Transfer Restless Multi-Armed Bandit Policy for Energy Efficient Heterogeneous Cellular Network

Navikkumar Modi, Philippe Mary, Christophe Moy

RESEARCH

Transfer Restless Multi-Armed Bandit Policy for Energy Efficient Heterogeneous Cellular Network

Navikkumar Modi[1†], Philippe Mary[2*] and Christophe Moy[3]

[*]Correspondence:
philippe.mary@insa-rennes.fr
[2]Univ. Rennes, INSA de Rennes, CNRS, IETR - UMR 6164, 20 avenue des Buttes de Coesmes, F-35000, Rennes, France
Full list of author information is available at the end of the article
[†]Work performed during the PhD thesis of NM at CentraleSupelec, France [1]

Abstract

This paper proposes a learning policy to improve the energy efficiency (EE) of heterogeneous cellular networks. The combination of active and inactive base stations (BS) that allows for maximizing EE is identified as a combinatorial learning problem and requires high computational complexity as well as a large signaling overhead. This paper aims at presenting a learning policy that dynamically switches a BS to ON or OFF status in order to follow the traffic load variation during the day. The network traffic load is represented as a Markov decision process, and we propose a modified upper confidence bound algorithm based on restless Markov multi-armed bandit framework for the BS switching operation. Moreover, to cope with initial reward loss and to speed up the convergence of the learning algorithm, the transfer learning concept is adapted to our algorithm in order to benefit from the transferred knowledge observed in historical periods from the same region. Moreover, based on our previous work, a convergence theorem is provided for the proposed policy. Extensive simulations demonstrate that the proposed algorithms follow the traffic load variation during the day and contribute to a performance jump-start in EE improvement under various practical traffic load profile. It also demonstrates that proposed schemes can significantly reduce the total energy consumption of cellular network, e.g. up to 70% potential energy savings based on a real traffic profile.

Keywords: Energy efficiency, green cellular networks, upper confidence bound, reinforcement learning, transfer learning, multi-armed bandit.

1 Introduction

The increasing popularity of portable smart devices has flared up rising traffic demand for radio access network and has been arousing massive energy consumption, which leads to the exhaustion of energy resources and causes a potential increase of CO2 emissions. Data centers, back-haul routers and cellular access networks are the main source of energy consumption in the information and communication technology industry, which is equivalent of 2% to 10% of the global overall power consumption of human activity [2]. In cellular networks, the energy consumption of base stations (BS) is about 60% to 80% of the overall power consumption of the cellular network [3]. Besides, cellular network operators require to spend more than 10 billion dollars to meet current energy consumption of the cellular network [4, 5], thus there exists both environmental and high economical pressures for cellular network operators to take into account an energy efficiency aspect of the network. The main reason for such high energy consumption is because BS, and more generally cellular networks, are designed on a peak traffic load basis.

In fact, due to the traffic load variation in time domain and dynamic distribution of cellular users among cells in space domain, there are opportunities for some BS to be put in sleep mode in order to achieve higher energy efficiency (EE). The side BS components, controller, air-conditioner are the main sources of energy consumption, rather than transmit power which consumes only 3.1% of the BS power consumption [6]. Recent studies on the real temporal traffic have stated that BS are largely underutilized, e.g. traffic load can be below 10% of peak load during 30% and 45% of the day during weekdays and weekends respectively [7]. Thus, instead of just turning off radio transceivers, the BS operators may prefer to turn off the underutilized BS and transfer the imposed traffic loads to neighbor active BS during low traffic periods such as night time and/or weekend, which reduces the energy consumption [4].

Recently, there has been a rising interest on the works dealing with switch ON and OFF BS according to the traffic load, however, it is essential for network operators to guaranty radio coverage and quality of service (QoS) to the cellular users. Dynamically switching the BS' operation mode to ON and OFF with respect to traffic load fluctuation is considered to be one of the effective methods to reduce total energy consumption of cellular network while maintaining good QoS. Moreover, BS operation mode switching decision cannot be made individually at each BS level, since it does not only depend on the load of the cell of interest but also on the load of its neighbors. For instance, a BS may not be turned to sleep mode while its neighboring BS are overloaded, even if its own traffic load is very low. The problem of EE maximization with BS switching operation is a famous combinatorial class of problem. In machine learning, combinatorial problems are mostly addressed with centralized decision made by a central controller taking into account a global information, i.e. channel state information and traffic load information.

In this work, the best BS deployment is learnt in order to maximize the network EE under QoS constraints, by switching ON/OFF some BS. The EE maximization problem is tackled under the multi-armed bandit (MAB) approach where arms are represented by the deployment configurations. MAB is a class of sequential decision making paradigm where, given a set of arms, a user selects an arm at each slot in order to collect some reward. The most important property of MAB paradigm is that a player does not need to know a prior information about each arms' reward distribution, making MAB an interesting solution for EE maximization problem. In this paper, we focus on a *restless upper confidence bound* policy which has been proven to be efficient for opportunistic spectrum access (OSA) problem [8, 9, 10], where selecting an arm leads to two different rewards associated with it. In this paper, the algorithm we proposed in [10], i.e. RQoS-UCB policy, is adapted to the problem of finding the optimal BS configuration that maximizes the observed energy efficiency in the long run. Moreover, the transfer learning (TL) concept [11, 12] is applied in our context, where the temporal dependence in the traffic load between two days is exploited in order to increase the convergence speed of the current learning.

1.1 Related Work

Recently, there has been a substantial body of work on traffic load-aware BS adaptation, and the authors in [13, 14], have validated the possibility of improving EE

and also showed the energy saving gains by simulations. In [15, 16, 17], authors proposed to dynamically adjust the sleeping status of BS, depending on the learnt and predicted traffic load of the network. The works in [18, 19] introduced some BS switching strategies for dynamic BS operations depending on daily traffic variation. However, reliable prediction of BS traffic load is still an important challenge for network operators, which limits its usefulness in practical applications. An alternative energy-efficient procedure is the relay station switching technique employed in [20], where certain BS being turned to sleep mode and switched on the low-powered relay station mode during the low-traffic intervals. On the contrary, authors in [21] introduced reinforcement learning (RL) algorithms as an application of dynamic BS switching operation, however, these algorithms are highly dependent on the a priori knowledge of the traffic load.

As stated in [22, 23, 12], the problem of EE maximization with BS switching operation is a combinatorial problem, and it has been proven to be NP-hard. Instead of directly addressing this problem, the authors in [24, 25], adopted fixed BS switching patterns and then evaluated the call blocking probability and the outage probability. In [4, 26, 17], some greedy algorithms have been introduced to tackle BS switching operation without presenting sufficient theoretical guarantees of convergence to optimal configuration. The authors in [26] have taken forward a greedy algorithm to handle the trade-off between the energy consumption and the revenue in heterogeneous cellular networks. Then, [16, 12] used Markov decision process (MDP) to model the traffic load prediction and used a RL approach, named actor-critic algorithm [27], to predict the traffic load of the network without prior information about it. Moreover, authors in [12] extended the actor-critic algorithm by including the TL concept [11] leading to Transfer Actor CriTic (TACT) algorithm in order to use the knowledge acquired in previous learning phases. Actor-critic based algorithms provide good performance but are generally more complex than upper confidence bound (UCB) algorithms. Moreover, the existing works for BS energy saving problem often lack for theoretical analysis on the convergence. On the contrary, decentralized schemes for dynamic BS switching operation [28, 29, 30, 31] are more beneficial as they do not require a central controller, but demand more information exchanges. However, all the existing decentralized schemes do not present theoretical analysis on the convergence, which makes them less appealing from theoretical point of view.

We assert that problems such as channel allocation in dynamic spectrum access can be of the same nature than the problem of base station switching, i.e. restless MAB. As a consequence, works dealing with learning strategies for OSA scenario for instance can be related to our approach. In [32] authors tackled the problem of MAB with Markovian rewards that can be applied to OSA or base station switching for green networking. In OSA scenario, the most common addressed optimization problem is to find the band with the highest probability to be vacant. In that case, rewards are generally modeled as binary, but some works have also dealt with continuous rewards rating the quality of the bands for instance. The authors of [33, 34] considered the problem of finding the channel that gives the best data rate when data rates on channels are drawn from a Markovian distribution. But in these works, channel quality and availability have never been considered separately. In

our previous works [8, 10], we proposed a new restless upper confidence bound for Markovian settings in OSA problem. The proposed scheme, named restless quality of service upper confidence bound (RQoS-UCB), allows the radio for learning about the spectrum opportunities, i.e. bands that are less used by a primary network for instance, and also on a quality indicator of the bands that have been identified as unoccupied. The proposed scheme has been proven to have a logarithmic regret which is the best behavior a learning policy may have, and its ability to converge toward the best band, in terms of availability and quality as well, has been shown.

1.2 Contributions

This paper tackles the problem of EE maximization from the restless MAB framework, considering varying traffic load. The paper includes the following contributions:

- This paper adapts the UCB policy in [10] to fit with the EE maximization problem. In particular, the state reward is fed-back when the BS configuration fulfils a set of constraints of the EE maximization problem. The soft reward is matched with the energy efficiency of the network.
- The proposed algorithm includes a transfer learning stage in order to speed up the convergence toward the best deployment configuration.

To the best or our knowledge, MAB-UCB has never been applied to the dynamic configuration by switching ON and OFF BS in order to maximize the network EE. Our algorithm that learns on the state of the network and on the energy efficiency, is proven to be efficient to solve the green networking problem.

1.3 Paper Structure

The remainder of the paper is organized as follows. Section 2 introduces the system description and EE maximization problem formulation. In Section 3, the traffic load variation is formulated as an MDP and the EE maximization algorithm, energy efficiency maximization-upper confidence bound (EEM-UCB), is presented. Moreover, the TL concept is embedded in the proposed EEM-UCB algorithm to form the TLEEM-UCB algorithm. Section 4 numerically evaluates and compares the proposed schemes with the state of the art methods and presents the validity and effectiveness. Finally, Section 5 concludes this paper and presents future way of researches.

2 Methods and Problem Formulation

Beforehand, Table 1 summarizes the notations in this paper.

2.1 Network Model

In this work, we consider a heterogeneous wireless cellular network comprising of a mixture of macro and small cells, each governed by a macro or micro BS respectively, where set of BS $\mathcal{Y} = \{1, 2, \cdots, Y\}$ lies in a two dimensional area in $\mathbb{R}^2$. In addition, we assume that there exists a central controller, which can timely know the traffic load in the network at each instant and can predict the energy efficiency of BS at next stage. Let us assume that all BS operate in an open access mode, i.e. any MS is allowed to connect to any BS whatever it belongs to the micro or macro

Table 1 List of the main symbols in the paper.

