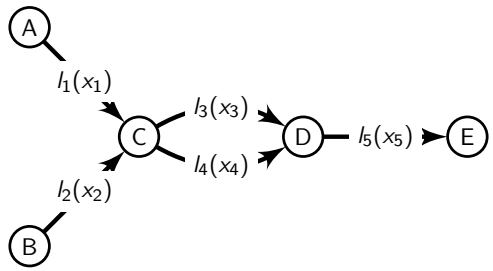
An Online Learning Approach to a Multi-player N-armed Functional Bandit

Sam O'Neill, Ovidiu Bagdasar, Antonio Liotta

University of Derby

June 15 – 21, 2019

Numerical Computations: Theory and Algorithms
The 3rd International Conference and Summer School



Overview

1. Routing Games
2. Bandit Machines
3. Online Learning for Multi-player N-armed Functional Bandits
4. Results
5. Conclusions/Future Work

Route N players between their origin/destination such that each player has no incentive to change path.

> *"The journey times on all routes are equal, and less than those which would be experienced by a single vehicle on any unused route"*
>
> [Wardrop, 1952]

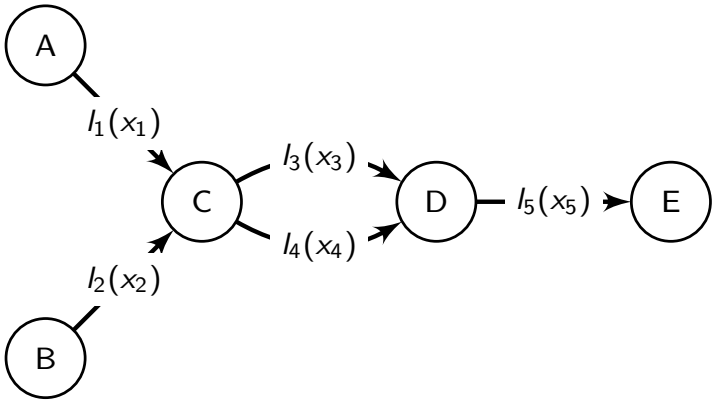
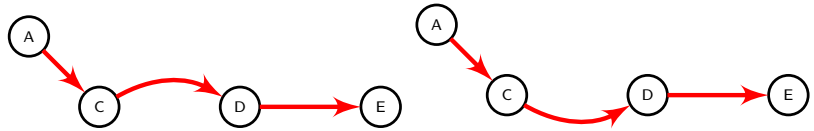


Figure 1: [Sheffi, 1985]

Example latency functions

$$l_1(x_1) = 1$$
$$l_2(x_2) = 2$$
$$l_3(x_3) = 2 + x_3$$
$$l_4(x_4) = 1 + 2x_4$$
$$l_5(x_5) = 1$$

Paths between A and E



Paths between B and E

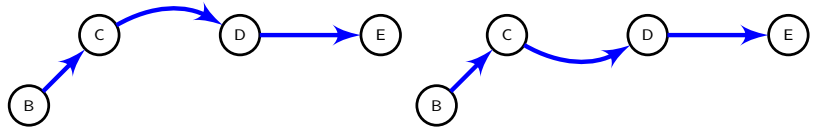


Figure 2: Strategy sets for (A,E) and (B,E)

Notation

$G = (V, E)$ - Graph

$\mathcal{N} = \{1, \ldots, N\}$ - Set of N players

x_e - Number of players using edge e

$l_e : \mathbb{N} \to \mathbb{R}$ - Edge latency functions

$l = (l_e)_{e \in E}$ - Vector of all edge latency functions

(o_i, d_i) - Origin/destination pair for player i

$a_i \subseteq E$ - A strategy (path) for player i

$\mathcal{A}_i$ - Set of strategies (paths) for player i, $a_i \in \mathcal{A}_i$

$\mathcal{A} = \prod_{i \in \mathcal{N}} \mathcal{A}_i$ - Set of all strategy profiles

