Contextual Bandit Approach for Energy Saving and Interference Coordination in HetNets

Jose A. Ayala-Romero*, Juan J. Alcaraz*, Andrea Zanella†, Michele Zorzi†
*Department of Information and Communications Technologies, Technical University of Cartagena, Spain
†Department of Information Engineering, University of Padova, Italy
email: josea.ayala@upct.es, juan.alcaraz@upct.es, zanella@dei.unipd.it, zorzi@dei.unipd.it

***Abstract*—This paper addresses the joint problem of energy saving and interference coordination in heterogeneous networks (HetNets) using a contextual bandit formulation. We propose a semi-distributed scheme consisting of a learning agent and local controllers. The learning agent comprises a neural network (NN) classifier and a Multi-Armed Bandit (MAB) algorithm. The NN classifier is dynamically trained to choose a subset of configurations (i.e., feasible configurations in terms of QoS) based on the context information (network state). Then, the MAB algorithm picks one control (i.e., global configuration parameters) among those selected by the NN classifier, with the aim of improving the energy efficiency. These global configurations are interpreted by the local controllers on each network sector. This scheme allows the learning agent to progressively learn the best policy by observing the network state and the performance of the chosen configurations in terms of energy consumption and QoS. Our numerical results show an energy saving close to 20% with respect to a default policy and an improvement of 13% with respect to addressing energy saving and interference coordination separately.**

I. Introduction

A promising step towards increasing the network capacity is based on the dense deployment of small cells, thus realizing the so-called Heterogeneous Networks (HetNets), considered as one of the key technologies in 5G systems [1]. Nevertheless, the densification of HetNets poses two main challenges: the increment of the energy consumption due to the increase in the number of cells, and the higher inter-cell interference. These two issues are intertwined, in that the energy saving mechanisms can affect the inter-cell interference (e.g., when switching cells on and off) and, on the other hand, the interference coordination mechanisms determine the utilization of the radio resources in the different cells, with an impact on their energy consumption. Nonetheless, so far these two issues have been addressed separately.

In contrast, this paper proposes a learning framework that jointly considers the energy efficiency and the interference coordination functionalities. Figure 1 provides a high-level description of our proposal. Its main elements are: the controlled HetNet comprising m sectors, a central entity (Global Controller) and m Local Controllers (LC), one per sector. The Global Controller receives performance metrics and context data from the network and makes global control decisions based on this information. The global control, which comprises both interference and energy saving decisions, is broadcast to the Local Controllers. Each Local Controller decides how to effectively translate the global control into a local decision based on the specific pico eNodeB (eNB) deployment within its sector. The main feature of the Global Controller is that it

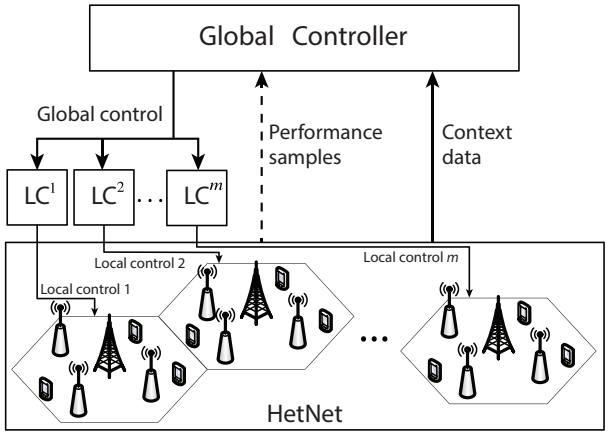

Fig. 1. System scheme comprising the Global Controller that learns efficient global controls from the network state and performance metrics, and m Local Controllers which compute local controls for each sector.

is capable of learning from the observed performance of past decisions to gradually approach the most efficient operation policy. This learning process is performed by a novel algorithm based on a contextual bandit approach [2]. The main challenge addressed by our proposal is to minimize the performance losses associated to the learning process. This is accomplished by means of two strategies: first, the use of two decision levels (global and local) which notably reduces the dimension of the control space at the Global Controller, and second, the introduction of a classification algorithm for discarding the controls that are considered unable to meet the QoS requirements.

