Decentralized learning implications on the performance of dense WLANs

Universitat Pompeu Fabra Barcelona

Francesc Wilhelmi

Co-authors: B. Bellalta, C. Cano, G. Neu, A. Jonsson & S. Barrachina-Muñoz

Geneva, 30/01/18

Table of contents

Outline

Problem description

Spatial Reuse (SR) enhancement in dense Wireless Networks

- Transmit Power Control (TPC)
- Carrier Sense Threshold (CST) adjustment
- Dynamic Channel Selection (DCA)

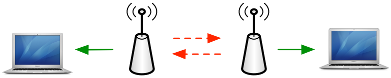


Figure 1: Limited performance

Figure 2: Enhanced spatial reuse

Context - Use case

- Dense IEEE 802.11 WLANs
  - Unplanned (chaotic deployments)
  - Decentralized (local information only)
- Online learning through adversarial Multi-Armed Bandits (MABs)

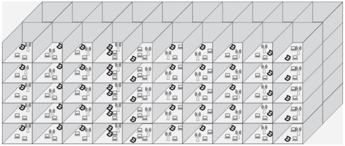


Figure 3: TGax residential scenario. Image retrieved from [1].

Why MABs?

1. Uncertainty → no information exchange
2. Adversarial setting → the reward is influenced by the environment
3. Complex interactions:

Action	**Effect**			
	Parallel Transmissions	Data Rate	Collisions probability (by hidden node)	Energy Consumption
↑ Power	↓	↑	↓	↑
↓ Power	↑	↓	↑	↓
↑ CCA	↑	-	↑	↑*
↓ CCA	↓	-	↓	↓*

Table 1: Effects of TPC and CST adjustment

Need to find an **approximation** of the optimal solution, rather than computing it.

Outline

Related Work

Surveys

- Self-Organized Networks (SONs) [1]
- Cognitive radio [2]
- Wireless Sensor Networks (WSN) [3, 4]
- Ad-hoc networks [5]

Related to this problem

- Q-learning for channel selection [6-9] and power adjustment [10, 11]
- MABs to Power control in D2D networks [12, 13]
- MABs to DCA & TPC [14]
- Structured MABs for combinatorial optimization problems [15, 16]
- MABs for decentralized channel access [17, 18]

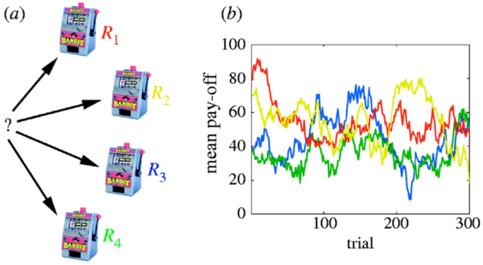
Outline

The Multi-Armed Bandit problem

Formal definition

A game in which the following steps are repeated in $t = 1, 2, \ldots, T$:

1. The environment fixes an assignment of rewards $r_{a,t}$ for each action $a \in [K] \stackrel{\text{def}}{=} \{1, 2, \ldots, K\}$,
2. the learner chooses action $a_t \in [K]$,
3. the learner obtains and observes reward $r_{a_t,t}$



MABs application into Decentralized WLANs

Use case

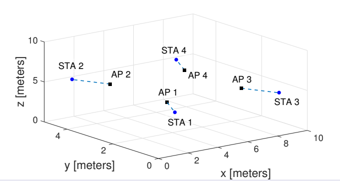
- Adversarial setting (N WLANs make actions simultaneously)
- Actions consist in {channel, tx. power, CCA} combinations
- The reward is **selfishly** set as the own throughput
- Action-selection procedure: Thompson sampling [19]

Outline

Selfish learning - Scenario

Symmetric grid

- Grant equal opportunities to WLANs
- Study characteristic types of interaction
  - Three variations of the scenario
  - Different spatial distributions and possible configurations



Optimal solutions

- **S1:** all WLANs must use CCA_{max}
- **S2:** all WLANs must use CCA_{max} and $Power_{min}$
- **S3:** all WLANs must listen to the others