Symbol	Meaning
$\mathcal{Y}$	Set of BS $\mathcal{Y} = \{1, \cdots, Y\}$
$\mathcal{Y}_n^{on}$	Set of active (ON) BS at the $n-$th iteration
$\mathcal{I}_k(n)$	Cell coverage of BS k at time n
x_k	Denote the locations of the MS in coverage $\mathcal{I}_k(n)$ of the $k-$th BS at time n
$\Lambda(x_k, n)$	Traffic arrival rate at location x_k in BS k following a Poisson point process at the $n-$th iteration
$1/h(x_k, n)$	Average call duration (or file size) at $n-$th iteration at x_k
$L_k(n)$	Instantaneous traffic load served by the BS $k \in \mathcal{Y}_n^{on}$
$\Theta_k(x_k, n)$	Service rate at location x_k from BS k at the $n-$th iteration
$\text{SINR}_k(x_k, n)$	Received SINR at active MS location x_k from BS k at the $n-$th iteration
$\rho_k(n)$	System load of BS k at the $n-$th iteration
ρ_{th}	System load threshold
P_k^{tx}, P_f^k and P_T^k	Transmit, fixed and total operational power of BS k
$EE(n)$	Network energy efficiency (EE) in bits per joule
$\Theta^{\min}$	Prescribed minimum data rate to continue data transmission
$\mathcal{M} =< \mathcal{S}, \mathcal{K}, \mathcal{P}, R >$	MDP Tuple: state space, action space, transition probability and reward
$\mathcal{P}^i = \{P_{k,l}^i, k, l \in \mathcal{S}\}$	state transition probabilities of the $i-$th action
$\mathcal{A}(n)$	$\mathbf{a}^i(n) = \left[a_1^i(n), \cdots, a_Y^i(n)\right]$ the controller decides an action for all BS, i.e. ON or OFF
$T^i(n)$	total number of times action i has been selected up to iteration n
$b(n)$	total number of completed blocks up to iteration n
n_2	total number of iterations in SB2 block up to block $b(n)$
T_2^i	total number of times action i has been selected during SB2 block up to n_2 iteration
h_2	total number of iterations in historic period
H_2^i	total number of times action i has been selected in SB2 block in historic period
$S^i(n)$ and $S^{i,h}(n)$	state observed due to action i at the $n-$th iteration, in the current and source task respectively
$R_S^i(n)$ and $R_S^{i,h}(n)$	immediate EE reward with action i in state S at $n-$th iteration in the current and source task respectively
$G_S^i(T^i(n))$	$\frac{1}{T^i(n)} \sum_{k=1}^{T^i(n)} R_S^i(k)$, the empirical mean of EE rewards
$G_{\max}^S(n)$	$\max_{i \in \mathcal{K}} G_S^i(T^i(n))$, maximum expected average EE
$M^i(n, T^i(n))$	$G_{\max}^S(n) - G_S^i(T^i(n))$
$B^i(n, T^i(n))$	EEM-UCB policy index giving the BS configuration status to activate
α and β	exploration coefficients with respect to state and reward, respectively
ζ^i	state that determines regenerative cycles for action i
π_S^i	stationary distribution for state S of the Markov chain associated with action i
μ^i	$\sum_{S \in \mathcal{S}} S^i G_S^i \pi_S^i$, global mean reward, i.e. taking into account the reward as well as the state of each action i
$\Delta\mu^i$	$\mu^* - \mu^i$
$\hat{\pi}_S^i$, $\hat{\pi}_{\max}$, $\pi_{\min}^i$, $\pi_{\min}$	$\max\left\{\pi_S^i, 1 - \pi_S^i\right\}$, $\max_{S \in \mathcal{S}, i \in \mathcal{K}} \hat{\pi}_S^i$, $\min_{S \in \mathcal{S}} \pi_S^i$, $\min_{i \in \mathcal{K}} \pi_{\min}^i$
$r_{\max}$	$\max_{S \in \mathcal{S}, i \in \mathcal{K}} r_S^i$
$S_{\max}$	$\max_{i \in \mathcal{K}} \lvert S^i \rvert$, where $\lvert S^i \rvert$ stands for the cardinality of the state space of action i
$M_{\min(\max)}$	$\min(\max \text{ resp.})_{i \in \mathcal{K}} M^i\left(n_2, T_2^i(n_2)\right)$
ε^i, $\varepsilon_{\min}$	$1 - \lambda_2^i$, being the eigenvalue gap of the $i-$th action, $\min_{i \in \mathcal{K}} \epsilon^i$
$\Omega_{k,l}^i$	mean hitting time of state l starting from an initial state k for the ith action
$\Omega_{\max}^i$, $\Omega_{\max}$	$\max_{k,l \in \mathcal{S}, k \neq l} \Omega_{k,l}^i$, $\max_{i \in \mathcal{K}} \Omega_{\max}^i$

tier [25]. We focus on the downlink communication as mostly considered for the mobile Internet application. The network area is divided according to the Voronoi tessellation with BS acting as seeds for each cell. Each cell coverage in wireless cellular network is denoted as $\mathcal{I}_k(n), k = 0, 1, 2, \cdots$ at time slot n. At a given time slot, the set of active BS, denoted as $\mathcal{Y}^{\text{on}}$ defines a partition of the space. Each MS in the network connects to its nearest BS, as explained in Section 2.1.2. As the set of active BS is changing from a time instant to another, MSs connect always to the nearest BS, micro or macro. Each configuration of $\mathcal{Y}^{\text{on}}$ leads to a certain rate and energy consumption, whose computation is detailed in the following, and we aim at finding the configuration maximizing the energy efficiency, while guaranteeing a minimum data rate to all users.

2.1.1 Traffic Profile

Let $x_k \in \mathcal{I}_k(n)$ be the two-dimensional Cartesian coordinates, denoting the locations of MS in the coverage of the $k-$th BS at time slot n. An MS is referred as active when it is receiving a call. When the call ends, the MS becomes inactive and is departed from the network. Traffic load of a BS is measured in terms of the number of active MSs and their respective call duration.

At each time slot n, new and handover call at x_k follows a Poisson point process with arrival rate per time-unit $\Lambda(x_k, n)$. The associated call duration (or file size) is assumed to be exponentially distributed with mean $1/h(x_k, n)$. Then the instantaneous traffic load at location x_k can be expressed as $L(x_k, n) = \frac{\Lambda(x_k,n)}{h(x_k,n)}$ at time slot n [29]. By setting different arrival rates or call holding time for MSs located in different cells, this model can capture temporal and spatial traffic variability. Thus, when the set of BS $\mathcal{Y}_n^{on}$ is switched ON at time slot n, the instantaneous traffic load served by BS $k \in \mathcal{Y}_n^{on}$ can be expressed as:

$$L_k(n) = \sum_{x_k \in \mathcal{I}_k(n)} \frac{\Lambda(x_k, n)}{h(x_k, n)}.$$

On the contrary when BS k is turned OFF, the instantaneous traffic load served by BS k is defined as zero, i.e. $L_k(n) = 0$. The total arrival rate of a BS k is the composition of all Poisson arrivals at different locations in $\mathcal{I}_k$, which again forms a Poisson process [35]. Moreover, the daily traffic profile of the whole cellular network repeats periodically as recorded by several works [7, 20]. This model will be useful when considering the performance of the learning algorithms during the day.

2.1.2 BS Selection Rule

An MS is assumed to connect with the nearest BS, in order to suffer from the least path loss during the wireless transmission. An active MS located at x_k is connected with and served by the BS $k, k \in \mathcal{Y}_n^{on}$ which presents the best received signal strength at each time slot n[1], and where $\mathcal{Y}_n^{on}$ is the set of active BS at instant n.

2.1.3 Channel Model

The service rate of an active MS at location x_k provided by the k-th BS at the n-th time slot is assumed to be equal to the Shannon capacity:

$$\Theta_k(x_k, n) = B_a \cdot \log_2 \left(1 + \mathrm{SINR}_k(x_k, n)\right) \tag{1}$$

where B_a denotes the system bandwidth, $\mathrm{SINR}_k(x_k, n)$ is the received signal to interference plus noise ratio (SINR) at x_k from BS k at the $n-$th time slot, and is defined as

$$\mathrm{SINR}_k(x_k, n) = \frac{g_{k,x_k}(n) P_k^{tx}}{\phi g_{k,x_k}(n) P_k^{tx} + \sum\limits_{m \in \mathcal{Y}_n^{on} \setminus \{k\}} g_{m,x_k}(n) P_m^{tx} + \sigma^2} \tag{2}$$

where $g_{k,x_k}(n)$ is the average channel gain from BS k to active MS at location x_k at the $n-$th time slot. The channel gain only comprises path loss in this paper, but log-normal shadowing and fading can be taken into account easily without changing the principle of the learning policy that will be introduced in the next section. Moreover, P_k^{tx} is the transmission power of BS k, σ^2 is the noise power

[1]Denote that an other user association metric could also be used. The optimal user association problems has been well addressed in [36, 37, 38], however, we focus on the BS sleeping scheme rather than user association due to the space limitation.

and $\sum_{m \in \mathcal{Y}_n^{on} \setminus \{k\}} g_{m,x_k}(n) P_m^{tx}$ is the interference power experienced by MS x_k from its neighboring BS at the $n-$th time slot. The parameter ϕ is the orthogonality (or self interference) factor, $\phi \in [0,1]$, and $\phi g_{k,x_k}(n) P_k^{tx}$ models intra-cell interference [38].

2.1.4 System Load

In order to satisfy the QoS requirement of MSs, a BS should provide a certain amount of resources (e.g., time or frequency) in order to absorb the MSs traffic load and provide enough service rate to users. From the system's perspective, the system load of BS k at the $n-$th time slot is estimated as the fraction of resource to serve the total traffic load in its coverage [29]

$$\rho_k(n) = \sum_{x_k \in \mathcal{I}_k(n)} \frac{L(x_k, n)}{\Theta_k(x_k, n)}. \tag{3}$$

The system load denotes the fraction of time required to serve the total traffic load in the coverage of the k-th BS. Eventually, our main goal is to choose the set of active BS that maximizes the global network energy efficiency without having a prior on traffic load statistic. We will give the details in Section 3.

2.1.5 Power Consumption Model

The total power consumed $P_T^k(n)$ by each BS k at the $n-$th time slot can be expressed as [39]:

$$P_T^k(n) = a_k P_k^{tx}(n) + P_f^k \tag{4}$$

where, $P_k^{tx}(n)$ denotes the transmission power of BS k at the $n-$th time slot and P_f^k denotes the static power consumption independent of $P_k^{tx}(n)$ and includes all electronic circuit power dissipation due to site cooling, signal processing hardware as well as battery backup systems. a_k is a BS power scaling factor which reflects both amplifier and feeder losses.