We consider a HetNet in which each macro cell contains multiple pico eNBs in each sector. The pico eNBs can be in active (on) or sleeping mode (off), while the macro eNBs are always active. We apply the enhanced Inter Cell Interference Coordination (eICIC) mechanism proposed by the 3GPP for LTE-A Networks, which will be described in more detail in Sec. II-A. The eICIC parameters are the Cell Range Expansion (CRE) bias and the Almost Blank Subframe (ABS) ratio.

Some previous works have addressed the problem of energy efficiency in HetNets. One of the most usual approaches is to formulate the problem as a Markov Decision Process (MDP) [3], [4]. The inherent computational complexity of the MDPs implies the use of approximate dynamic programming approaches such as Reinforcement Learning (RL) [4]. The complexity and scalability of a learning approach is directly related with the dimensions of the state and control spaces

(curse of dimensionality). Our proposal overcomes this issue by applying two strategies: 1) using two decision levels (global and local), which makes the dimension of the control space independent of the number of macro sectors, and 2) projecting the controls on a low-dimensional space, which makes the control space scale linearly with the number of pico eNBs, instead of exponentially.

Other works address the problem of interference coordination in HetNets using learning algorithms [5], [6]. For example, the authors in [6] propose Q-learning algorithms for learning ABS ratio and CRE bias. However, to the best of our knowledge, our work is the first that *jointly* addresses energy saving and interference coordination in HetNets. In contrast to approaches relying on mathematical models of the network [7], our proposal is *model-free*, i.e., it operates without making any assumption about the network [8]. The main contributions of our proposal are summarized as follows:

- To jointly control the energy saving and the interference coordination mechanisms in HetNets with QoS guarantees.
- To learn efficient configurations during network operation thanks to a novel contextual bandit approach that combines a Multi-Armed Bandit algorithm with a Neural Network classifier.

We evaluate our proposal in an LTE-A network simulator following the 3GPP guidelines [9]. The next section details the interference coordination mechanism and the consumption model associated to this technology.

II. Interference Management and Power Consumption Model

A. Interference Management in LTE-A: eICIC

eICIC is an interference coordination mechanism for heterogeneous networks defined in 3GPP Release 10 (LTE-A). To minimize inter-cell interference, the eICIC schedules the radio resources for pico and macro eNBs in different time periods (subframes). It comprises two main features: *Cell Range Expansion (CRE)* and *Almost Blank Subframe (ABS)*.

The CRE increases the pico eNBs footprint by adding a bias to their received signal reference power. It is intended to balance the offloading (from macro to pico eNBs) in the network. However, the User Equipments (UEs) located in the extended region (CRE region) will generally have a poor channel quality due to the high interference received from the macro eNB. ABS is motivated by the need to improve the performance of UEs located in CRE regions and consists of reserving certain subframes for pico cell traffic only, muting the macro eNBs in those resources (Almost Blank Subframes). The ABS ratio defines the portion of muted subframes over the total number of subframes (muted and not). We consider synchronized muting, as recommended by the 3GPP [10]. That is, the eICIC controls are applied globally to a cluster of macro eNBs with homogeneous traffic profile.

B. eNB Power Consumption Model

The eNB consumption model used in this work is based on 3GPP guidelines [11]. The power consumption of some of the components of an eNB depends on its load. Thus, it is common to assume a linear relationship between RF output power and power consumption of eNB transceivers (TRXs) [11]. The power consumption model of a pico eNB j is given by:

$$C_{\mathrm{p}}^{j} = e^{j} \cdot N_{\mathrm{TRX}} \cdot (P_0 + R^{j} \cdot P_{\max}) + (1 - e^{j}) \cdot N_{\mathrm{TRX}} \cdot P_{\mathrm{sleep}} + \Delta \quad (1)$$

$$\Delta = \begin{cases} \beta \cdot P_0 & \text{when eNB } j \text{ is switched on} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where:

- $e^{j} = 1$ when the pico eNB j is active and $e^{j} = 0$ otherwise.
- N_{TRX} is the number of TRXs.
- $P_{\max}$ is the maximum RF output.
- P_0 represents the power consumption at zero RF output power.
- $R^{j} \in [0, 1]$ is the load factor of the pico eNB j and depends on the ABS ratio, the CRE bias, the traffic intensity and the location of UEs.
- P_{sleep} is the power consumption of TRX components in sleep mode.
- β is the portion of P_0 needed to switch on the pico eNB TRXs.