Selfish learning - Average throughput

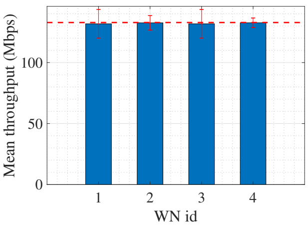


Figure 4: S1

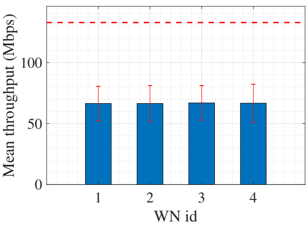


Figure 5: S2

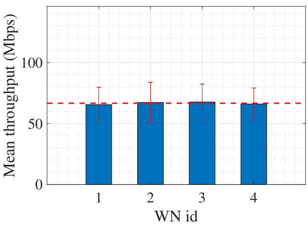


Figure 6: S3

Convergence in terms of average throughput is reached in all the cases. However, it does not always match with the optimal solution.

Selfish learning - Individual throughput evolution

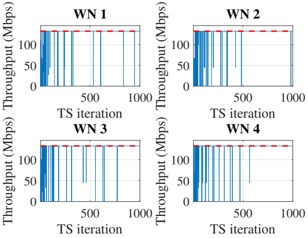


Figure 7: S1

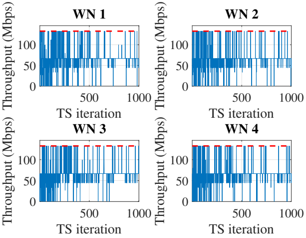


Figure 8: S2

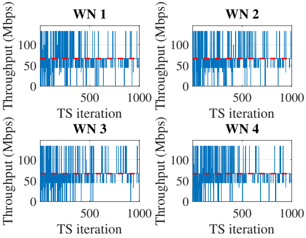


Figure 9: S3

A higher variability is observed in S2 and S3 (WLANs alternate good and bad performance).

Selfish learning - Actions probabilities

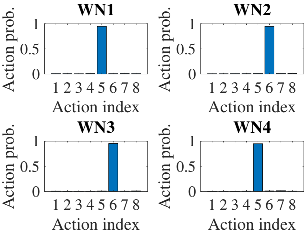


Figure 10: S1

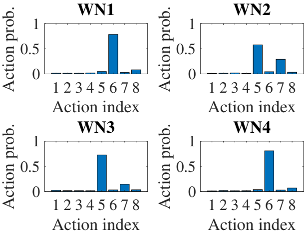


Figure 11: S2

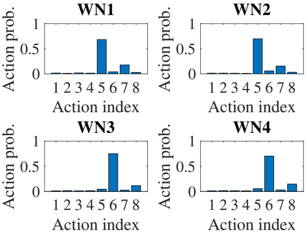


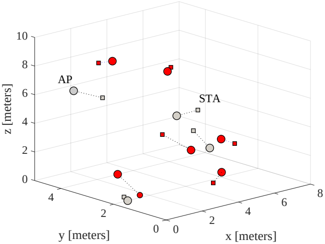
Figure 12: S3

WLANs in S1 rapidly achieve a single preferred action, while in S2 and S3 they use more actions.

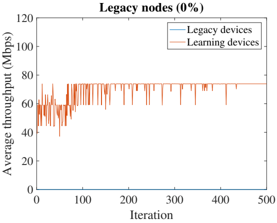
Impact on legacy networks - Scenario

Random scenario

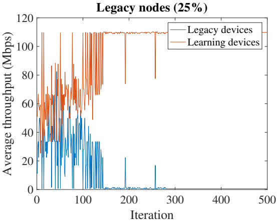
- We test different % of legacy networks (randomly chosen)
- Goal: to study the performance of both learning and legacy networks



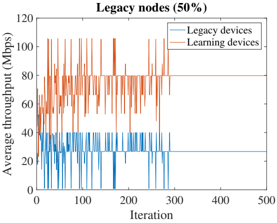
Impact on legacy networks - Results



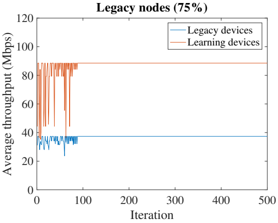
(a) 0/8 legacy (0%)



