No-Regret Learning in Unknown Games with Correlated Payoffs

Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, Andreas Krause

ETH Zürich

AUTOMATIC CONTROL LABORATORY

Learning & Adaptive Systems

Motivation

Consider *learning* to play an *unknown* repeated game.
- *Bandit* algorithms: slow convergence rates.
- *Full-information* algorithms: improved rates but are often unrealistic.

Under some regularity assumptions and a *new feedback model*, we propose **GP-MW** algorithm. **GP-MW** improves upon bandit regret guarantees while not relying on full-information feedback.

Set-Up

player i other players

- At each time t:
  - player i picks action $a_t^i \in \mathcal{A}_i$
  - other players pick actions a_t^{-i}
  - player i receives reward $r^i(a_t^i, a_t^{-i})$
- After T time steps, player i incurs **regret**:

$$R^i(T) = \max_{a \in \mathcal{A}^i} \sum_{t=1}^{T} r^i(a, a_t^{-i}) - \sum_{t=1}^{T} r^i(a_t^i, a_t^{-i})$$

- Reward function $r^i : \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \to [0,1]$ is **unknown**
- Each time t, player i **observes**:
  1) $\tilde{r}_t^i = r^i(a_t^i, a_t^{-i}) + \epsilon_t^i$, ϵ_t^i σ_i-sub-Gaussian   *(noisy bandit feedback)*
  2) a_t^{-i} *(actions of the other players)*
- Regularity (smoothness) assumption: $r^i(\cdot)$ has a bounded RKHS norm w.r.t. a kernel function $k^i(\cdot,\cdot)$

Key Idea

Use ***Gaussian Process (GP) confidence bounds*** *to* ***emulate the full-information*** *feedback:*

- Player i can use the observed data $\{a_\tau^i, a_\tau^{-i}, \tilde{r}_\tau^i\}_{\tau=0}^{t-1}$ to build a *shrinking* Upper Confidence Bound on $r^i(\cdot)$:

$$UCB_t(\cdot) = \mu_t(\cdot) + \beta_t^{1/2}\sigma_t(\cdot)$$

- $\mu_t(\cdot)$ and $\sigma_t(\cdot)$ are the *posterior mean* and *covariance* functions computed using standard **GP regression**.

Main Results

GP-MW algorithm for player i

Initialize mixed strategy : $\mathbf{w}_1 = [1/K_i, \dots, 1/K_i] \in \mathbb{R}^{K_i}$

For $t = 1, \dots, T$:
- *Sample action:* $a_t^i \sim \mathbf{w}_t$
- *Observe:* *noisy reward* $\tilde{r}_t^i$ & *opponents actions* a_t^{-i}
- *Compute optimistic full-info. feedback* $\mathbf{r}_t \in \mathbb{R}^{K_i}$:

$$\mathbf{r}_t[k] = \min\{UCB_t(a_k, a_t^{-i}), 1\}, \quad k = 1, \dots, K_i$$

- *Update mixed strategy :*

$$\mathbf{w}_{t+1}[k] \propto \mathbf{w}_t[k] \cdot \exp\left(\eta \cdot \mathbf{r}_t[k]\right), \quad k = 1, \dots, K_i$$

- *Update GP model based on the new observed data*

Def. *Maximum information gain:*

$$\gamma_T = \max_{x_1, \dots, x_T} I(\mathbf{r}_T; r^i)$$

Mutual information btw. $r^i(\cdot)$ and $\mathbf{r}_T = [r^i(x_t) + \epsilon]_{t=1}^T$

- γ_T grows with domain's dimension d. E.g., $\gamma_T = \mathcal{O}((\log T)^{d+1})$ for SE kernels

Theorem. Assume $\|r^i\|_{k^i} \leq B$. If player i uses **GP-MW**, with $\beta_t = B + \sqrt{2\gamma_{t-1} + \log(2/\delta)}$ and $\eta = \sqrt{(8 \log K_i)/T}$. Then, w.p. $(1-\delta)$,

$$R^i(T) = \mathcal{O}\left(\sqrt{T \log K_i} + B\sqrt{T\gamma_T} + \gamma_T\sqrt{T}\right)$$

- For $a^i \in \mathbb{R}^{d_i}$ and Lipschitz rewards: $R^i(T) = \mathcal{O}(\sqrt{d_i T \log(d_i T)} + \gamma_T \sqrt{T})$