2.2 Problem Formulation

The energy efficiency of a cell k in bits per joule at instant n, is the ratio between the data sum-rate of the cell over the power used to run the cell. The network EE is then the aggregate EE of each cell and can be expressed as [40]:

$$EE(n) = \sum_{k \in \mathcal{Y}_n^{on}} \frac{\sum_{x_k \in \mathcal{I}_k(n)} \Theta_k(x_k, n)}{P_T^k(n)} \tag{5}$$

EE maximization in the cellular network, without power allocation strategy, can be reduced to find the set of active BS that maximizes (5). The problem can be

formally written as

$$\mathcal{Y}_n^{on^*} = \arg\max_{\mathcal{Y}_n^{on}} \left[\sum_{k \in \mathcal{Y}_n^{on}} \frac{\sum\limits_{x_k \in \mathcal{I}_k(n)} \Theta_k(x_k, n)}{P_T^k(n)} \right] \tag{6a}$$

$$s.t. \quad 0 \leq \rho_k(n) \leq \rho_{th}, \forall k \in \mathcal{Y}_n^{on} \tag{6b}$$

$$\Theta_k(x_k, n) \geq \Theta^{\min}, \forall x_k \in \mathcal{I}_k(n), \forall k \in \mathcal{Y}_n^{on} \tag{6c}$$

$$\mathcal{Y}_n^{on} \neq \emptyset \tag{6d}$$

Like in [29, 12], a system load threshold $\rho_{th} \leq 1$ is introduced as a constraint, (6b), in order to keep the system stable. Indeed, the service rate of a user, i.e. $\Theta_k(x_k, n)$, should be sufficient to absorb the traffic load at x_k. If not, some transmissions may be delayed and should be taken into account in the model, which is out of scope of the paper. For instance, low threshold value ρ_{th} indicates that BS would operate in a more conservative manner with low delay and low call dropping probability for MSs since all calls can be routed to users. On the contrary, with a high threshold ρ_{th} value close to 1, the data rate of users is just enough to avoid overflow implying a limited power consumption but with an increasing call dropping probability. The constraint (6c) guarantees a minimum data rate $\Theta^{\min}$ to each active user and constraint (6d) states that there is at least one active BS.

The above problem can be proven to be NP-complete by reducing from a vertex cover problem [22, 29]. Finding the set of active BS maximizing network EE by an exhaustive search is very costly in computational resources since $2^Y - 1$ ON/OFF combinations have to be tried, specifically when the number of BS is large. This problem can rather be tackled under MAB approach where a specific combination is tried at each iteration and a reward (EE of the system) is collected. In the next section, we will show how this principle can lead to a good state.

3 RL for Energy Efficient Network

3.1 System Model

The dynamic BS switching problem is modeled as an MAB under Markovian settings. Figure 1 illustrates the principle of the learning policy with MAB approach where an arm represents different configuration of BS' activity. We defined an MDP for BS switching operation as a tuple $\mathcal{M} =< \mathcal{S}, \mathcal{K}, \mathcal{P}, R >$, where $\mathcal{S}$ denotes the state space, $\mathcal{K}$ denotes the action space, $\mathcal{P}$ denotes a state transition probability matrix, and finally R is a reward function associated with $\mathcal{S}$, $\mathcal{K}$ and $\mathcal{P}$. At each iteration, the controller chooses an action i among $|\mathcal{K}| = 2^Y - 1$ possible actions, i.e. $\mathbf{a}^i(n) = \left[a_1^i(n), \cdots, a_Y^i(n)\right]$ where $a_k^i(n) = 1$ if BS k is switched ON in action number i at time n and 0 otherwise. This action leads the network to a given state $S^i(n) \in \{0, 1\}$, where $S^i(n) = 1$ if all constraints from (6b) to (6d) are satisfied and 0 otherwise.

Due to the random process governing the time evolution of the traffic load, the state $S^i(n)$ transforms into $S^i(n+1)$ at the next time instant according to a transition probability measure for arm i, i.e. $\mathcal{P}^i = \{P_{k,l}^i, k, l \in \mathcal{S}, i \in \mathcal{K}\}$. Moreover, the

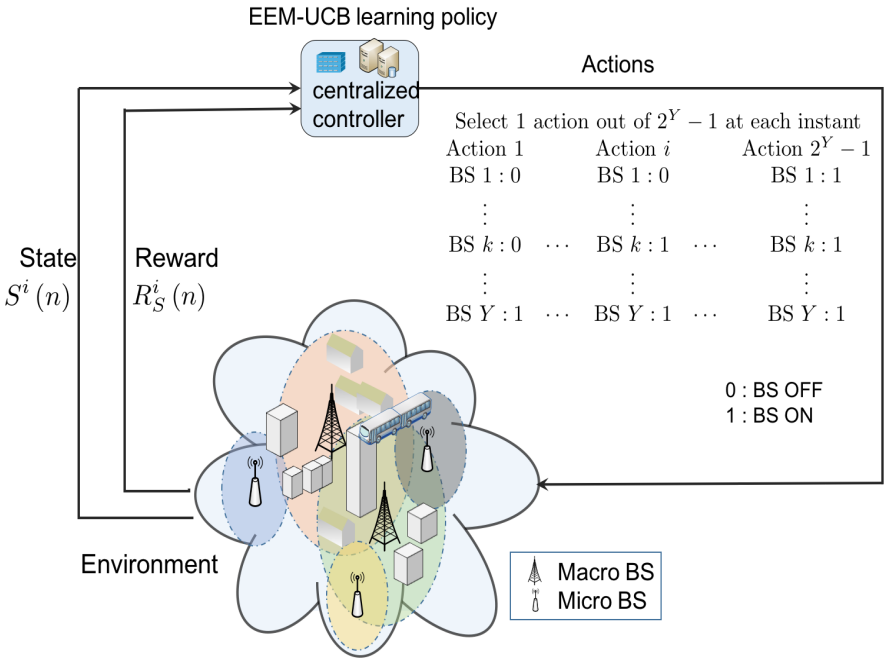


Figure 1 Reinforcement learning (RL) framework for BS switching operation.

Markovian process is considered as stationary, on a short time period e.g. 1 hour, and hence the distribution of this MDP is such as $\pi_S^i(n) = \pi_S^i, \forall n$. The reward achieved in state S^i from the BS switching operation i after n time slot is, without loss of generality, equal to the value of the state, $S^i(n)$, i.e. 0 or 1. In addition, we consider that the network EE achieved by switching BS status according to the action number i is the second reward, i.e. $R_S^i(n) = EE(n)$, computed from (5) for a given environment state $S^i(n)$. The reward on EE and the state are fed back to the controller in order to decide the next action to take. The mean reward μ^i associated with BS switching operation i under stationary distribution $\boldsymbol{\pi}_S^i$ is given by: $\mu^i = \sum_{S \in \mathcal{S}} S^i G_S^i \pi_S^i$, where

$$G_S^i(T^i(n)) = \frac{1}{T^i(n)} \sum_{p=1}^{T^i(n)} R_S^i(p) \tag{7}$$

where $T^i(n)$ refers to the number of times the BS switching operation i has been performed by the controller up to time n. The policy $\mathcal{A}$ is a one-to-one mapping such as at each time slot n, a BS switching operation i is selected:

$$\begin{aligned} \mathcal{A} : \mathbb{N} &\longrightarrow \mathcal{K} \\ n &\longmapsto i \end{aligned}$$

The goal of a RL policy is to minimize its regret on the long run, i.e.

$$\Phi^{\mathcal{A}}(n) = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} S^{\mathcal{A}(t)}(t) G^{\mathcal{A}(t)}_{S^{\mathcal{A}(t)}}(t)\right] \tag{8}$$

where the expectation $\mathbb{E}$ is taken over the states and observed reward. Let $S^{\mathcal{A}(t)}$ be the state observed by using the policy $\mathcal{A}$ at time slot t. Moreover, μ^* is the optimal mean reward obtained by always selecting the best action at each time t.

3.2 Restless Energy Efficiency Maximization Upper Confidence Bound (EEM-UCB)

In this section, we adopt the restless UCB policy, i.e. RQoS-UCB, that we proposed in [10] for learning the best bands in OSA context based on their probability to be free and the quality of the band. The principle of the algorithm is adapted to the current problem, where the policy aims at finding an optimal set of active BS which maximizes the energy efficiency of the network, and will be named EEM-UCB.

When dealing with an MAB problem, one should first ask if it belongs to rested or restless category. In the former category, the state of the Markov chains corresponding to arms that are not played does not evolve with time and only the Markov chain of the selected arm does. In the later, states of all Markov chains continue to evolve whatever they are selected or not. Our problem fits with the later category. Indeed, the traffic request of users does not depend on selected BS however, the selected configuration definitely influences the traffic load of the network by distributing the data flow among BS. EEM-UCB algorithm operates in a block structure as represented in Fig. 2 which is based on regenerative cycle [41, 42]. Each block is divided into three sub-blocks, SB1, SB2 and SB3. For each arm i, a *regenerative* state ζ^i is defined, i.e. 0 or 1 in our case, and SB1 comprises all time slots from the selection of configuration i to the first visit of the state ζ^i. SB2 contains all time slots from the first visit to ζ^i up to, but excluded, the second visit to ζ^i. The last block is only the second visit to the state ζ^i. The selection of the active BS set is based on an index computation $B^i(n)$ for configuration i and will be formally expressed in the next section. The computation of the index occurs after the completion of SB3. The reason of this structure lies on the restless nature of arms which evolve even if they are not played. The distribution of the state reward obtained by playing a given arm, is function of the time elapsed since the last time the same arm has been played. In order to deal with an homogenous Markov chain, the stay in a given arm should be sufficiently long in order to reconstruct a sample path with the same statistical characteristic of the Markov chain governing the arm [42]. It is worth noting, however, that this structure does not prevent to collect rewards, state and EE, in any blocks. This sub-division just comes for mathematical convergence proof.

At a given time slot n, policy $\mathcal{A}$ selects the BS switching operation that has the highest policy index $B^i(n, T^i(n))$ at time n. This action may transform the current state $S^i(n)$ of the network to another state $S^i(n+1)$ with certain probability $\mathcal{P}$. The new reward $R^i_S(n)$, i.e. energy efficiency, is fed back to the controller. Then, the policy $\mathcal{A}$ updates the policy index $B^i(n+1, T^i(n+1))$ with the empirical average on the state S^i and the empirical mean of the energy efficiency experienced so

far. The algorithm repeats the above procedure until convergence to optimal BS switching operation during each hour of operation. The formal description of the index computation is given in Section 3.3.

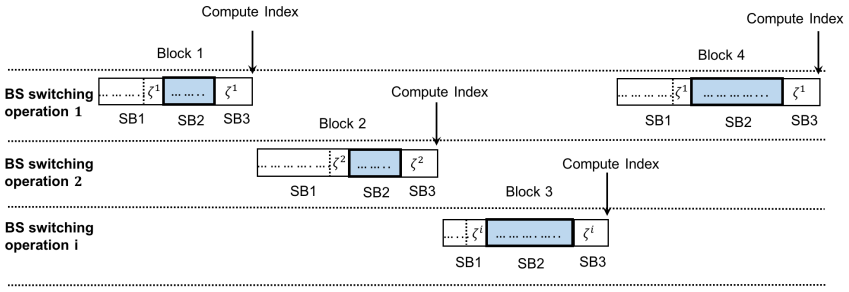


Figure 2 Regenerative sample path for EEM-UCB

3.3 Transfer Learning EEM-UCB (TLEEM-UCB) Policy

The previous strategy may suffer from traffic load variation from one day to another at a given period of time due to the variation of Poisson arrival rate between two consecutive days. This rather advocates for learning from scratch at each new day with the new, unknown, statistic characterizing the underlying Markovian process. In that case, the network would loose time to re-learn the best deployment configuration it learnt the day before at the same hour. Another strategy would consist in using the previous knowledge the controller learnt during some historical periods to find the current optimal BS switching operation. This strategy would make even more sense as per Poisson arrival rate does not change too much between two consecutive days as we will see in Section 4.

