ s.t. for each player $i = 1,\dots,N$:

$$\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\mathbf{a}\sim\rho(z_t)} r^i(\mathbf{a}, z_t) \geq \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\mathbf{a}\sim\rho(z_t)} r^i(\pi(z_t), a^{-i}, z_t) \quad \forall \pi \in \Pi^i$$

(Under ρ, no player has incentive in choosing any other policy π)

Def. Optimal Contextual Welfare:

$$\mathrm{OPT} = \max_{\pi^1 \in \Pi^1, \dots, \pi^N \in \Pi^N} \frac{1}{T}\sum_{t=1}^{T} \Gamma(\pi^1(z_t), \dots, \pi^N(z_t), z_t)$$

Game Welfare function

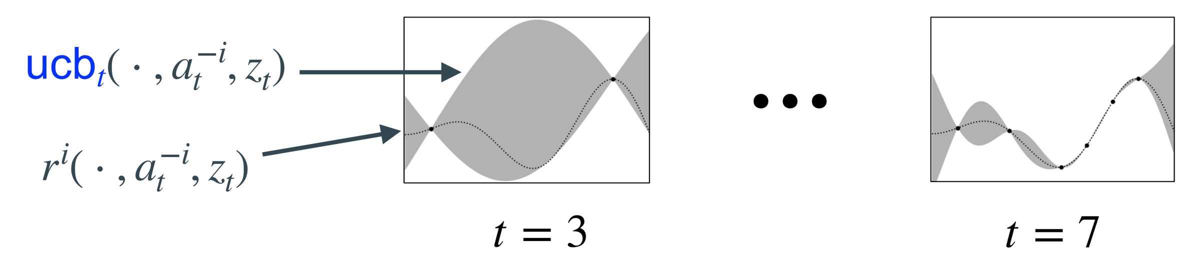
Thm (informal) When $R_c(T)/T \to 0, \forall i$, the game approaches a c-CCE and approximately optimal contextual welfare

- Extends main results [1,2] to the larger class of contextual games

No-Regret Strategies (for a generic player i)

Contextual game is a special *adversarial contextual bandit* problem with sequence of reward functions $\{r^i(\cdot, a_\tau^{-i}, z_\tau)\}_{\tau=1}^{T}$

—> Use **kernel-based regularity assumptions** on r^i and learn it using **kernel-ridge regression**:



c.GP-MW (meta) algorithm

Input: K actions, kernel function k
For $t = 1, \dots$:
- Observe context z_t
- Compute distribution $\mathbf{p}_t(z_t)$ using $\{\mathrm{ucb}_\tau(\cdot), a_\tau^{-i}, z_\tau\}_{\tau=1}^{t-1}$
- Sample action $a_t^i \sim \mathbf{p}_t(z_t)$
- Update $\mathrm{ucb}_t(\cdot)$ based on observed game data

Finite (small) number of contexts

Assume contexts set $\mathcal{Z}$ is finite

Strategy 1: $$\mathbf{p}_t(z_t)[a] \propto \exp\Big(\eta_t \cdot \sum_{\tau=1}^{t-1} \mathrm{ucb}_\tau(a, a_\tau^{-i}, z_\tau) \cdot \mathbf{1}\{z_\tau = z_t\}\Big)$$

Exploiting contexts similarity

$\mathcal{Z} \subseteq \mathbb{R}^c$ and assume r^i and optimal policy are Lipschitz w.r.t. $\mathcal{Z}$

Strategy 2: Iteratively build an ϵ-net of the contexts space and

$$\mathbf{p}_t(z_t)[a] \propto \exp\Big(\eta_t \cdot \sum_{\tau=1}^{t-1} \mathrm{ucb}_\tau(a, a_\tau^{-i}, z_\tau) \cdot \mathbf{1}\{z_\tau \in \mathrm{Ball}(z_t)\}\Big)$$