$a_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N)$ - Strategies of other players

$\mathbf{a} = (a_i; a_{-i}) = (a_1, \ldots, a_i, \ldots a_N) \in \mathcal{A}$ - Strategy profile

The number of players choosing edge e,

$$x_e = \sum_{i \in \mathcal{N}} \sum_{e_i \in a_i} \mathbb{1}(e_i = e).$$

Cost to player i picking strategy $a_i \in \mathcal{A}_i$,

$$u_i(a_i; a_{-i}) = \sum_{e \in E} \sum_{e_i \in a_i} \mathbb{1}(e_i = e) \cdot l_e(x_e).$$

Total Cost of strategy profile $\mathbf{a}$,

$$U(\mathbf{a}) = \sum_{i \in \mathcal{N}} u_i(a_i; a_{-i})$$

Strategy (a_i, a_{-i}) is a Nash equilibrium if,

$$u_i(a_i; a_{-i}) \leq u_i(a_i'; a_{-i}) \quad \forall a_i' \in \mathcal{A}; \forall i \in \mathcal{N}$$

- Routing games belong to a broader class known as congestion games
- Congestion games have at least one pure strategy Nash equilibrium [Rosenthal, 1973]
- Found by minimising the potential function

$$\Phi = \sum_{e \in E} \sum_{i=1}^{x_e} l_e(x_e)$$

- Can be approximated by minimising the function

$$\Phi \approx \sum_{e \in E} \int_0^{x_e} l_e(\omega) d\omega$$

- If the functions l_e are convex then the optimal solution to the approximation is unique in terms of edge (x_e) loadings [Beckmann et al., 1956]

- Strategy set (possible paths) grows exponentially
- Many path solutions can result in equilibrium
- Unrealistic paths employed in the solution
- Traditional methods employ full information
- User equilibrium is a strong assumption, lacks consideration of 'human factors'. Realism?
- A player's appetite for exhaustive search [Gigerenzer and Selten, 2001]

> *"The capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problems whose solution is required for objectively rational behavior in the real world — or even for a reasonable approximation to such objective rationality."*

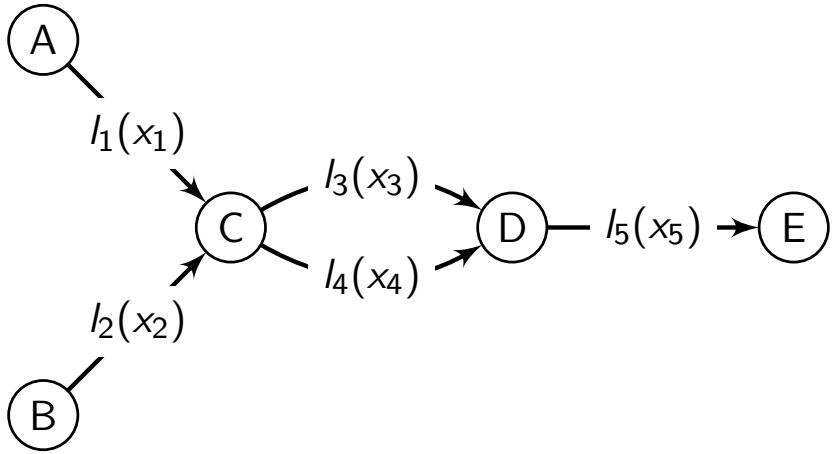
[Simon, 1957]

eg. restrict the strategy set of a player to 'reasonable' paths.