The motivation for transfer learning is to utilize previous learnt features on a given task (source task) in order to speed up the learning phase of different features on a target task as illustrated in Fig. 3. In other words, the controller uses the BS deployment learnt in previous time period for the current task with its own statistical characteristics. We hence propose a new policy update method, named Transfer Learning EEM-UCB (TLEEM-UCB) policy that is detailed in Algorithm 1. In the source policy, the reward achieved in state $S^{i,h}$ from a BS switching operation $i \in \mathcal{K}$ during H_2^i time slots is $S^{i,h}(H_2^i)$, where H_2^i is the number of time slots the BS switching operation i has been selected in SB2 block in the source task as reminded in Table 1. Meanwhile, the observed reward associated with energy efficiency by selecting a BS switching operation i is $R_S^{i,h}(H_2^i)$ during H_2^i source task observations in SB2.

At the end of each block b, Algorithm 1 returns a BS switching operation index maximizing the policy index, $B^{i,h}(n_2, T^i)\ \forall\ i \in \mathcal{K}$, which has to be selected for the next block of operation, i.e. steps 2 and 3. The index computation is done according to three terms:

$$B^{i,h}(n_2, T_2^i) = \bar{S}^{i,h}(T_2^i) - Q^{i,h}(n_2, T_2^i) + A^{i,h}(n_2, T_2^i), \quad \forall i \tag{9}$$

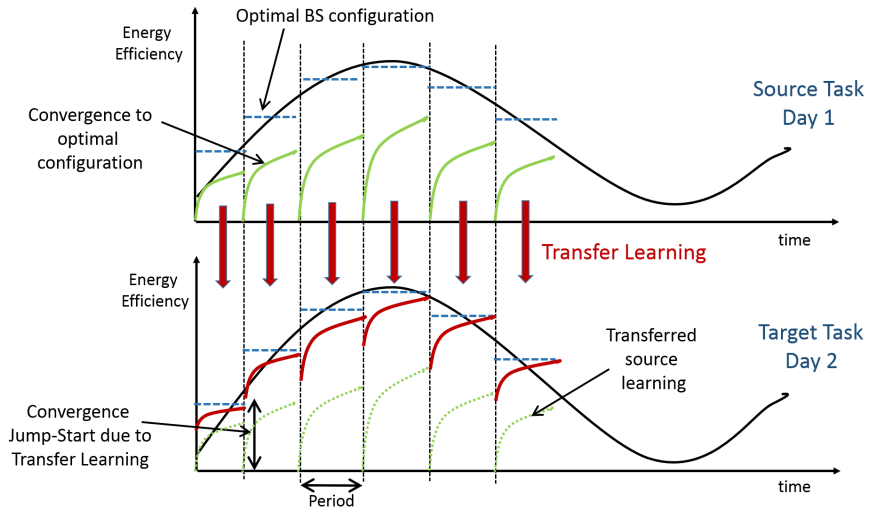


Figure 3 Transfer learning for Transfer EEM-UCB (TLEEM-UCB) policy.

where $\bar{S}^{i,h}(T_2^i)$ is the empirical mean of the observed states obtained with action i considering the time period in the source task and in the current task. As reminded in Table 1, T_2^i is the number of times action i has been selected in SB2 block up to time n_2 in the current task. The empirical mean is expressed as

$$\bar{S}^{i,h}(T_2^i) = \frac{\sum_{t=1}^{T_2^i} S^i(t) + \sum_{t=1}^{H_2^i} S^{i,h}(t)}{T_2^i + H_2^i}, \forall i. \tag{10}$$

The second term, i.e. $Q^{i,h}(n_2, T_2^i)$, is computed similarly than in RQoS-UCB policy [10] but including source task observations:

$$Q^{i,h}(n_2, T_2^i) = \frac{\beta M^{i,h}(n_2, T_2^i) \ln(n_2 + H_2^i)}{T_2^i + H_2^i}, \forall i, \tag{11}$$

where,

$$M^{i,h}(n_2, T_2^i) = G^S_{\max} - G_S^{i,h}(T_2^i), \quad \forall i,$$

and $G_S^{i,h}(T_2^i) = \frac{1}{T_2^i}\sum_{k=1}^{T_2^i} R_S^i(k) + \frac{1}{H_2^i}\sum_{k=1}^{H_2^i} R_S^{i,h}(k)$ denotes the empirical mean of EE reward, i.e. R_S^i, collected in the current task in SB2 block by applying action i in state S plus the total mean EE reward gathered in source task. Moreover, $G^S_{\max} = \max_{i\in\mathcal{K}} G_S^{i,h}(T_2^i)$ is the maximum reward within the set of BS switching operations from current and historical observations in state S. Finally, the bias term $A^{i,h}(n_2, T_2^i)$, is defined as

$$A^{i,h}(n_2, T_2^i) = \sqrt{\frac{\alpha \ln(n_2 + H_2^i)}{T_2^i + H_2^i}}, \quad \forall i. \tag{12}$$

Algorithm 1 TLEEM-UCB policy

Require: **Transferred Observations :** h : total historic time, h_2 : total historic time in SB2 block, H_2^i : total historic time action i has been selected, b^h : total blocks in historical observations $R_S^{i,h}(t)$, $S^{i,h}(t)$ $\forall i \in \mathcal{K}, 1 \leq t \leq H_2^i$: Reward and state observed in historic data,
Current policy initialization: $b = 0$, $n = 0$, $n_2 = 0$, $T_2^i = 0$, α, β, ζ^i $R_S^i(0)$ and $S^i(0)$.
Ensure: $\mathcal{A}(n+1)$
while (1) **do**
  $B^{i,h}(n_2, T_2^i) = \bar{S}^{i,h}(T_2^i) - Q^{i,h}(n_2, T_2^i) + A^{i,h}(n_2, T_2^i)$, $\forall i$
  $\mathcal{A}(n) = \arg\max_i B^{i,h}(n_2, T^i)$
  while $S^i(n_2) \neq \zeta^i$ **do**
    $n = n + 1$ and $\mathcal{A}(n) = i$ // Start SB1 sub-block
    Activate configuration i and Observe $S^i(n_2)$
  end while
  $n = n + 1$, $n_2 = n_2 + 1$, $T_2^i = T_2^i + 1$ and $\mathcal{A}(n) = i$; // End of SB1, start SB2
  Observe current state $S^i(n_2)$ and update $R_S^i(n_2)$
  Update $\bar{S}^{i,h}(T_2^i)$, $Q^{i,h}(n_2, T_2^i)$ and $A^{i,h}(n_2, T_2^i)$ as of (10), (11) and (12), respectively
  while $S^i(n_2) \neq \zeta^i$ **do**
    $n = n + 1$, $n_2 = n_2 + 1$, $T_2^i = T_2^i + 1$ and $\mathcal{A}(n) = i$; // Start SB2 sub-block
    Observe current state $S^i(n_2)$ and update $R_S^i(n_2)$
    Update $\bar{S}^{i,h}(T_2^i)$, $Q^{i,h}(n_2, T_2^i)$ and $A^{i,h}(n_2, T_2^i)$ as of (10), (11) and (12), respectively
  end while
  $b = b + 1$, $n = n + 1$ and $\mathcal{A}(n) = i$ // Start of SB3 sub-block
end while

It is worth noting that it exists a class of bandit algorithms that uses side information. This kind of bandits is sometimes called contextual bandit or bandit with feedback [43]. Expert systems described in [44] can also be seen as a generalization of learning with side observations. The main idea of bandits with side information is that at each time instant, and before taking a decision, the player is able to observe a realization of a random variable, or a linear function of it, that is called side information, in order to produce a next estimate closer to the real value that is searched. On the other hand, transfer learning aims at using the index B^i of each arm i, computed previously, to initialize the algorithm in order to achieve a jump-start in the convergence rate, that makes transfer learning a quite different approach than bandits with side information.

3.4 Convergence Analysis of TLEEM-UCB

In this section, the total number of suboptimal plays is upper-bounded and established under the following condition 1 on the arms.

Condition 1 *All arms are finite-state, irreducible, aperiodic Markov chains whose transition probability matrices have irreducible multiplicative symmetrization, and the state of non-played arms may evolve.*

Let us consider $G_q^i \geq \frac{1}{\hat{\pi}_{\max} + \pi_q^i}$ and $\beta \geq 84 S_{\max}^2 r_{\max}^2 G_{\max}^2 \hat{\pi}_{\max}^2 / \left(\epsilon_{\min} \Delta\mu_i^R M_{\min}\right)$. We present an upper bound on the total expected number of plays of suboptimal arms in Theorem 1.

Theorem 1 *Assume all arms follow condition 1. Let* $\pi_{\min}$, $\hat{\pi}_{\max}$, $S_{\max}$, $r_{\max}$, $\varepsilon_{\min}$, $M_{\min}$, $\Delta\mu^i$ *and* $\Omega_{\max}^i$ *defined as in Table 1. The total expected number of plays of suboptimal BS configuration is upper-bounded by:*

$$\mathbb{E}[T^{i,h}(n)] \leq \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + 1\right)\left(l^+ + \frac{4}{\pi_{\min}} \sum_{t=1}^{\infty} (t + H_2^*)^{-2}\right)$$

where,

$$l^{+} = \max\left(0, \frac{4\alpha \ln\left(n_2 + H_2^i\right)}{(\Delta\mu^i)^2} - H_2^i\right) \tag{13}$$

Proof A sketch of proof of Theorem 1 is provided in Appendix A and follows the same steps as in [10, Th. 1] considering transferred observations. □

Note that the above bound reduces to the bound of EEM-UCB policy, which is the bound of RQoS-UCB policy [10, Th. 1], when the transferred knowledge is not available (i.e. $H_2^i = 0, \forall i$).

3.5 Complexity and scalability issues

TLEEM-UCB is an index based algorithm. The complexity of the index computation is small. At each iteration in SB2, it requires the evaluation of (10), (11) and (12). (10) is nothing but a moving average that only requires one addition and one division at each iteration since other values have been recorded during previous iterations. (11) requires a log operation, two multiplications (but one with a constant which is less complex than a multiplication between two varying terms), one division and a moving average for evaluating M at each iteration. Finally, the bias term (12) requires a multiplication (with a constant) of the log term already computed once, a division and a square root evaluation at each iteration. The low amount of computation and the long period between two iterations makes it negligible compared to the simplest signal processing operation to be done at PHY layer for instance.

The algorithm complexity of TLEEM-UCB is linear with the number of combinations, but the later is exponential with the number of base stations, i.e. $2^Y - 1$. But this complexity is entirely concentrated in the first initialization phase where the algorithm explores all combinations once in order to give an index to each configuration. Once this has been done, only one BS configuration is tested for the index computation at a given time, hence with the computational complexity mentioned above. Moreover and thanks to transfer learning, initialization phase does not need to be repeated each day, since algorithm uses the best indexes previously learnt in historical periods, to start the new learning phase. A large network will impact the convergence time of the algorithm, since the best configuration needs to be found in a larger set cardinality, but it does not increase the computational complexity. The convergence time would be far too long for a network with 50 base stations for instance. However, one can imagine to have a learning algorithm to control a cluster of few base stations and not the entire network. Coordination among clusters could be done in a higher level in the network, but this is beyond the scope of the paper. Finally, it is worth noting that actor-critic algorithm in [12] and decentralized greedy in [28] belong to the larger class of Q-learning algorithms whose algorithmic complexity is significantly larger than the computation of an index in an UCB policy.