Note that Δ captures the consumption associated to switching on a sleeping pico eNB. The power consumption of the macro eNB i is given by

$$C_{\mathrm{m}}^{i} = N_{\mathrm{TRX}} \cdot (P_0^{\mathrm{m}} + R^{i} \cdot P_{\mathrm{max}}^{\mathrm{m}}) \cdot (1 - \gamma) + N_{\mathrm{TRX}} \cdot P_0^{\mathrm{m}} \cdot \gamma \quad (3)$$

where γ denotes the ABS ratio. Given the influence of the ABS ratio and the CRE bias on C_{p}^{j} and C_{m}^{i}, our proposal includes these parameters in the control of energy consumption, as explained below.

III. System Model and Problem Formulation

A. System Model

We consider an LTE-A access network comprising M sectorized macro eNBs denoted by the set $\mathcal{M}$ and P pico eNBs for each macro sector. Let $\mathcal{P}^{m}$ be the set of pico eNBs overlapping the macro sector m, and $\mathcal{P} = \{\mathcal{P}^{1}, \mathcal{P}^{2}, \ldots, \mathcal{P}^{M}\}$ the set of all pico eNBs in the network. We denote the ABS ratio and the CRE bias by $\gamma \in \Gamma$ and $\phi \in \Phi$, respectively, where Γ and Φ are the finite sets comprising all available configurations for these parameters. The time is divided into time stages denoted by $k \in \{0, 1, \ldots\}$.

1) States: Let e_k^{j} be the state of the pico eNB $j \in \mathcal{P}^{m}$ at stage k, where $e_k^{j} = 1$ when the pico eNB is switched on and $e_k^{j} = 0$ otherwise. Let $p_k^{m} = \{e_k^{j}\}_{j \in \mathcal{P}^{m}} \in \mathcal{E}^{m}$ ($\mathcal{E}^{m} = \{0, 1\}^{P}$) be the vector indicating the on/off state of all the pico eNBs in $\mathcal{P}^{m}$. Let $p_k = (p_k^{1}, \ldots, p_k^{M}) \in \mathcal{E}$ be the state of all picos in the network at stage k where $\mathcal{E} = \{0, 1\}^{P \cdot M}$. Let $\lambda_k \in \Lambda$ denote the aggregate traffic load in the network at stage k, where $\Lambda = [0, \lambda^{\max}]$. We define the system state at stage k as $x_k = (\lambda_k, p_{k-1}) \in \mathcal{X}$, where $\mathcal{X} = \Lambda \times \mathcal{E}$ is the state space.

2) Controls: We denote $u_k = (p_k, \gamma_k, \phi_k) \in \mathcal{U}$ as the network control, where $\mathcal{U} = \mathcal{E} \times \Gamma \times \Phi$ is the control space.

Given x_k, the decision maker selects a control u_k based on its previous knowledge. The next state $x_{k+1} = (\lambda_{k+1}, p_k)$ depends on the current control and the traffic at the next stage, which is unknown in advance.

3) Feedback functions: We define two feedback functions: $C : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ providing the aggregated power consumption of macro and pico eNBs in the network (using the model in Sec. II-B) and $Q : \mathcal{X} \times \mathcal{U} \to [0, 1]$ which gives the ratio of UEs in the network with satisfactory QoS that, in this study, we consider as a minimum guaranteed throughput. Thus, the feedback function Q indicates the portion of UEs with throughput above a selected threshold. Note that the values obtained from the feedback functions C and Q are random variables due to the randomness of UE locations and traffic demands.

B. Contextual Bandit Problem Formulation

We define a policy as a function $\pi : \mathcal{X} \to \mathcal{U}$ which maps states into controls. A learning agent following a policy π operates as follows: (i) the learning agent obtains the network state x_k at stage k and selects the control u_k that policy π prescribes for x_k. (ii) The network operates according to the control u_k during stage k, gathering performance measures from each eNB in order to obtain the feedback values (C_k, Q_k) that are sent back to the learning agent at the end of stage k. (iii) The learning agent receives the feedback and updates the policy π accordingly.