Stochastic and private contexts

Assume $z_t \sim \zeta$, and is private information to player i

Strategy 3: $$\mathbf{p}_t(z_t)[a] \propto \exp\Big(\eta_t \cdot \sum_{\tau=1}^{t-1} \mathrm{ucb}_\tau(a, a_\tau^{-i}, z_t)\Big)$$

Bounds on contextual regret

Strategy 1: $\mathcal{O}\big(\sqrt{T|\mathcal{Z}|\log K} + \gamma_T\sqrt{T}\big)$

Strategy 2: $\mathcal{O}\big(L^{\frac{c}{c+2}} T^{\frac{c+1}{c+2}} \sqrt{\log K} + \gamma_T\sqrt{T}\big)$

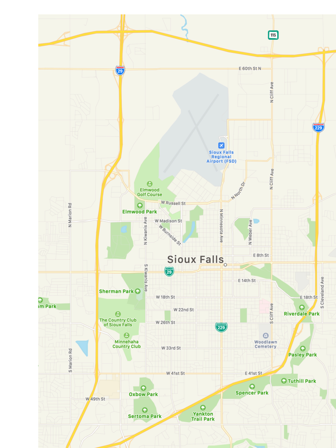
Strategy 3 (pseudo-regret): $\mathcal{O}\big(\sqrt{T\log K} + \gamma_T\sqrt{T}\big)$

Max info. gain [3] (e.g. $\gamma_T \leq \mathcal{O}((\log T)^{d+1})$ for SE kernels, d=input dim)

c.GP-MW exploits the correlation in the game and its regret scales only logarithmically with K

Contextual Traffic Routing Game

- Each agent i wants to send d_i units from origin O_i to destination D_i, $i = 1,\dots,528$

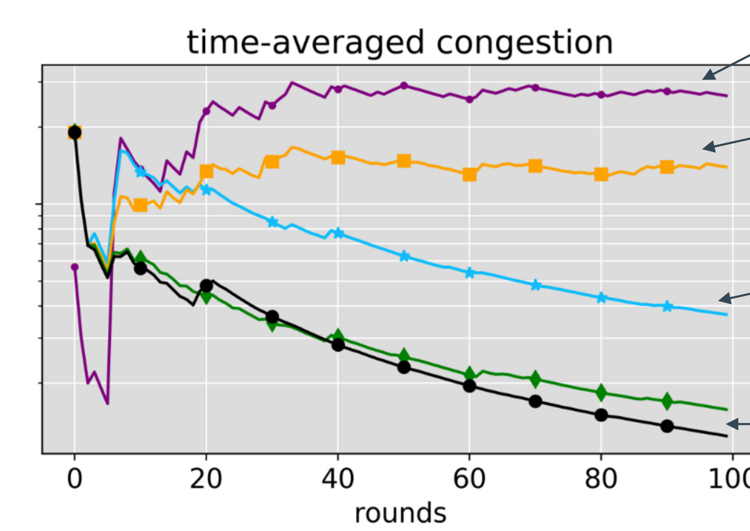


- $z_t \in \mathbb{R}^{76}$ = Network edges' capacity, randomly generated at each round

$$r^i(a_t^i, a_t^{-i}, z_t) = -\sum_{e=1}^{76} a_t^i[e] \cdot t_e(a_t^i + a_t^{-i}, z_t[e])$$

travel-time func. of edge e

- Sioux-Falls Network data and congestion model taken from [4]



References

[1] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 2000.

[2] T. Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM*, 2015.

[3] N. Srinivas, A. Krause, S. M Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML*, 2010.

[4] Transportation Network Test Problems. http://www.bgu.ac.il/ bargera/tntp/.

[5] G. Neu and J. Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual bandits. *CoLT*, 2020.

[6] P. G. Sessa, I. Bogunovic, M. Kamgarpour, and A. Krause. No-regret learning in unknown games with correlated payoffs. In *NeurIPS*, 2019.