(b) 2/8 legacy (25%)



(c) 4/8 legacy (50%)



(d) 6/8 legacy (75%)

Outline

Decentralized learning

Challenges

- Finding equilibriums
- Fairness and asymmetries in a network

Opportunities

- Behavior inference (work in progress)
- Usage of constraints to favor fairness
- Exploitation of problem's characteristics for fast convergence (contextual, combinatorial bandits)

Centralized and Collaborative learning

Challenges

- Increased complexity
- Communication overheads (is it worth?)

Opportunities

- More control
- Less variability
- Correlated equilibria

Any questions?

?

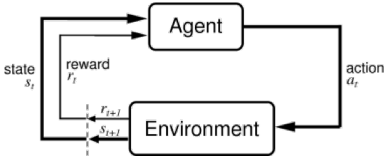
Francesc Wilhelmi
francisco.wilhelmi@upf.edu
PhD student
Department of Communication and Information Technologies
Universitat Pompeu Fabra (Barcelona)

Backup: Reinforcement Learning

Goal

An agent attempts to learn a policy given the observations it does. The goal is to maximize the expected future cumulative reward.

- No supervisor (only reward signal)
- Delayed feedback & sequentiality
- Actions affect the environment



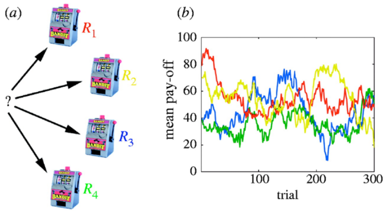
$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}\}$

- $\mathcal{S}$: set of states
- $\mathcal{A}$: set of actions
- $\mathcal{R}$: set of rewards
- $\mathcal{T}$: transitions probabilities

Backup: Multi-Armed Bandits

Frames the exploration/exploitation trade-off. The hidden reward distributions must be learned while maximizing the gains.

- Action-selection strategies to cope with hidden distributions (ε-greedy, EXP3, UCB…)
- Several variants (contextual, adversarial, stochastic, restless…)
- States-independent
- Reward becomes regret: $R_n = \sum_{t=1}^{n} l_{t,I_t} - \min_{i \in K} \sum_{t=1}^{n} l_{t,i}$



Backup: Thompson sampling

Thompson sampling [19] is a Bayesian action-selection technique

- It constructs a probabilistic model of the rewards and assumes a prior distribution of the parameters of said model
- Keeps track of the posterior distribution of the rewards, and pulls arms randomly in a way that the drawing probability of each arm matches the probability of the particular arm being optimal
- For the sake of practicality, we aim to apply Thompson sampling using a Gaussian model for the rewards with a standard Gaussian prior as suggested in [20].
- In adversarial wireless networks, it has been shown to perform better than using the magnitude of the reward [9]

Backup: Applied Thompson sampling

Algorithm 1: Implementation of Multi-Armed Bandits (Thompson sampling) in a WN

```
Function Thompson Sampling (SNR, 𝒜);
Input  : SNR: information about the Signal-to-Noise Ratio received at the STA
         𝒜: set of possible actions in {a_1, ..., a_K}
initialize: t = 0, for each arm a_k ∈ 𝒜, set r̂_k = 0 and n_k = 0
while active do
    For each arm a_k ∈ 𝒜, sample θ_k(t) from normal distribution 𝒩(r̂_k, 1/(n_k+1))
    Play arm a_k = argmax_{k=1,...,K} θ_k(t)
    Observe the throughput experienced Γ_t
    Compute the reward r_{k,t} = Γ_t/Γ*, where Γ* = B log_2(1 + SNR)
    r̂_{k,t} ← (r̂_{k,t} n_{k,t} + r_{k,t}) / (n_{k,t} + 2)
    n_{k,t} ← n_{k,t} + 1
    t ← t + 1
end
```

References

[1] Afaqui, M. Shahwaiz, et al. "Evaluation of dynamic sensitivity control algorithm for IEEE 802.11 ax." Wireless Communications and Networking Conference (WCNC), 2015 IEEE. IEEE, 2015.