Our goal is to learn, stage by stage, a policy π satisfying our dual objective: to minimize the power consumption while satisfying the QoS requirement. We define the following cost function capturing the tradeoff between these two objectives:

$$\rho(x, u) = C(x, u) + \delta \cdot \max(0, Q_{\min} - Q(x, u)). \quad (4)$$

The first term accounts for the consumption in the network and the second one is a penalty that is applied whenever the QoS threshold ($Q_{\min}$) is not satisfied. The coefficient δ is a weighting factor that regulates the importance of this penalty.

We define the optimal policy π^* as the one that minimizes the average cost per stage in the long term, i.e.,

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmin}} \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \rho(x_k, \pi(x_k)) \quad (5)$$

where Π is the set of all possible policies. The pseudo-regret (referred to as regret henceforth) of a policy π over N stages is given by

$$R_\pi(N) = \sum_{k=0}^{N} E[\rho(x_k, \pi^*(x_k))] - E[\rho(x_k, \pi(x_k))]. \quad (6)$$

This metric accumulates the loss incurred when selecting a suboptimal control at each stage and can then be used to assess the performance of a policy. That is, the lower the regret of a policy, the closer the policy to the optimal one.

However, in order to find policies minimizing (6), it is necessary to deal with the curse of dimensionality since the dimensions of the state and control spaces ($\mathcal{X}$ and $\mathcal{U}$) make the problem unmanageable as the network size grows. In particular, the size of the set $\mathcal{E}$ grows exponentially with the number of pico eNBs per sector and with the number of sectors ($|\mathcal{E}| = 2^{P \cdot M}$). In the next section we present our proposal aimed at addressing these problems.

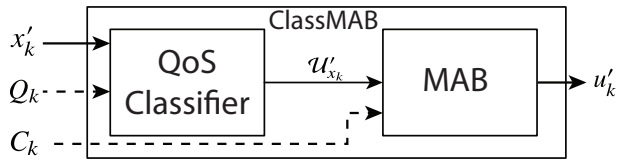

Fig. 2. Detailed scheme of the Global Controller composed of a QoS classifier, which pre-selects the set of controls $\mathcal{U}'_{x_k}$ that are considered capable of meeting the QoS requirements, and the Multi-Armed Bandit module, which performs the learning process considering only the controls provided by the classifier at each stage. The context data is denoted by x_k and Q_k and C_k are the performance measurements.

IV. Contextual Bandit Proposal

As described in the introduction and in Fig. 1, the Global Controller (*ClassMAB*) is in charge of the learning and control functions. Fig. 2 shows the internal structure of *ClassMAB*, which comprises two elements, a QoS Classifier and Multi-Armed Bandit module. *ClassMAB* and Local Controllers are described in this section.

Let us first introduce some notation regarding the global states and controls. We define $p' = \sum_i p_i^m \in \mathcal{E}' = \{0, \ldots, P\}$ as the projection of the parameter p^m, i.e., the total number of active pico eNBs in every sector[1]. We define $x' = (\lambda, p') \in \mathcal{X}'$ and $u' = (p', \gamma, \phi) \in \mathcal{U}'$ as the global state and the global control, respectively. Note that the dimensions of the global state and control spaces ($\mathcal{X}'$ and $\mathcal{U}'$) are now linear with P and independent of M. The use of global controls is a reasonable assumption when all sectors in $\mathcal{M}$ are selected with homogeneous traffic profile, e.g., city center, outskirts of the city, business area, etc. The data flow of our proposal is detailed as follows (Fig. 2):

- The classifier receives the global network state x'_k at stage k. Then, it computes the set $\mathcal{U}'_{x_k} \subset \mathcal{U}'$ of available controls satisfying the QoS, where $\mathcal{U}'_{x_k} = \{u' \in \mathcal{U}' : E[Q(x_k, u')] > Q_{\min}\}$. The set $\mathcal{U}'_{x_k}$ is sent to the MAB.
- The MAB selects a global control $u' \in \mathcal{U}'_{x_k}$ trying to minimize the energy consumption. This control is sent to the *LCs*.
- Each *LC* receives the global control u'_k and generates a local control u_k^m for its respective sector $m \in \mathcal{M}$.
- Each sector $m \in \mathcal{M}$ operates with its corresponding local control u_k^m during stage k. At the end of the stage, *ClassMAB* obtains a global feedback computed from the information provided by the eNBs of each sector. Specifically, the classifier and the MAB obtain a QoS (Q_k) and a consumption (C_k) measure, respectively.
- Using this feedback, *ClassMAB* updates its knowledge for the next iteration.