Summary

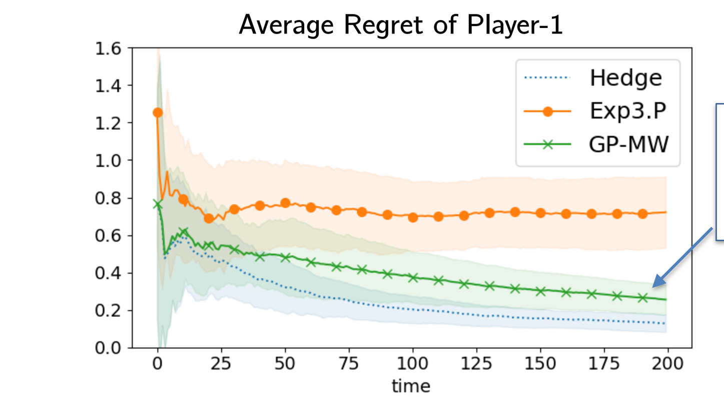
	Full-information	Bandit	Proposed model
Feedback:	$\{r^i(a, a_t^{-i}), \forall a \in \mathcal{A}^i\}$	$r^i(a_t, a_t^{-i})$	$r^i(a_t^i, a_t^{-i}) + \epsilon_t^i$, a_t^{-i}
Regret:	$\mathcal{O}(\sqrt{T \log K_i})$ Hedge [Freund and Schapire '97]	$\mathcal{O}(\sqrt{TK_i \log K_i})$ Exp3 [Auer et al. '02]	$\mathcal{O}(\sqrt{T \log K_i} + \gamma_T\sqrt{T})$ GP-MW [This paper]

Unrealistic feedback, since $r^i(\cdot,\cdot)$ is unknown

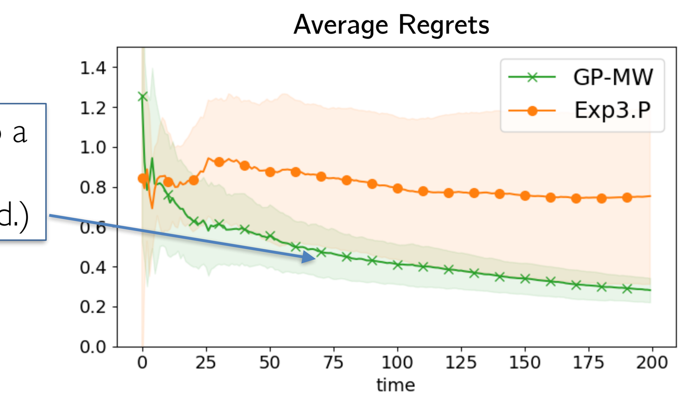
Scales badly with K_i

Experiments

- Random zero-sum games:



GP-MW leads to a smaller regret (shades = ± 1 std.)

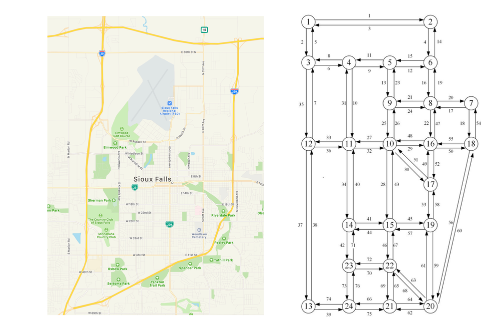


- Player-1 uses Hedge, Exp3.P, or GP-MW. Player-2 plays random actions.
- Player-1 uses GP-MW, Player-2 uses Exp3.P

- Repeated traffic routing:
  - 528 agents, $K_i = 5$ possible routes for each agent
  - Agents want to miminize traveltimes:

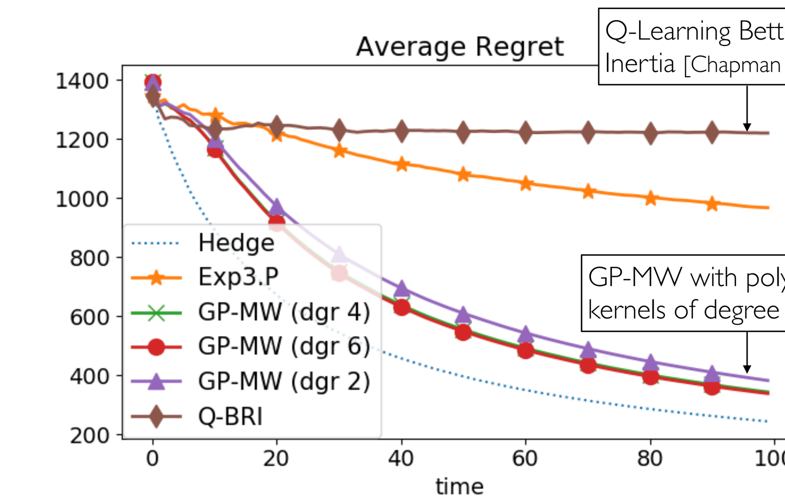
$$r^i(a^i, a^{-i}) = -\,\mathbf{traveltime}^i(a^i, a^{-i})$$