Figure 3: Bandit Machines

- K one armed bandit machines with associated rewards
- Player pulls one bandit per round and receives a reward
- Player objective: Maximise cumulative reward after a number of rounds
- Exploration 'vs' Exploitation
- A good online learning algorithm seeks to maximise cumulative reward

[Cesa-Bianchi and Lugosi, 2006]



- $\mathcal{N}$ players
- Graph $G = (V, E)$
- A set E of edges with latency functions $l_e(x_e)$
- Each round each player $i \in \mathcal{N}$ pulls a bandit, a combination of edges $a_i \subseteq E$ (path) from the strategy set $\mathcal{A}_i$ (set of **reasonable** paths for (o_i, d_i) pair)
- Players move **simultaneously** each round
- Players either receive the played path or all available path costs (bandit and semi-bandit feedback)
- Players seek to minimise their cumulative travel time over T rounds

Bandit Feedback

After a player makes their bandit (path) choice, they observe only the corresponding reward.

$$\hat{u}_i^t = \hat{r}_{ia_i}(a_{-i}^t) = u_i(a_i^t; a_{-i}^t) + \xi_{ia_i}^t$$

Semi-bandit Feedback

After a player makes their bandit (path) choice, they are allowed to observe all possible rewards.

$$\hat{\mathbf{r}}_\mathbf{i}^\mathbf{t} = (\hat{r}_{ia_i}(a_{-i}^t))_{a_i \in \mathcal{A}_i}$$

[Belmega et al., 2018]

Exponential weights with semi-bandit feedback

Algorithm 1 Exponential weights with semi-bandit feedback [EW-SB]

Require: $\gamma_t = t^{-\frac{1}{\alpha}}\ \forall t \in [1, \dots, T]$, $W_i^0 \in \mathbf{1}^{|\mathcal{A}_i|}\ \forall i \in \mathcal{N}$

for $t = 1, \dots, T$ **do**
  for each player i in $\mathcal{N}$ **do**
    $\mathbf{P}_i^t = \frac{W_i^t}{\sum_{j=1}^{|\mathcal{A}_i|} W_{ij}^t}$ ▷ Calculate probability distribution for strategies
    $a_i^t \sim \mathbf{P}_i^t$ ▷ Sample action from probability distribution
  end for
  for each player i in $\mathcal{N}$ **do**
    $\hat{\mathbf{r}}_\mathbf{i}^\mathbf{t} = (\hat{r}_{ia_i}(a_{-i}^t))_{a_i \in \mathcal{A}_i}$ ▷ Observe estimated reward for strategies
    $W_i^{t+1} = W_i^t \cdot \exp\left(\frac{\gamma_t \hat{\mathbf{r}}_\mathbf{i}^\mathbf{t}}{|\mathcal{A}_i|}\right)$ ▷ Update weights
  end for
end for

- Based on Hedge/Exponential Weights algorithm
- Each player maintains a belief about a set of strategies
- Players sample based on these beliefs
- Beliefs updated via semi-bandit feedback

Exponential weights with bandit feedback

Algorithm 3 Exponential weights with bandit feedback [EW-B]

Replace lines $7-8$ in algorithm 1 with:

$\hat{u}_i^t = u_i(a_i^t; a_{-i}^t) + \xi_i^t$ ▷ Observe estimated feedback for played strategy

$$\hat{r}_{ia_i}^t = \begin{cases} \frac{\hat{u}_i^t}{\mathbf{P}_{ia_i}^t}, & \text{if } a_i = a_i^t. \\ 0, & \text{otherwise.} \end{cases} \quad \forall a_i \in \mathcal{A}_i$$

▷ Estimate feedback vector $\hat{\mathbf{r}}_\mathbf{i}^\mathbf{t}$

$W_i^{t+1} = W_i^t \cdot \exp\left(\frac{\gamma_t \hat{\mathbf{r}}_\mathbf{i}^\mathbf{t}}{|\mathcal{A}_i|}\right)$ ▷ Update weights

- Beliefs updated via bandit feedback
- Feedback vector for all strategies estimated via importance sampling
- Can be shown to be an unbiased estimator [Cohen et al., 2017]

ϵ-greedy with semi-bandit feedback

Algorithm 2 ϵ-greedy with semi-bandit feedback [ϵG-SB]