The following subsections explain in detail each module of our proposal.

[1]Note that p' is a global network parameter, i.e., the projection of p^m gives the same value of p' for all $m \in \mathcal{M}$.

A. QoS Classifier

The QoS Classifier denoted by Θ generates at each stage k the set of controls $\mathcal{U}'_{x_k} \subset \mathcal{U}'$ given the state x'_k, i.e.:

$$\mathcal{U}'_{x_k} = \{u' \in \mathcal{U}' : \Theta(\lambda_k, u') = 1\} \tag{7}$$

where Θ is the classifier's outcome function which equals 1 if control u' is expected to satisfy the QoS given the traffic intensity λ_k (extracted from the global network state x'_k), and 0 otherwise.

The classifier is based on a neural network (NN) composed of two hidden layers of 50 nodes each. A rectified linear unit (ReLU) is used as a nonlinear activation function in both hidden layers, and a sigmoid activation function is used in the output layer. The NN receives at each stage k the feedback Q_k from the network. Using this feedback, the NN is trained using the gradient-based algorithm in [12] to minimize the cross-entropy cost function.

B. Multi-Armed Bandit with Constrained Control Availability

The MAB selects a control u'_k from the set $\mathcal{U}'_{x_k}$ provided by the QoS Classifier at each stage. Classical MAB algorithms [13] assume a fixed set of controls at every stage which makes these algorithms unsuitable for this setting. We propose for this task a modification of the descent ε-greedy MAB algorithm referred to as Control Constrained Descent ε-greedy (CCD ε-greedy) algorithm, shown in Algorithm 1.

Algorithm 1 CCD ε-greedy algorithm

Input parameters: ε_0
for each stage k **do**
  Receive the set of available controls $\mathcal{U}'_{x_k}$
  Generate the set $\mathcal{U}'_{\text{ini}} \subset \mathcal{U}'_{x_k}$ of unexplored controls
  if $\mathcal{U}'_{\text{ini}}$ is empty **then**
    if random number $\in [0, 1] < \varepsilon_0/k$ **then**
      Randomly and uniformly pick one control u' from the set $\mathcal{U}'_{x_k}$
    else
      Select the control $u' \in \mathcal{U}'_{x_k}$ with minimal expected reward
    end if
  else
    Randomly and uniformly pick one control u' from the set $\mathcal{U}'_{\text{ini}}$
  end if
  Obtain the consumption sample C_k of the configuration u'_k
  Update the expected reward of u'_k with C_k.
end for

C. Local Controllers (LCs)

Each Local Controller $m \in \mathcal{M}$ (LC^m) generates a control u^m_k from the global control u'_k at each stage k. As we discussed in Section II-A, the eICIC controls are inherently global, and therefore, the task of *LCs* is focused on the projection $\mathcal{E}' \to \mathcal{E}$.

We define the *cell adjacency* value d^j of a pico eNB $j \in \mathcal{P}^m$ in sector m as the weighted average of the distances to the remaining eNBs in the sector:

$$d^j = w \cdot d^j_m + (1 - w) \cdot d^j_p \tag{8}$$

where d^j_m is the distance to macro eNB m, d^j_p is the average distance to the other pico eNBs in sector m and $w \in [0, 1]$ is a weighting factor which, in this study, is set to $w = 0.4$.

Let $\mathcal{D}^m = \{d^i\}_{i \in \mathcal{P}^m}$ denote the set containing the cell adjacency values of all pico eNBs in $\mathcal{P}^m$. Our strategy consists of switching off first the pico eNBs with lower values of cell adjacency since these eNBs can offload their traffic to near eNBs with less degradation in the channel quality.

The LC^m can compute the values of each element e^j of p^m from the control p' as follows:

$$e^i = \begin{cases} 1 & \text{if } i \in \mathcal{I} \text{ where } \mathcal{I} = \xi_{p'}(\mathcal{D}^m) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where the operator $\xi_{p'}(\mathcal{D}^m)$ gives us the indices of the p' greatest values in $\mathcal{D}^m$, that is, the indices of the p' pico eNBs with the greatest values of cell adjacency.