Simulated with BPR congestion model



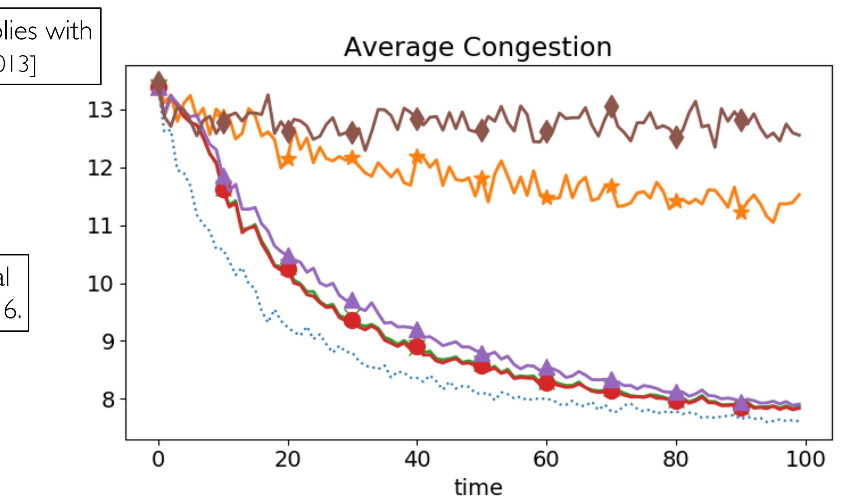
Sioux Falls Network [http://www.bgu.ac.il/ bargera/tntp/]

At every round each agent observes:
1) Incurred travel time, subject to noise
2) Total occupancy on each edge (i.e., $a_t^i + a_t^{-i}$)



Q-Learning Better Replies with Inertia [Chapman et al. 2013]

GP-MW with polynomial kernels of degree 2,4,or 6.



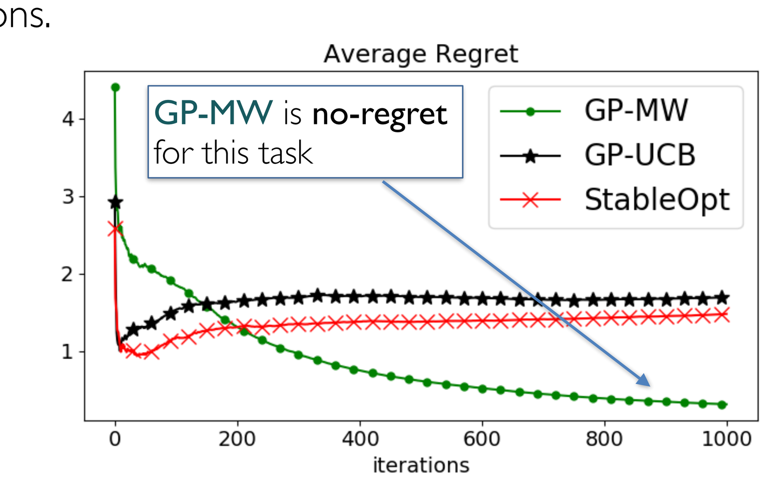
GP-MW leads to smaller regret and reduces the congestion

- Sequential movie recommendation: [https://grouplens.org/datasets/movielens/100k/.]

Don't know a-priori who will see our recommendations.

At every round:
- We select a movie a_t
- Adversary selects a user u_t
- Rating: $r(a_t, u_t) = a_t^\top \mathbf{r}_{u_t} + \epsilon$

Bayesian Optimization baselines:
- GP-UCB [Srinivas et al. 2010] : $a_t = \arg\max_a \max_u UCB_t(a, u)$
- StableOpt [Bogunovic et al. 2018] : $a_t = \arg\max_a \min_u UCB_t(a, u)$



GP-MW is no-regret for this task

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 2003.
[2] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997.

This work was gratefully supported by Swiss National Science Foundation, under the grant SNSF 200021_172781, and by the ERC grant 815943.

{sessap, ilijab, maryamk, krausea}@ethz.ch