Require: $W_i^0 \in \mathbf{0}^{|\mathcal{A}_i|}\ \forall i \in \mathcal{N}$

for $t = 1, \ldots, T$ **do**
  for each player i in $\mathcal{N}$ **do**
    if $\epsilon_t \sim unif(0,1) < \epsilon$ **then**
      $a_i^t \sim unif\{1, |\mathcal{A}_i|\}$ ▷ Choose at random with probability $p = \frac{1}{|\mathcal{A}_i|}$
    else
      $a_i^t = \arg\max_{a_i \in \mathcal{A}_i}(W_{ia_i}^t)$
    end if
  end for
  for each player i in $\mathcal{N}$ **do**
    $\hat{\mathbf{r}}_\mathbf{i}^\mathbf{t} = (\hat{r}_{ia_i}(a_{-i}^t))_{a_i \in \mathcal{A}_i}$ ▷ Observe estimated feedback for strategies
    $W_i^{t+1} = W_i^t + \frac{1}{t+1}\left[\hat{\mathbf{r}}_\mathbf{i}^\mathbf{t} - W_i^t\right]$ ▷ Update average feedback
  end for
end for

- Each player maintains an average estimate of their strategies
- Players greedily choose the best strategy with probability $1 - \epsilon$
- Averages updated via semi-bandit feedback

ϵ-greedy with bandit feedback

Algorithm 4 ϵ-greedy with bandit feedback [ϵG-B]

Replace lines $10-11$ in algorithm 2 with:

$\hat{u}_i^t = u_i(a_i^t; a_{-i}^t) + \xi_i^t$ ▷ Observe estimated feedback for played strategy

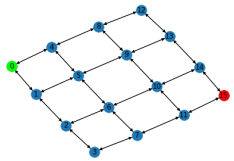
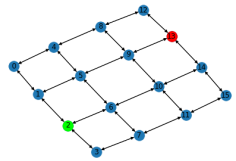
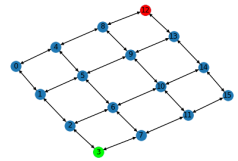
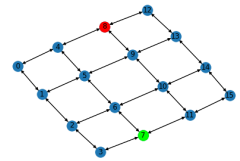
$$\hat{r}_{ia_i}^t = \begin{cases} \frac{1}{t+1}\left[\hat{u}_i^t - W_{ia_i}^t\right], & \text{if } a_i = a_i^t. \\ 0, & \text{otherwise.} \end{cases} \quad \forall a_i \in \mathcal{A}_i$$

▷ Estimate feedback vector $\hat{\mathbf{r}}_\mathbf{i}^\mathbf{t}$

$W_i^{t+1} = W_i^t + \hat{\mathbf{r}}_\mathbf{i}^\mathbf{t}$ ▷ Update average rewards

- Only the average estimate of the strategy played is updated

- Tested on a bidirectional lattice network with 16 vertices and 48 edges
- 250 players were routed between 4 origin destination pairs

- Edge latency function of the form,

$$l_e(x_e) = a_e + b_e \left(\frac{x_e}{c_e} \right)^{n_e}$$

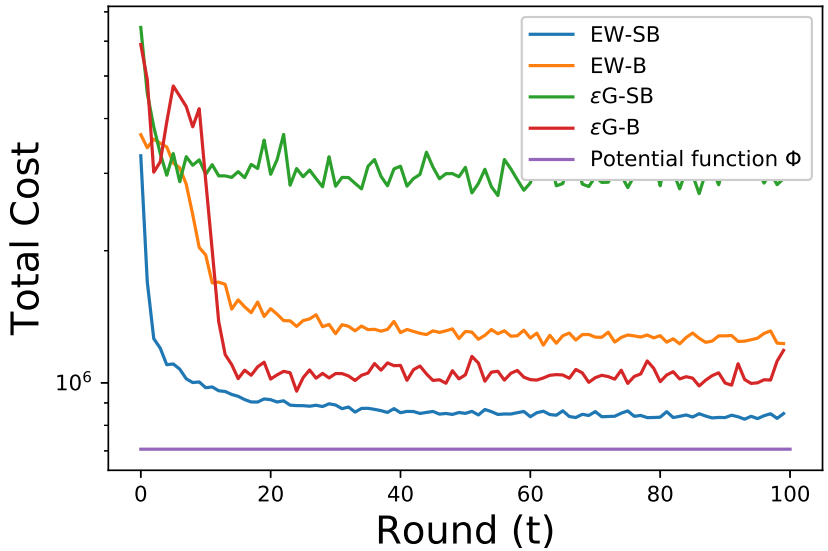
- 10 randomly generated instances of the lattice network
- Each player given the 10 cheapest routes at $t = 1$ as a strategy set
- 10 episodes of each instance
- Each episode consisted of a 100 rounds $T = 100$
- Tuned hyperparameters - $\alpha = 4$, $\epsilon = .3$