The algorithm relies on the feedback of energy consumption metric of each cell at the central entity. However, base stations already record the data rate and the

transmit power allocated to each user. By monitoring also its own power consumption, an estimation of energy efficiency can be computed. Computed EE only needs to be transmitted over a certain number of bytes, to the central controller leading to a negligible amount of overhead added on fronthaul links. However, the time needed to put decision into action is not equal to zero. The exact evaluation of the time needed to collect measurements, providing a reward, switch on/off a set of BS, would take a certain amount of time that depends on the data used to build a statistic, e.g. average consumed power, and other technological constraints. These features are, of course, of great importance in a real deployment experiment but require much more investigations including implementation in a real platform and are left for further works. We will see in the next section that the proposed algorithms converge around 3000 iterations. If the time lag between the collection of data and the configuration change is 1 second, convergence occurs after 1 hour. However, the algorithm continuously performs the index computation according to the received frames in the network and never stops running such that the base station configuration is continuously changed during the day according to the traffic measured in the network.

4 Results and Discussion

In this section, the performance of our proposed energy efficient dynamic BS operation algorithm is investigated through extensive simulations under practical configurations similar to [4, 12, 29]. We consider an heterogeneous cellular network topology consisting of 5 macro and 5 micro BS arbitrarily deployed in an area of 5×5 km^2. Furthermore, the call arrival rate at location x^k follows a Poisson point process with intensity $\Lambda(x^k, n)$ which may vary between source and target task as summarized in Table 2 and the average file size of each call is $1/h(x^k, n) = 100$ Kbyte.

Table 2 Simulation parameters

Parameter description	Value
Simulation area	5km $\times$ 5km
Maximum transmission power	Macro BS: 20W, Micro BS: 1W
Maximum operational power	Macro BS: 865W, Micro BS: 38W
BS Height	Macro BS: 32m, Micro BS: 12.5m
Intra-cell interference factor	0.01
Channel bandwidth	1.25MHz
Path loss model	COST 231
Arrival rate $\Lambda(x^k)$ in source task	0.05×10^{-4}
Arrival rate $\Lambda(x^k)$ in target task	0.05×10^{-4} to 2×10^{-4}
MSs call holding time $1/[h(x^k)]$	100Kbyte
System load threshold ρ_{th}	0.6
Minimum bit rate requirement $\Theta^{\min}$	122kbps
Exploration parameters of RQoS-UCB	$\alpha = 0.25$ and $\beta = 0.32$

Maximal macro and micro BS transmission powers are set to 20 and 1 W respectively, while the maximum operational power consumption for macro and micro BS are 865 W and 38 W, respectively. The COST 231 modified path loss model is used for radio propagation environment, with macro and micro BS heights are set to 32 m and 12.5 m, respectively similar than in [4, 29, 12]. In order to guarantee system reliability, system load threshold $\rho_{th} = 0.6$ is considered for all BS [29] and the minimum bit rate $\Theta^{\min}$ is set to 122 kbps [40] for each active user. The intra-cell

interference factor ϕ is set to 0.01 and the exploration parameters for EEM-UCB and TLEEM-UCB policies are $\alpha = 0.25$ and $\beta = 0.32$. As per [4], a homogeneous user distribution with intensity $\Lambda = 10^{-4}$ corresponds to 10% of BS utilizations in a case where all BS are switched ON, this value is taken as reference in the analysis on the influence of traffic load variation on the performance of proposed policy. Table 2 summarizes all the parameters used for the simulations.

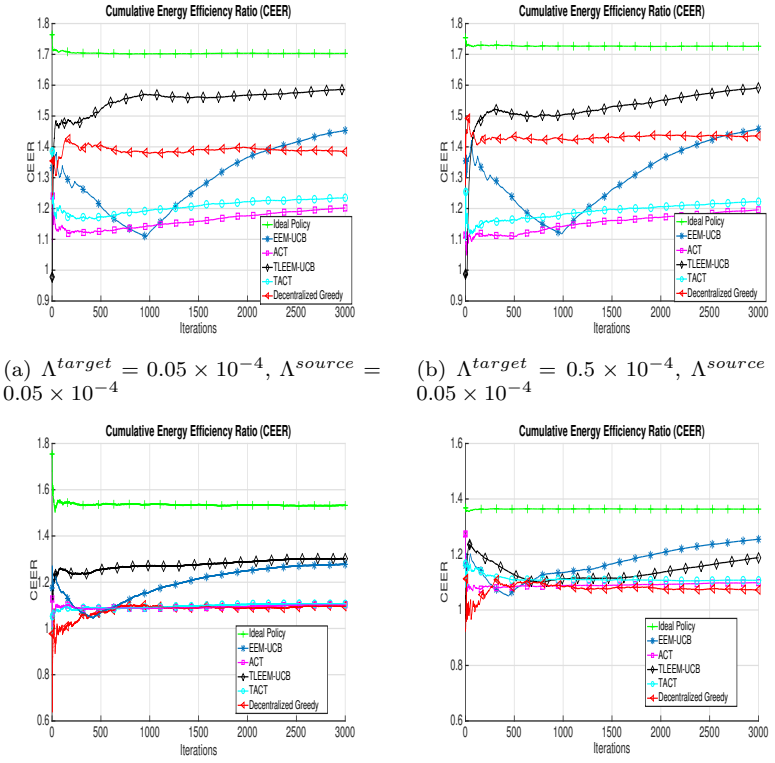
4.1 Convergence Analysis

Fig. 4 compares the convergence behaviors of the proposed EEM-UCB and TLEEM-UCB algorithms w.r.t. the Actor CriTic (ACT) [45], decentralized greedy [28] and Transfer Actor CriTic (TACT) [12] policies. The cumulative energy efficiency ratio (CEER) is presented for all policies in Fig. 4 which is defined as

$$\text{CEER}_{\pi} = \frac{\text{EE policy } \pi}{\text{EE when all BS are ON}}$$

Moreover, the global optimal solution achieved by an exhaustive search, and referred as ideal policy, is also shown in Fig. 4. The figure shows the behaviors of the policies in terms of CEER after 3000 iterations for 4 configurations of arrival rates in source, i.e. Λ^{source}, and current tasks i.e. Λ^{target}. These curves can be seen as the evolution of the network EE at a given hour of a day with a given arrival rate.

As depicted in Fig. 4, the network utilities of all algorithms tends to increase with time since their confidence on the best deployment strategy increases as the time elapses. However, the performance of all algorithms largely depends on the difference between the source and target task arrival rates. Our policies, EEM-UCB and TLEEM-UCB, converge towards the ideal policy, while ACT, TACT and decentralized greedy algorithms achieve a far suboptimal solution after 3000 iterations. The lower convergence rate of ACT and TACT algorithms is clear regarding these results. Our policy TLEEM-UCB generally performs better than all the others except when the source and target arrival rates are quite different, i.e. Fig. 4(d). From Fig. 4(a) to Fig. 4(d), the source arrival rate is fixed to $\Lambda^{source} = 0.05 \times 10^{-4}$ and the target arrival rate varies from $\Lambda^{target} = 0.05 \times 10^{-4}$ to $\Lambda^{target} = 2 \times 10^{-4}$. The transfer learning procedure is the most beneficial when $\Lambda^{source} = \Lambda^{target}$, Fig. 4(a) since TLEEM-UCB achieves performance jump start in the beginning and quickly converge towards the best configuration. On the contrary on Fig. 4(d), when the source and target arrival rates are significantly different, i.e. $\Lambda^{source} = 0.05 \times 10^{-4}$ and $\Lambda^{target} = 2 \times 10^{-4}$, transferred knowledge impacts the learning in a negative way and thus TLEEM-UCB performs worse than EEM-UCB. In that case, it is more beneficial to learn from scratch since the previously computed indexes B^i $\forall i$ has to be forgotten to learn a better configuration. From these results, we can state that temporal knowledge transfer improves the convergence speed of classical MAB approaches, but it also affects in a negative manner if traffic loads in a source and target environments are significantly different. The execution time needed for convergence does not exceed few minutes in a standard simulation platform using Matlab, since one iteration basically consists in the computation of the index $B^{i,h}$, which does not exceed few milliseconds.



(a) $\Lambda^{target} = 0.05 \times 10^{-4}$, $\Lambda^{source} = 0.05 \times 10^{-4}$

(b) $\Lambda^{target} = 0.5 \times 10^{-4}$, $\Lambda^{source} = 0.05 \times 10^{-4}$

(c) $\Lambda^{target} = 1 \times 10^{-4}$, $\Lambda^{source} = 0.05 \times 10^{-4}$

(d) $\Lambda^{target} = 2 \times 10^{-4}$, $\Lambda^{source} = 0.05 \times 10^{-4}$

Figure 4 Performance comparison under various homogeneous Poisson point process traffic intensity with transferred knowledge estimated from a source task with traffic intensity $\Lambda^{source} = 0.05 \times 10^{-4}$.

Fig. 5 presents the improvement in CEER of our algorithm when using TL concept compared to non transferred knowledge, i.e.

$$\text{CEER performance of improvement} = \frac{\text{CEER}_{TLEEM-UCB} - \text{CEER}_{EEM-UCB}}{\text{CEER}_{EEM-UCB}}$$

On can observe that the TL concept allows a performance jump-start at the early iterations compared to the simple EEM-UCB. The maximum rate of improvement is around 500 iterations and is as much better than the source and target arrival rates are similar. For instance, a gain about 28% is achieved after 500 iterations when $\Lambda^{source} = \Lambda^{target} = 0.05 \cdot 10^{-4}$ but reduces to only 5% when $\Lambda^{target} = 2 \times 10^{-4}$. For this setting, the improvement of TLEEM-UCB w.r.t. EEM-UCB is even negative after 3000 iterations, i.e. -5%, meaning that TL has a negative impact on the long-run on the network EE compared to EEM-UCB.

Finally, Fig. 6 shows how the network energy efficiency decreases when the number of BS increases. In that figure, the percentages of macro and micro BS are 50-50%, and the same settings than on Table 2 are used. Network EE reduces because the optimal configuration is not necessarily achieved after 3000 iterations, specially for high number of BS. Hence the selected configuration is not the optimal one and the gap increases as the number of BS increases as it can be inferred with the

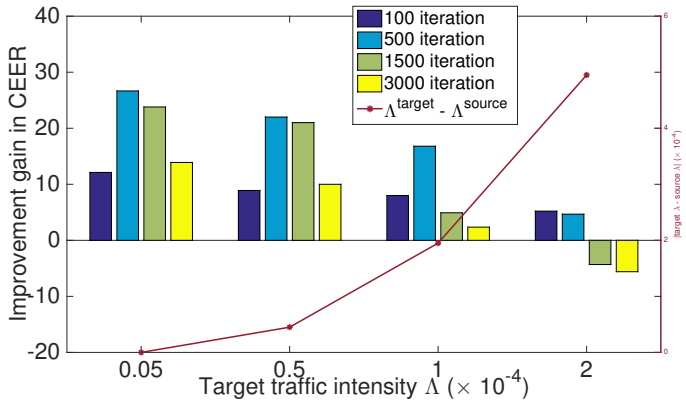


Figure 5 Improvement gain of TLEEM-UCB policy w.r.t. EEM-UCB policy for different target arrival rate. The bars corresponding to the left Y-axis reflect the gain in CEER while the right Y-axis represents the difference $\Lambda^{target} - \Lambda^{source}$.

ideal policy. The larger the number of base stations, the larger the exploration space cardinality making the convergence to the optimal configuration longer. EE achieved with the ideal policy always increases or remains constant w.r.t. the number of BS. Indeed, the larger the BS density, the larger the data rate, at least up to a limit where the interference generated by co-channel transmissions prevents from increasing the spectral efficiency. Hence, always selecting the best configuration of active BS increases the data rate and hence EE, in this configuration. We can also note the gain of learning policies compared to a scenario where all BS are always ON. It is also worth mentioning that the achieved EE with Decentralized Greedy finishes to outperform TLEEM-UCB when the number of BS is larger than 15 in this configuration, due to a more efficient search of a local optimum when the problem dimension begins to be high. However, Decentralized Greedy requires to exchange informations between nodes and hence the traffic overhead increases as the number of BS increases.