V. Numerical results

A. Description of the Simulation Framework

The simulation framework is based on 3GPP guidelines for the evaluation of LTE networks [9]. The network layout comprises 5 sectorized macro eNBs (120 degrees) and several pico eNBs overlapping with each macro coverage area. We simulate the central sector using the remaining sectors to emulate the aggregated interference of a larger network. The wireless channel is composed of the pathloss and the stochastic shadow fading. The aggregated interference at each UE receiver consists of the power received from all interfering eNBs in the sector (picos and macro) plus the interference from the macro eNBs from other sectors.

Each incoming UE genenerates one throughput measurement, which is defined according to the 3GPP guidelines [9]. A UE satisfies the QoS requirement if its throughput is above $T_{\min} = 100$ kbps. The QoS function Q gives the ratio of UEs satisfying $T_{\min}$. The power consumption model is defined in Section II-B and the values of its parameters are shown in Table I.

The number of pico eNBs per sector is $P = 6$. The sets of available configurations of eICIC parameters are $\Gamma = \{0, \frac{1}{8}, \frac{2}{8}, \ldots, \frac{7}{8}\}$ and $\Phi = \{0, 6, 9, 12, 18\}$. The remaining simulation parameters are shown in Table I. The simulation framework has been developed using Python. The neural networks are implemented using TensorFlow.

B. Simulation results

In this section, we provide numerical results for our proposal and compare its performance with the following benchmarks:

- *Oracle* provides an upper bound in terms of regret. It selects, for each network state x_k at stage k, the optimal control $u^*_k \in \mathcal{U}'$, which is found by exhaustive search.
- *Default configuration* is a fixed control where energy saving and interference mechanisms are deactivated.
- *NeuralBandit* [14] implements a contextual bandit algorithm based on neural networks. It is aimed at learning the cost function $\rho(x', u')$ for each arm u' given the context x'. It comprises, for each arm, a neural network composed

Network layout	5 sectorized macro eNBs, 500 m ISD, P = 6 pico eNBs per sector
LTE frame duration	Subframe 1 ms, Protected-subframe pattern 8 ms, Frame 10 ms
Transmit power	Macro eNB 46 dBm, pico eNB 30 dBm
Antenna pattern	Macro: $A_H(\phi) = -min[12(\frac{\phi}{\phi_{3dB}})^2, A_m]$, A_m = 70 degrees A_m = 25 dB, Pico: Omnidirectional
Antenna gains	macro: 14 dBi; pico: 5 dBi; UE: 0 dBi
Macro to UE path loss	128.1 + 37.6 · $\log_{10}(R[\text{Km}])$ where R is the macro eNB to UE distance
Pico to UE path loss	149.7 + 36.7 · log10(R[Km]) where R is the pico eNB to UE distance
Shadow fading	Lognormal distribution with 10 dB standard deviation
Minimum distances	Macro - pico: 70 m; Macro - UE: 35 m; Pico - pico: 40 m; Pico - UE: 10 m
Macro consumption parameters	N_{TRX} = 6, P_0^{m} = 130 W, $P_{\text{max}}^{\text{m}}$ = 20 W
Pico consumption parameters	N_{TRX} = 2, P_0 = 56, P_{max} = 6.3, P_{sleep} = 39 W

TABLE I
SIMULATION PARAMETERS

of two hidden layers with 20 nodes each. At each stage, an arm is selected according to an ε-greedy policy with decreasing ε. Then, the selected arm is trained using [12] with the feedback measures C_k and Q_k.

- *ClassMAB (ES)* evaluates our proposal only controlling the energy saving mechanism, i.e., $\gamma = \phi = 0$.

Other state-of-the-art contextual bandits algorithms, e.g. [2], assume that the expected value of each arm is linear with respect to the context. However, our cost function (4) shows a threshold structure, making these algorithms unsuitable for this application.

Our simulations are aimed at assessing the performance of the algorithms in two phases: a *training* phase composed of 1200 epochs with 200 stages of variable traffic intensities, and a *test* phase where the learning state of the algorithms is frozen setting a greedy policy (i.e., using the control with the lowest expected regret). Both *ClassMAB* and *NeuralBandit* start with an initialization period where each arm is selected once and are configured with $\varepsilon_0 = 30$. The weighting factor of the cost function is set to $\lambda = 1 \cdot 10^6$ and $Q_{\text{min}} = 0.6$.