Github Repository - https://bit.ly/2WGCNJr

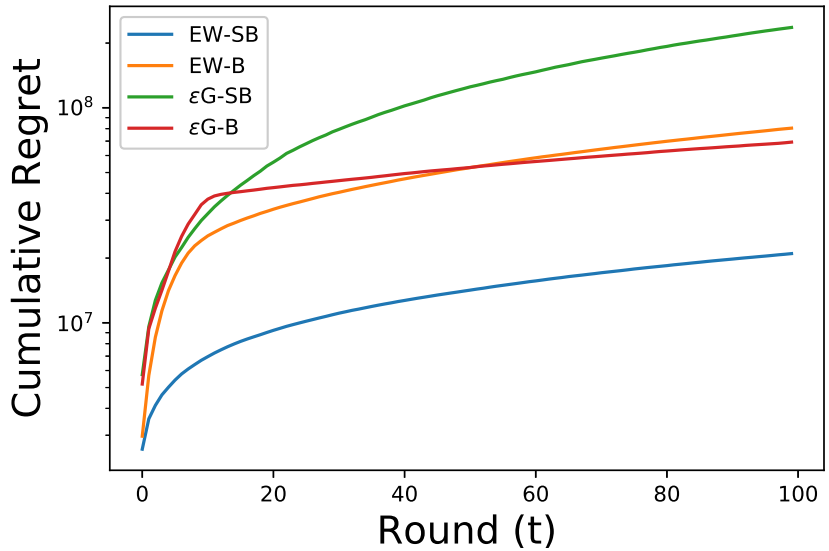
Cumulative regret of an algorithm at time t,

$$R_t = \sum_{s=1}^{t} [U(\mathbf{a}^{\mathbf{s}}) - U_{\Phi}]$$

where $U(\mathbf{a}^{\mathbf{s}})$ is the total cost at time s of strategy $\mathbf{a}$ and U_{Φ} is the total cost of the Nash equilibrium given by minimising Φ