4.2 Performance under Periodic Traffic Load

We also investigate the effectiveness of the proposed learning framework when traffic loads periodically fluctuates. As stated in Section 2, real traffic load follows a periodical pattern that can be approximated by a sinusoidal function as in [29].

Fig. 7 compares the network EE achieved with our policies, i.e. EEM-UCB and TLEEM-UCB, with the previously introduced state-of-the-art algorithms, i.e. ACT, TACT and Decentralized Greedy, when traffic load is fluctuating during the day. On can first remark that all policies behave the same, except Decentralized Greedy which is inferior, at high traffic load from noon to 22h. Indeed, in high traffic load all BS need to be switched ON in order to satisfy the demand and hence all learning policies logically converge to the full deployment. Decentralized Greedy tends to switch OFF some BS, even in high traffic load, in order to save energy leading to a loss in EE. On the other hand, in lower traffic period, i.e. night time, less BS need to be switched ON to meet the QoS requirements and hence learning strategies make sense to optimize the network EE. In these time slots (1am to

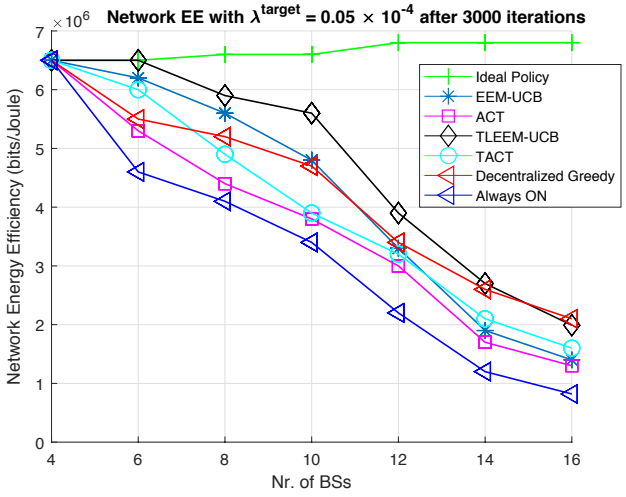


Figure 6 Network EE obtained with several learning algorithms with respect to the number of BS after 3000 iterations.

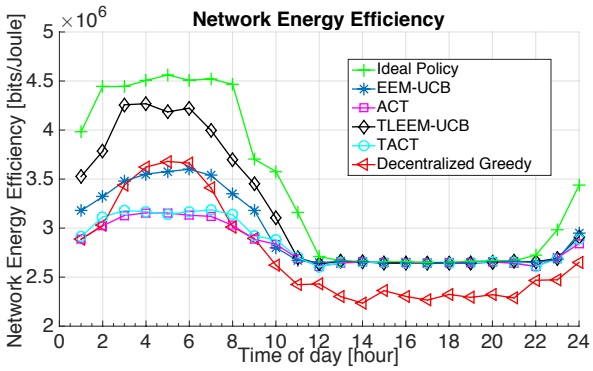


Figure 7 Network EE obtained with several learning algorithms with respect to time

8am), TLEEM-UCB achieves significantly higher EE compared to other algorithms from literature and reaches 95% of the maximum achievable EE, i.e. ideal policy. Moreover, TLEEM-UCB outperforms its counterpart, i.e. TACT, of about 34%. It also confirms that transferred learning improves the performance compared to non transferred knowledge policy, i.e. EEM-UCB, of about 23%.

Fig. 8 depicts the average percentage of energy savings achieved by the learning algorithms and the ideal policy during one day. The energy saving percentage is measured w.r.t. to the energy expenditure of a full deployment. As shown in Fig. 8, a large amount of energy saving is achieved by the proposed TLEEM-UCB policy, e.g. about 70% during low traffic load period (night time). Moreover, the difference between the ideal policy and TLEEM-UCB policy is less than 5%. On the contrary, ACT, TACT and EEM-UCB algorithms achieve only about 60% of energy saving. Decentralized Greedy procedure allows the most important energy saving gain which nearly equals the ideal policy performance in the night time. On can also remark that the later policy allows 20% of energy saving during high traffic period by

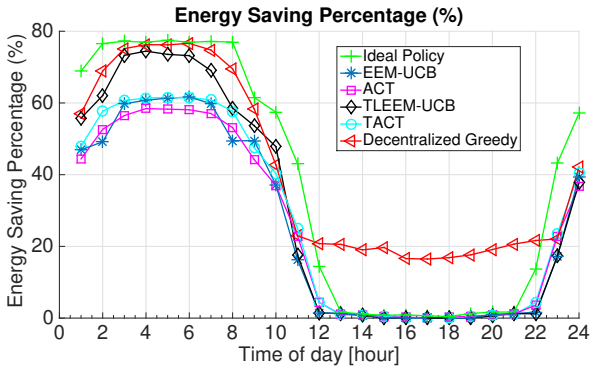


Figure 8 Average percentage of energy saving achieved by several learning algorithms with respect to time of day.

putting more BS into sleep mode, since the energy consumption is privileged. This improvement comes at the cost of user experience and comparatively less network EE as it has been observed in Fig. 7.

The impact of learning algorithms on the actual deployment of the network is also of great interest for operators. Figs. 9 and 10 represent the average number of active BS and the average number of switches that are performed at each time of the day, respectively. Fig. 9 gives insight on the average number of BS that is needed to be switched ON in order to meet the traffic variation along the day. As expected, the average number of BS needed at the night time is less than the one required at the peak period, leading to an increase of network EE and a decrease of energy consumption in night time as corroborated by Figs. 7 and 8 respectively. During the night time, Decentralized Greedy, TACT and ACT are the policies activating the less number of BS in that order. Our policies come after with an average of 5 BS switched ON, close to the optimal average number around 5.5, allowing higher EE than their counterparts. During the peak load in the afternoon, almost all policies activate the whole set of BS. Decentralized Greedy fluctuates around 8.5 BS in average allowing larger energy saving gain but lower EE. It is worth noting that the proposed policies, i.e. EEM-UCB and TLEEM-UCB, activate more micro BS than macro BS to cope with the varying traffic load and to save energy in the same time.

The results presented in Fig. 10 are important because of some practical constraints, i.e. time needed to turn ON/OFF the power amplifier (PA), lifetime of the PA. Indeed, if a learning policy requires to switch PA too often in each time slot, then it will significantly reduce the lifetime of PA and may cause additional power loss due to the initial burst of power consumption when an equipment is switched ON (non taken into account in this work). Our proposed policy, TLEEM-UCB, requires an average about 2 BS mode switches at each time slot in low load period (night time) which is significantly less compared to 5 mode switches with ACT and TACT algorithms in the same period while a little more than 3 switches are needed without TL, i.e. EEM-UCB. All algorithms but Decentralized Greedy do not require BS switches during high traffic periods. Whereas Decentralized Greedy, requires between 1 and 2 BS switches all along the day, irrespective of the traffic load.

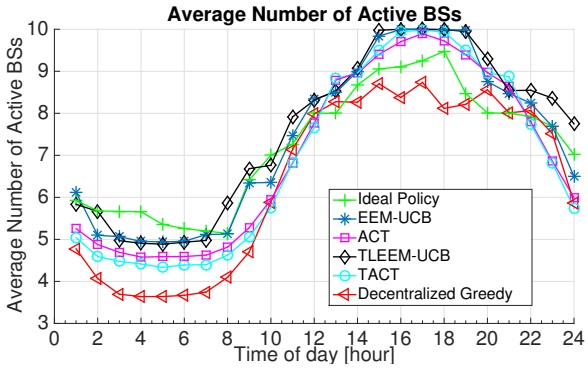


Figure 9 Average number of active BS suggested by several learning algorithms with respect to time of day.

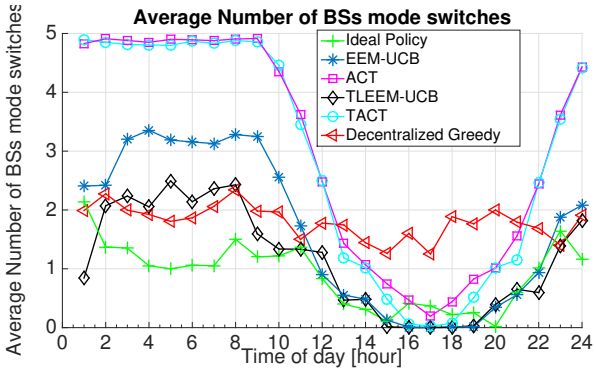


Figure 10 Average number of BS mode switch required at each time slot during one hour with respect to time of day.

To conclude this part and to shed the light on the transfer learning feature, let us assume that the average traffic load variation is very small from one day to another at the same day-time, i.e. $\Lambda^{\text{source}} \approx \Lambda^{\text{target}}$, which is a reasonable assumption, excepted between a week-day and a week-end day or between two consecutive days with occasional and exceptional events as it has been reported in [29]. As mentioned in Section 3.5, if one second is taken between two configuration switches, stable configuration is roughly achieved after one hour, which may appear relatively high. However, the applicability of TLEEM-UCB has to be thought on the long run, e.g. one week. Indeed, let us consider the particular time-range 10:00-11:00 am during the week. On Monday, the algorithm runs during one hour and saves the configuration achieved at 11:00 am. The next day at 10:00 am, the network just applies the configuration learned the day before and keeps like it is all along the week on the range 10:00-11:00 am without running again the algorithm. This strategy could be applied for each one hour-slot of the day during the first or two first days of the week and network just applies the computed configurations at each time slot for the rest of the week. Of course, this strategy does not work if important variations of the average traffic at a given time and between two days are observed, as it can be inferred in Fig. 4(d).

5 Conclusion

In this paper, the problem of BS switching operation for EE maximization in heterogeneous wireless cellular network has been tackled under the restless MAB framework. A reinforcement learning algorithm originally proposed in OSA scenario, has been adapted to deal with the optimal BS deployment in order to increase the global network EE. Furthermore and in order to increase the convergence rate of our EEM-UCB algorithm, we proposed to use the learnt knowledge acquired in previous time periods, leading to TLEEM-UCB policy. Our proposed algorithm has been proven to converge to an optimal solution as long as Markov chains governing the arms obey to certain conditions. Extensive numerical analysis shown the ability of our proposed policy to converge to the optimal deployment, maximizing EE. Transfer learning has been shown to be an effective solution to increase the convergence rate of our UCB algorithm when source and target arrival rates are not too different. Moreover, our policies have been shown to be able to follow a practical periodic traffic fluctuation. TLEEM-UCB can achieve 95% of EE achieved by the optimal BS configuration and up to 70% energy saving gain when traffic load is low (night time). Future work may include other index-based policies, such as Thomson sampling or Bayesian-UCB, that are known for their high performance in terms of regret in other scenarios. Moreover, spatial knowledge transfer between cells may also be of great interest for operators in a dynamic environment.