Fig. 3 shows the accumulated regret during the training phase. The regret slope of *ClassMAB* tends to zero at the end of the training phase, i.e., it selects controls very close to the optimum. *NeuralBandit* shows a slower learning rate and a steeper slope in its long term regret. *ClassMAB (ES)* obtains a constant regret slope, which reflects the loss incurred when neglecting the interference coordination mechanism. The regret of the default configuration grows linearly, exceeding that of our proposal at the end of the training period. Fig. 4 shows the value of the cost function at each stage. We can observe the fast convergence of *ClassMAB*, and its ability to operate at smaller cost values compared to other benchmarks.

The test phase was simulated for a one day period, using a stage duration of 10 minutes (a total of 144 stages). At each stage, a random traffic intensity was generated according to the traffic profile shown in Fig. 5. Fig. 6 shows the regret measured during the test phase. After training, *ClassMAB* incurs the lowest regret. *NeuralBandit* obtains the second best result despite its slower learning rate during the training phase. Fig. 7 shows the consumption at each stage. Note that, in general, the consumption pattern resembles the traffic profile (Fig. 5). The crosses on the consumption curves indicate the stages where the selected control does not satisfy the QoS. The benchmarks not using the interference coordination

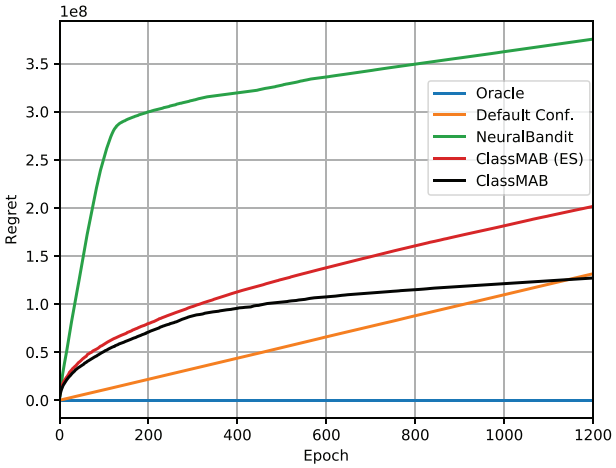

Fig. 3. Regret measured during the training phase. The incurred regret at each epoch is the summation of the regret of each one of its corresponding 200 stages.

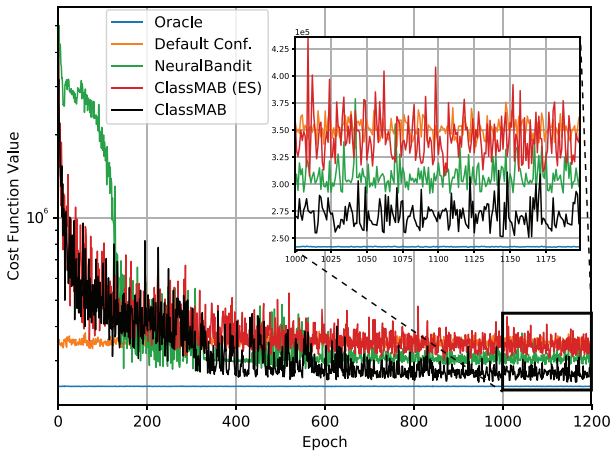

Fig. 4. Evolution of the cost function during the training phase.

mechanism, besides suffering higher regret, occasionally miss the QoS requirements even using higher consumption controls. This highlights the importance of a synergistic and joint use of interference coordination and energy saving mechanisms. *ClassMAB* shows a power consumption closer to the optimum and always satisfies the QoS requirement. The numerical results of the test phase are summarized in Table II where we show the energy savings with respect to the default configuration and the ratio of stages where QoS has been satisfied.