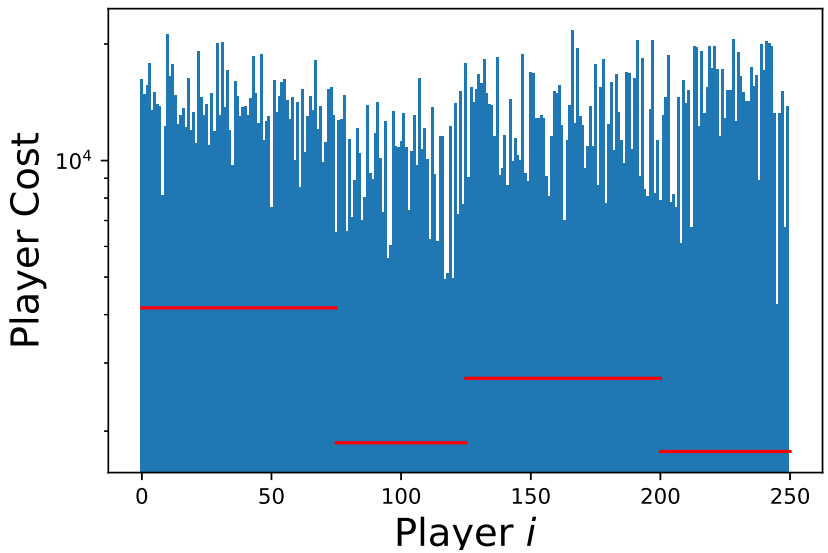


(a) Total Cost

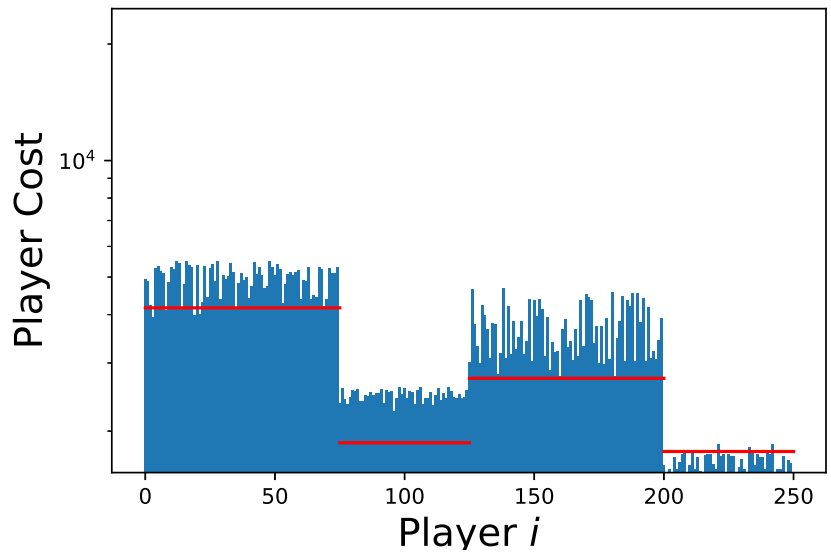


(b) Cumulative Regret

Towards Equilibrium



(a) Player costs after initial round



(b) Player costs after T rounds

Figure 5: Log-lin plots illustrating the convergence of players costs for the 4 origin/destination pairs (EW-SB)

Conclusions

- An equilibrium pattern emerges from multi-player interaction
- Equilibrium is similar to the theoretical one given by Φ
- ϵ-greedy algorithm performs poorly under semi-bandit feedback as averages become skewed
- Results are reasonable for both algorithms with bandit feedback which most resembles human decision making
- Inefficient in terms of computational complexity

Future Work

- Complexity analysis
- Regret analysis on bounded rewards
- Incorporation of other 'behavioural traits'
- Epsilon-optimization


Beckmann, M. J., McGuire, C. B., and Winsten, C. B. (1956).
Studies in the Economics of Transportation.
WileyRoyal Economic Society.


Belmega, E. V., Mertikopoulos, P., Negrel, R., and Sanguinetti, L. (2018).
Online convex optimization and no-regret learning: Algorithms, guarantees and applications.


Cesa-Bianchi, N. and Lugosi, G. (2006).
Prediction, learning, and games.
Cambridge University Press.


Cohen, J., Héliou, A., and Mertikopoulos, P. (2017).
Learning with bandit feedback in potential games.


Gigerenzer, G. and Selten, R. (2001).
Bounded rationality : the adaptive toolbox.
MIT Press.


Rosenthal, R. W. (1973).
A class of games possessing pure-strategy Nash equilibria.
International Journal of Game Theory, 2(1):65–67.


Sheffi, Y. (1985).
Urban transportation networks.
Prentice-Hall.


Simon, H. A. (1957).
Models of man: social and rational : mathematical essays on rational human behavior in a social setting.
John Wiley and Sons, New York.


Wardrop, J. G. (1952).
Road Paper. Some Theoretical Aspects Of Road Traffic Research.
Proceedings of the Institution of Civil Engineers, 1(3):325–362.