Appendix A: Sketch of Proof of Theorem 1

The regret of TLEEM-UCB policy is governed by the expected number of plays, $\mathbb{E}[T^{i,h}(n)]$, for any suboptimal BS switching operation i. Let l be a positive integer. Let us remind that $\mu^i = \sum_{S\in\mathcal{S}} S^i G_S^i \pi_S^i$. Following the steps as in [10] and including the historic time, the number of blocks a BS switching operation (action) i has been selected up to block $b(n)$ can be expressed as

$$F^{i,h}(b) = 1 + \sum_{t=2^Y}^{b} \mathbb{1}\{\mathbf{a}(t) = i\} \tag{14}$$

$$F^{i,h}(b) = l + \sum_{t=2^Y}^{b} \mathbb{1}\{\mathbf{a}(t) = i, T_2^i(t-1) \geq l\} \tag{15}$$

$$= l + \sum_{t=2^Y}^{b} \mathbb{1}\{B^{*,h}\left(t-1, T_2^*(t-1)\right) \leq B^{i,h}\left(t-1, T_2^i(t-1)\right), T_2^i(t-1) \geq l\} \tag{16}$$

$$\leq l + \sum_{t=2^Y}^{b} \mathbb{1}\{\exists \omega^i : l \leq \omega^i \leq t-1, B^{i,h}(\omega^i, t) > \mu^*\}$$
$$+ \mathbb{1}\{\exists \omega^* : 1 \leq \omega^* \leq t-1, B^{*,h}(\omega^*, t) \leq \mu^*\} \tag{17}$$

where the lower bound in the summation in (14) comes from the fact that each BS configuration are tried at least once, (15) comes from the fact that each action has been sensed at least l blocks up to block b. (16) comes from the reason why suboptimal action i is chosen, i.e. the index of the optimal action at block $t-1$,

$B^{*,h}\left(T_2^*(t-1), t-1\right)$, is below the index of the suboptimal action i. Moreover (16) is upper bounded by (17) because these two conditions are not exclusive. Taking the expectation on both sides and using union bound we get:

$$\mathbb{E}[F^{i,h}(b)] \leq l + \sum_{t=1}^{\infty}\sum_{\omega^i=l}^{t-1}\mathbb{P}(B^{i,h}(\omega^i,t) > \mu^*) + \sum_{t=1}^{\infty}\sum_{\omega^*=1}^{t-1}\mathbb{P}(B^{*,h}(\omega^*,t) \leq \mu^*) \quad (18)$$

The summation over t starts from 1 instead of 2^Y because it does not change the validity of the upper bound. Let remind that $G_S^{i,h}(T_2^i) = \frac{1}{T_2^i}\sum_{k=1}^{T_2^i} R_S^i(k) + \frac{1}{H_2^i}\sum_{k=1}^{H_2^i} R_S^{i,h}(k)$ denotes the empirical mean of quality observations R_S^i for action i in state S, $G_{\max}^S = \max_{i\in\mathcal{K}} G_S^{i,h}(T_2^i)$ is the maximum empirical reward and G^* is the empirical mean of the reward of the optimal action $*$, optimal in terms of both state and EE reward μ^*. it does not necessarily mean that $G^* = G_{\max}^S$. Moreover, let's remind that $\Delta\mu^i = \mu^* - \mu^i$. let's choose $l = \left\lceil \frac{4\alpha \ln\left(n_2 + H_2^i\right)}{(\Delta\mu^i)^2} \right\rceil$ and proceed from (18):

$$\begin{aligned}\mathbb{E}[F^{i,h}(b)] \quad &\leq \quad \left\lceil \frac{4\alpha \ln\left(n_2 + H_2^i\right)}{(\Delta\mu^i)^2} \right\rceil + \sum_{t=1}^{\infty} \sum_{\omega^i=\left\lceil \frac{4\alpha \ln\left(n_2 + H_2^i\right)}{(\Delta\mu^i)^2} \right\rceil}^{t-1} \mathbb{P}(B^{i,h}(\omega^i,t) > \mu^*) \\ &+ \quad \sum_{t=1}^{\infty}\sum_{\omega^*=1}^{t-1}\mathbb{P}(B^{*,h}(\omega^*,t) \leq \mu^*) \end{aligned} \quad (19)$$

We first start bounding the first part of (19), i.e. $\mathbb{P}(B^{i,h}(\omega^i,t) > \mu^*)$. Substituting $B^{i,h}(\omega^i,t)$ by its expression, and following the same steps than in [10, Appendix A] but using the number of times action i has been selected in SB2 block, i.e. H_2^i, we end up with

$$\mathbb{P}(B^{i,h}(\omega^i,t) > \mu^*) \leq \sum_{S\in\mathcal{S}} N_{\mathbf{h}^i} \exp\left(- \frac{\left(\omega^i + H_2^i\right)\left(\frac{\frac{\Delta\mu^i}{2} + D^{i,h}(\omega^i,t)}{r_S^i |\mathcal{S}| G_S^{i,h} \hat{\pi}_S^i}\right)^2 \varepsilon^i}{28} \right) \quad (20)$$

where $\hat{\pi}_S^i = \max\left\{\pi_S^i, 1-\pi_S^i\right\}$, $\hat{\pi}_{\max} = \max_{i\in\mathcal{K}} \hat{\pi}_S^i$ and $\epsilon^i = 1 - \lambda_2^i$ is the eigenvalue gap of action i, defined as the difference between 1 and the second largest eigenvalue of the i-th Markov chain. Moreover, (20) follows from [46, Th. 3.3] and from [47, Lem. 2.1] by considering $n = \omega^i$, $f(X_t^i) = \frac{\mathbb{1}\{S_t^i = S\} - G_S^{i,h}\pi_S^i}{G_S^{i,h}\hat{\pi}_S^i}$. The conditions of [46, Th. 3.3] are fulfilled if $G_S^{i,h} \geq \frac{1}{\hat{\pi}_{\max} + \pi_S^i}$. Consider an initial distribution $\mathbf{h}^i$ as defined in [41], $N_{\mathbf{h}^i}$ can be upper-bounded by $1/\pi_{\min}$ where $\pi_{\min} = \min_{S\in\mathcal{S}} \pi_S^i$. By following the same steps than in [10, Appendix A] we get from (20),

$$\mathbb{P}(B^{i,h}(\omega^i,t) > \mu^*) \leq \frac{\left|\mathcal{S}^i\right|}{\pi_{\min}} \left(t + H_2^i\right)^{-\frac{\Delta\mu^i \beta M_{\min} \varepsilon_{\min}}{28 S_{\max}^2 r_{\max}^2 G_{\max}^2 \hat{\pi}_{\max}^2}} \quad (21)$$

where $G_{\max} \equiv G_{\max}^S$, $r_{\max} = \max_{S\in\mathcal{S}, i\in\mathcal{K}} r_S^i$, $M_{\min} = \min_{i\in\mathcal{K}} M^{i,h}\left(\omega^i\right)$, $S_{\max} = \max_{i\in\mathcal{K}} \left|\mathcal{S}^i\right|$ and $\epsilon_{\min} = \min_{i\in\mathcal{K}} \epsilon^i$. Inserting (21) into first part of (19), and following

the same steps than in [10] we end up with

$$\sum_{t=1}^{\infty}\sum_{\omega^i=l}^{t-1}\mathbb{P}(B^{i,h}(\omega^i,t)\geq\mu^*)\leq\frac{|\mathcal{S}^i|}{\pi_{\min}}\sum_{t=1}^{\infty}\left(t+H_2^i\right)^{-2} \tag{22}$$

where (22) is obtained for of $\beta\geq 84S_{\max}^2r_{\max}^2G_{\max}^2\hat{\pi}_{\max}^2/\left(\varepsilon_{\min}\Delta\mu_i^RM_{\min}\right)$.

Similarly, one can bound the second part of (19) by following the same ideas than previously and applying the same steps than in [10, Appendix A] but introducing H_2^*, i.e. the number of times the best action has been chosen in historical period in SB2 block, we get

$$\mathbb{P}(B^{*,h}(\omega^*,t)\leq\mu^*)\leq\frac{|\mathcal{S}^*|}{\pi_{\min}}\left(t+H_2^*\right)^{-\frac{\varepsilon_{\min}\left(\alpha-2\sqrt{\alpha}\beta M_{\max}\right)}{28(S_{\max}G_{\max}r_{\max}\hat{\pi}_{\max})^2}} \tag{23}$$

where $M_{\max}=\max_{i\in\mathbb{K}}M^{i,h}\left(\omega^i\right)$. By choosing α such that $\frac{\varepsilon_{\min}\left(\alpha-2\sqrt{\alpha}\beta M_{\max}\right)}{28(S_{\max}G_{\max}r_{\max}\hat{\pi}_{\max})^2}\geq 3$ we obtain

$$\mathbb{P}(B^{*,h}(\omega^*,t)\leq\mu^*)\leq\frac{|\mathcal{S}^*|}{\pi_{\min}}\left(t+H_2^*\right)^{-3} \tag{24}$$

Substituting (24) into the second part of (19), we get

$$\sum_{t=1}^{\infty}\sum_{\omega^*=1}^{t-1}\mathbb{P}(B^{*,h}(\omega^*,t)\leq\mu^*)\leq\frac{|\mathcal{S}^*|}{\pi_{\min}}\sum_{t=1}^{\infty}\left(t+H_2^*\right)^{-2} \tag{25}$$

Furthermore, due to presence of transferred knowledge, we consider $l^+=\max\left(0,\frac{4\alpha\ln\left(n_2+H_2^i\right)}{(\Delta\mu^i)^2}-H_2^i\right)$ instead of l and the following bound follows from combining (22) and (25). Then, from (18):

$$\mathbb{E}[F^{i,h}(b)]\leq l^+ +\frac{|\mathcal{S}^*|}{\pi_{\min}}\sum_{t=1}^{\infty}\left(t+H_2^*\right)^{-2}+\frac{|\mathcal{S}^i|}{\pi_{\min}}\sum_{t=1}^{\infty}\left(t+H_2^i\right)^{-2} \tag{26}$$

Note that all observations in calculating the EEM-UCB indices come from the SB2 block. Let, SB2 block begin with observing regenerative state ζ^i and end with a return to the same ζ^i. The total number of time of sub-optimal action i is selected at the end of block $b(n)$ is estimated by considering the observations acquired in: i) the total number of plays of sub-optimal action i during SB2 block (upper-bounded by $1/\pi_{\min}^i$), ii) the total number of selections in SB1 before entering the SB2 block (upper-bounded by $\Gamma_{\max}^i$) and iii) Finally, one more selection resulting from the SB3 block which is state ζ^i. Thus, we have

$$\mathbb{E}\left[T^{i,h}(n)\right]\leq\left(\frac{1}{\pi_{\min}^i}+\Omega_{\max}^i+1\right)\mathbb{E}\left[F^i(b(n))\right]$$

Moreover, since $|\mathcal{S}^*|=\left|\mathcal{S}^i\right|=2$ and $S_{\max}=2$, $r_{\max}=1$ in our case, Theorem 1 follows.

