No-Regret Learning in Unknown Games with Correlated Payoffs

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

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

- At each time t:
  - player i picks action $a_t^i \in \mathscr{A}_i$
  - 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**:

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

- Reward function $r^i : \mathscr{A}^i \times \cdots \times \mathscr{A}^N \to [0,1]$ 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 $r^i(\cdot)$:

$$UCB_{t-1}(\cdot) = \mu_{t-1}(\cdot) + \beta_t^{1/2}\sigma_{t-1}(\cdot)$$

- $\mu_{t-1}(\cdot)$ and $\sigma_{t-1}(\cdot)$ are the *posterior mean* and *covariance* functions computed using standard **GP regression**.

Main Results

GP-MW algorithm for player i

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

For $t = 1, \ldots, T$:

- *Sample action:* $a_t^i \sim \mathbf{w}_t$
- *Observe:* *noisy reward* $\tilde{r}_t^i$ + *opponents actions* a_t^{-i}
- *Compute optimistic full-information feedback* $\mathbf{r}_t \in \mathbb{R}^{K_i}$:

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

- *Update mixed strategy via MWU:*

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

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

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

$$R^i(T) = \mathcal{O}\left(\sqrt{T \log K_i} + \sqrt{T \log(2/\delta)} + B\sqrt{T\gamma_T} + \sqrt{T\gamma_T(\gamma_T + \log(2/\delta))}\right)$$

- Extension to continuous actions and Lipschitz rewards.
- γ_T is the *(kernel-dependent)* ***maximum information gain*** *about* $r^i(\cdot)$.
  E.g., for S.E. kernels $\gamma_T = \mathcal{O}((\log T)^{d+1})$, where d is the domain dimension.

Summary

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

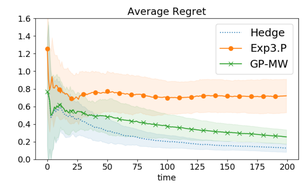
Unrealistic feedback, since $r^i(\cdot,\cdot)$ is unknown

Scales badly with K_i

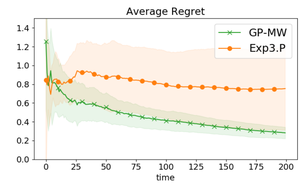
- Similar considerations apply to the continuous actions case.

Experiments

- Random matrix games:



Row player uses no-regret algorithms, column player plays random actions.

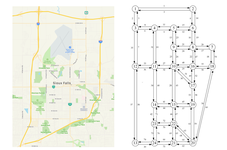


Row player uses GP-MW, column player uses EXP3.P [Auer et al. 2002].

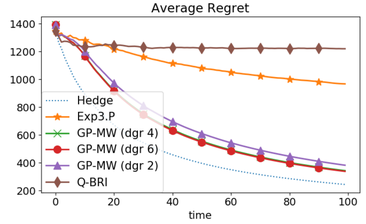
- Repeated traffic routing:

  - 528 agents in the network
  - Each agent wants to send d_i units from O_i to D_i
  - K_i = 5 possible routes for each agent.
  - Travel times computed using BPR congestion model [3]

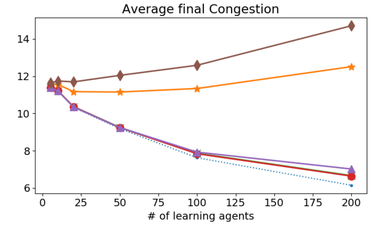
  At every round each agent observes:
  1) the incurred travel time
  2) $\psi(a_t^{-i})$ = Total occupancy on each edge caused by other agents.



Sioux Falls Network
Network data from [3]



N=100 *learning agents* implement no-regret algos. GP-MW leads to **smaller average regret**.



Increasing the number of *learning agents*, GP-MW **reduces the average congestion** (after T=100 iterations)

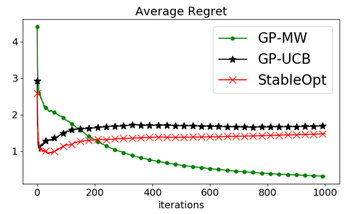
- Sequential movie recommendation:

[Movie-Lens 100k dataset]

We don't know a-priori which user will see our recommendations.

At every round: - We select a movie
- Adversary selects a user

Differently from existing Bayesian Optimization algorithms, **GP-MW is no-regret** for this task.



References and Acknowledgements

[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.
[3] Transportation test problems. http://www.bgu.ac.il/ bargera/tntp/.

This work was gratefully supported by Swiss National Science Foundation, under the grant SNSF 200021_172781, and by the European Union's Horizon 2020 ERC grant 815943.

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