[2] KLAINE, Paulo Valente, et al. A Survey of Machine Learning Techniques Applied to Self-Organizing Cellular Networks. IEEE Communications Surveys & Tutorials, 2017, vol. 19, no 4, p. 2392-2431.

[3] BKASSINY, Mario; LI, Yang; JAYAWEERA, Sudharman K. A survey on machine-learning techniques in cognitive radios. IEEE Communications Surveys & Tutorials, 2013, vol. 15, no 3, p. 1136-1159.

[4] ALSHEIKH, Mohammad Abu, et al. Machine learning in wireless sensor networks: Algorithms, strategies, and applications. IEEE Communications Surveys & Tutorials, 2014, vol. 16, no 4, p. 1996-2018.

[5] DI, Ma; JOO, Er Meng. A survey of machine learning in wireless sensor netoworks from networking and application perspectives. En Information, Communications & Signal Processing, 2007 6th International Conference on. IEEE, 2007. p. 1-5.

[6] FORSTER, Anna. Machine learning techniques applied to wireless ad-hoc networks: Guide and survey. En Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on. IEEE, 2007. p. 365-370.

[7] Nie, J., & Haykin, S. (1999). A Q-learning-based dynamic channel assignment technique for mobile communication systems. IEEE Transactions on Vehicular Technology, 48(5), 1676-1687.

[8] Li, H. (2009, October). Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: A two by two case. In Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on (pp. 1893-1898). IEEE.

[9] Sallent, O., Pérez-Romero, J., Ferrús, R., & Agustí, R. (2015, June). Learning-based coexistence for LTE operation in unlicensed bands. In Communication Workshop (ICCW), 2015 IEEE International Conference on (pp. 2307-2313). IEEE.

[10] Rupasinghe, N., & Güvenç, İ. (2015, March). Reinforcement learning for licensed-assisted access of LTE in the unlicensed spectrum. In Wireless Communications and Networking Conference (WCNC), 2015 IEEE (pp. 1279-1284). IEEE.

References

[11] Bennis, M., & Niyato, D. (2010, December). A Q-learning based approach to interference avoidance in self-organized femtocell networks. In GLOBECOM Workshops (GC Wkshps), 2010 IEEE (pp. 706-710). IEEE.

[12] Bennis, M., Guruacharya, S., & Niyato, D. (2011, December). Distributed learning strategies for interference mitigation in femtocell networks. In Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE (pp. 1-5). IEEE.

[13] Maghsudi, S., & Stańczak, S. (2015). Joint channel selection and power control in infrastructureless wireless networks: A multiplayer multiarmed bandit framework. IEEE Transactions on Vehicular Technology, 64(10), 4565-4578.

[14] Maghsudi, S., & Stańczak, S. (2015). Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting. IEEE Transactions on Wireless Communications, 14(3), 1309-1322.

[15] Wilhelmi, F., Cano, C., Neu, G., Bellalta, B., Jonsson, A., & Barrachina-Muñoz, S. (2017). Collaborative Spatial Reuse in Wireless Networks via Selfish Multi-Armed Bandits. arXiv preprint arXiv:1710.11403.

[16] Gai, Y., Krishnamachari, B., & Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. IEEE/ACM Transactions on Networking (TON), 20(5), 1466-1478.

[17] Combes, R., & Proutiere, A. (2015). Dynamic rate and channel selection in cognitive radio systems. IEEE Journal on Selected Areas in Communications, 33(5), 910-921.

References

[18] Liu, K., & Zhao, Q. (2010). Distributed learning in multi-armed bandit with multiple players. IEEE Transactions on Signal Processing, 58(11), 5667-5681.

[19] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25, 285-294.

[20] Agrawal, S., & Goyal, N. (2013, April). Further optimal regret bounds for thompson sampling. In Artificial Intelligence and Statistics (pp. 99-107).