Appendix B: Abbreviations

ACT	Actor-Critic
BS	Base Station
CEER	Cumulative Energy Efficiency Ratio
EE	Energy Efficiency
EEM-UCB	Energy Efficiency Maximization - Upper Confidence Bound
MAB	Multi-Armed Bandit
MDP	Markov Decision Process
MS	Mobile Station
OSA	Opportunistic Spectrum Access
PA	Power Amplifier
RL	Reinforcement Learning
SB	Sub-Block
SINR	Signal to Interference and Noise Ratio
TACT	Transfer Actor-Critic
TLEEM-UCB	Transfer Learning Energy Efficiency Maximization - Upper Confidence Bound
QoS	Quality of Service
RQoS-UCB	Restless Quality of Service - Upper Confidence Bound
UCB	Upper Confidence Bound

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and material
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Funding
This work has received a French government support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference No. ANR-10-LABX-07-01. The authors would also like to thank the Region Bretagne, France, for its support of this work.

Author's contributions
NM has provided the scientific and technical contents of the paper, deriving analytical results and numerical simulations. The three authors have contributed to the writing of the paper. All authors read and approved the final manuscript.

Acknowledgements
Not applicable.

Author details
[1]Brussels Airport Company, Brussels Airport, BE-1930, Zaventem, Belgium. [2]Univ. Rennes, INSA de Rennes, CNRS, IETR - UMR 6164, 20 avenue des Buttes de Coesmes, F-35000, Rennes, France. [3]Univ. Rennes, CNRS, IETR - UMR 6164, F-35000, Rennes, France.

References
1. Modi, N.: Machine learning and statistical decision making for green radio. PhD thesis, CentraleSupelec, Rennes (2017)
2. Marsan, M.A., Chiaraviglio, L., Ciullo, D., Meo, M.: Optimal energy savings in cellular access networks. In: IEEE International Conference on Communications Workshops (ICCW), pp. 1–5 (2009)
3. Fettweis, G.P., Zimmermann, E.: ICT energy consumption-trends and challenges. In: The 11th International Symposium on Wireless Personal Multimedia Communications (WPMC) (2009)

4. Son, K., Kim, H., Yi, Y., Krishnamachari, B.: Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks. IEEE Journal on Selected Areas in Communications **29**(8), 1525–1536 (2011)
5. Peng, C., Lee, S.-B., Lu, S., Luo, H., Li, H.: Traffic-driven power saving in operational 3G cellular networks. In: The 17th Annual International Conference on Mobile Computing and Networking (MobiCom), pp. 121–132. ACM, New York, NY, USA (2011)
6. Karl, H.: An overview of energy-efficiency techniques for mobile communication systems. Technical Report, Telecommunication networks Group, Technical University Berlin (Sept. 2003)
7. Oh, E., Krishnamachari, B., Liu, X., Niu, Z.: Toward dynamic energy-efficient operation of cellular network infrastructure. IEEE Communications Magazine **49**(6), 56–61 (2011)
8. Modi, N., Mary, P., Moy, C.: QoS driven channel selection algorithm for opportunistic spectrum access. In: IEEE Globecom Workshop on Advances in Software Defined Radio Access Networks and Context-aware Cognitive Networks (SDRANCAN), San Diego, USA (2015)
9. Robert, C., Moy, C., Wang, C.-X.: Reinforcement learning approaches and evaluation criteria for opportunistic spectrum access. In: IEEE International Conference on Communications (ICC), pp. 1508–1513 (2014)
10. Modi, N., Mary, P., Moy, C.: QoS driven Channel Selection Algorithm for Cognitive Radio Network: Multi-User Multi-armed Bandit Approach. IEEE Transactions on Cognitive Communications and Networking **3**(1), 49–66 (2017)
11. Taylor, M.E., Stone, P.: Transfer learning for reinforcement learning domains: A survey. J. Mach. Learn. Res. **10**, 1633–1685 (2009)
12. Li, R., Zhao, Z., Chen, X., Palicot, J., Zhang, H.: TACT: A transfer actor-critic learning framework for energy saving in cellular radio access networks. IEEE Transactions on Wireless Communications **13**(4), 2000–2011 (2014)
13. Niu, Z.: TANGO: traffic-aware network planning and green operation. IEEE Wireless Communications **18**(5), 25–29 (2011). doi:10.1109/MWC.2011.6056689
14. Chiaraviglio, L., Ciullo, D., Meo, M., Ajmone Marsan, M.: Energy-aware UMTS access networks. The 11th International Symposium on Wireless Personal Multimedia Communications (WPMC), 8–11 (2008)
15. Niu, Z., Wu, Y., Gong, J., Yang, Z.: Cell zooming for cost-efficient green cellular networks. IEEE Communications Magazine **48**(11), 74–79 (2010). doi:10.1109/MCOM.2010.5621970
16. Li, R., Zhao, Z., Wei, Y., Zhou, X., Zhang, H.: Gm-pab: A grid-based energy saving scheme with predicted traffic load guidance for cellular networks. In: IEEE International Conference on Communications (ICC), pp. 1160–1164 (2012). doi:10.1109/ICC.2012.6364637
17. Gong, J., Zhou, S., Niu, Z.: A Dynamic Programming Approach for Base Station Sleeping in Cellular Networks. IEICE Transactions on Communications **95**, 551–562 (2012). doi:10.1587/transcom.E95.B.551
18. Marsan, M.A., Chiaraviglio, L., Ciullo, D., Meo, M.: Optimal energy savings in cellular access networks. In: IEEE International Conference on Communications Workshops (ICCW), pp. 1–5 (2009). doi:10.1109/ICCW.2009.5208045
19. Marsan, M.A., Meo, M.: Energy efficient management of two cellular access networks. SIGMETRICS Perform. Eval. Rev. **37**(4), 69–73 (2010). doi:10.1145/1773394.1773406
20. Alam, A.S., Dooley, L.S., Poulton, A.S.: Traffic-and-interference aware base station switching for green cellular networks. In: 2013 IEEE 18th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), pp. 63–67 (2013)
21. Oh, E., Krishnamachari, B.: Energy savings through dynamic base station switching in cellular wireless access networks. In: IEEE Global Telecommunications Conference (GLOBECOM), pp. 1–5 (2010)
22. Karp, R.M.: Complexity of Computer Computations. In: Miller, R.E., Thatcher, J.W., Bohlinger, J.D. (eds.) Reducibility among Combinatorial Problems, pp. 85–103. Springer, Boston, MA (1972)
23. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York, NY, USA (1979)
24. Han, F., Safar, Z., Liu, K.J.R.: Energy-efficient base-station cooperative operation with guaranteed QoS. IEEE Transactions on Communications **61**(8), 3505–3517 (2013). doi:10.1109/TCOMM.2013.061913.120743
25. Soh, Y.S., Quek, T.Q.S., Kountouris, M., Shin, H.: Energy efficient heterogeneous cellular networks. IEEE Journal on Selected Areas in Communications **31**(5), 840–850 (2013)
26. Kim, J., Lee, H.W., Chong, S.: TAES: Traffic-aware energy-saving base station sleeping and clustering in cooperative networks. In: 13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pp. 259–266 (2015)
27. Konda, V., Borkar, V.: Energy-efficient base-station cooperative operation with guaranteed QoS. SIAM J. Contr. Optim. **38**(1), 94–123 (2013)
28. Wong, W.-T., Yu, Y.-J., Pang, A.-C.: Decentralized energy-efficient base station operation for green cellular networks. In: IEEE Global Communications Conference (GLOBECOM), pp. 5194–5200 (2012)
29. Oh, E., Son, K., Krishnamachari, B.: Dynamic base station switching-on/off strategies for green cellular networks. IEEE Transactions on Wireless Communications **12**(5), 2126–2136 (2013)
30. Zhou, S., Gong, J., Yang, Z., Niu, Z., Yang, P.: Green mobile access network with dynamic base station energy saving. ACM MobiCom **9**(262), 10–12 (2009)
31. Guo, W., O'Farrell, T.: Dynamic cell expansion with self-organizing cooperation. IEEE Journal on Selected Areas in Communications **31**(5), 851–860 (2013). doi:10.1109/JSAC.2013.130504
32. Tekin, C., Liu, M.: Online learning of rested and restless bandits. IEEE Transactions on Information Theory **58**(8), 5588–5611 (2012). doi:10.1109/TIT.2012.2198613
33. Oksanen, J., Koivunen, V., Poor, H.V.: A sensing policy based on confidence bounds and a restless multi-armed bandit model. In: 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), pp. 318–323 (2012)
34. Oksanen, J., Koivunen, V.: An order optimal policy for exploiting idle spectrum in cognitive radio networks. IEEE Transactions on Signal Processing **63**(5), 1214–1227 (2015). doi:10.1109/TSP.2015.2391072

35. Zhang, W.: Performance of real-time and data traffic in heterogeneous overlay wireless networks. In: Proceedings of the 19th International Teletraffic Congress (2005)
36. Hossain, M.F., Munasinghe, K.S., Jamalipour, A.: Distributed inter-bs cooperation aided energy efficient load balancing for cellular networks. IEEE Transactions on Wireless Communications **12**(11), 5929–5939 (2013)
37. Kim, H., de Veciana, G., Yang, X., Venkatachalam, M.: Distributed α-optimal user association and cell load balancing in wireless networks. IEEE/ACM Transactions on Networking **20**(1), 177–190 (2012)
38. Son, K., Chong, S., Veciana, G.D.: Dynamic association for load balancing and interference avoidance in multi-cell networks. IEEE Transactions on Wireless Communications **8**(7), 3566–3576 (2009)
39. Fehske, A.J., Richter, F., Fettweis, G.P.: Energy efficiency improvements through micro sites in cellular mobile radio networks. In: IEEE Globecom Workshops, pp. 1–5 (2009)
40. Alam, A., Dooley, L.: A scalable multimode base station switching model for green cellular networks. In: IEEE Wireless Communications and Networking Conference (2015)
41. Tekin, C., Liu, M.: Online algorithms for the multi-armed bandit problem with markovian rewards. In: The 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1675–1682 (2010). IEEE
42. Tekin, C., Liu, M.: Online learning in opportunistic spectrum access: A restless bandit approach. In: IEEE INFOCOM, pp. 2462–2470 (2011)
43. Wang, C.-C., Kulkarni, S.R., Poor, H.V.: Bandit problems with side observations. IEEE Transactions on Automatic Control **50**(3), 338–355 (2005). doi:10.1109/TAC.2005.844079
44. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning and Games. Cambridge University Press, New York, NY, USA (2006)
45. Li, R., Zhao, Z., Chen, X., Zhang, H.: Energy saving through a learning framework in greener cellular radio access networks. In: IEEE Global Communications Conference (GLOBECOM), pp. 1556–1561 (2012)
46. Lezaud, P.: Chernoff-type bound for finite markov chains. Annals of Applied Probability **8**, 849–867 (1998)
47. Anantharam, V., Varaiya, P., Walrand, J.: Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards. IEEE Transactions on Automatic Control **32**(11), 977–982 (1987)