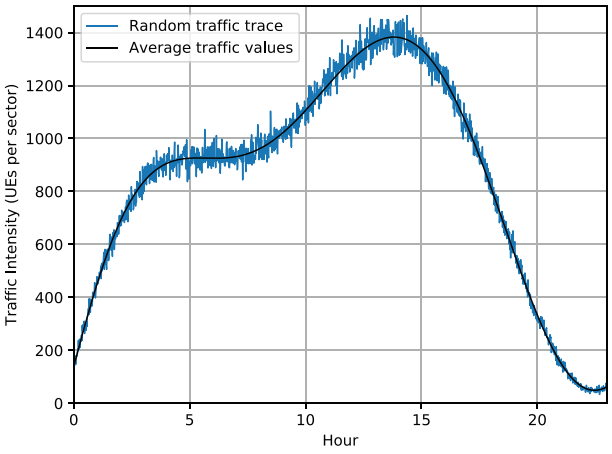

Fig. 5. Traffic pattern of one day considered in the test phase.

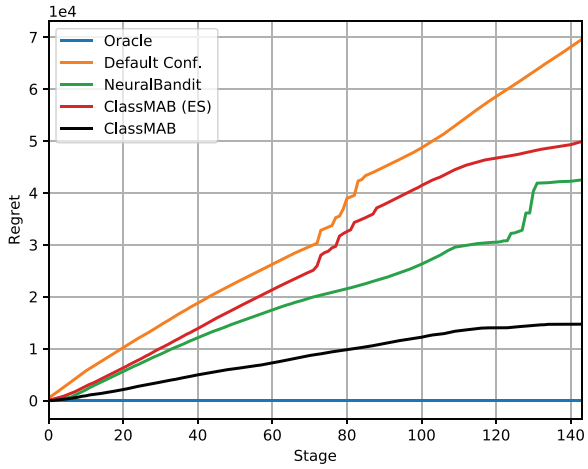

Fig. 6. Evolution of the regret in the test phase.

VI. Conclusion

This paper presented a novel contextual bandit approach addressing energy saving and interference coordination in LTE-A HetNets. In our proposal, the learning agent (Global controller) learns efficient global configurations based on network performance samples and context information (global state of the network). These global configurations are interpreted locally by the Local controllers on each network sector. Breaking down the decision into two levels allows the complexity of the control space to grow linearly rather than exponentially, favoring the scalability of our mechanism. Our proposal is a novel combination of a neural network classifier and a multi-armed bandit algorithm which is able to minimize the energy consumption providing QoS guarantees. Our numerical results show an energy saving close to 20% with respect to a fixed default policy. The considerable improvement associated to addressing energy saving and interference coordination problems jointly is also shown.

VII. Acknowledgements

This work was supported by project grant AEI/FEDER TEC2016-76465-C2-1-R (AIM). Jose A. Ayala-Romero acknowledges personal grant FPU14/03701.

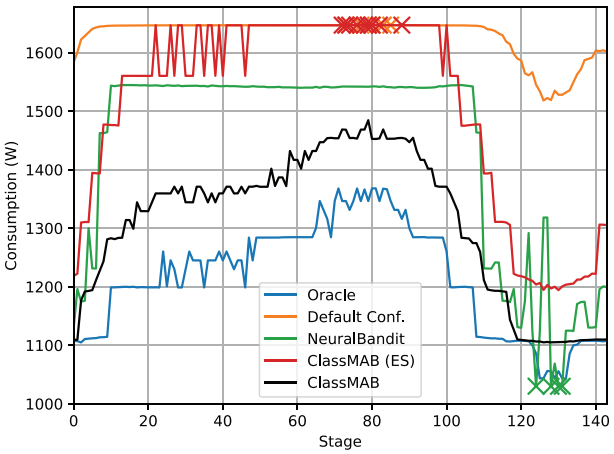

Fig. 7. Energy consumption during the test phase. The stages where the selected control does not satisfy the QoS are marked with crosses.

	Energy savings (%)	Ratio of QoS fulfillment
Oracle	25.48%	1
Default conf.	0%	0.9583
NeuralBandit	12.00%	0.9722
ClassMAB (ES)	7.50%	0.9375
ClassMAB	19.60%	1

TABLE II
Summary of numerical results

References

[1] N. Bhushan *et al.*, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, pp. 82–89, February 2014.

[2] L. Li *et al.*, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*. ACM, April 2010, pp. 661–670.

[3] L. Saker *et al.*, "Optimal control of wake up mechanisms of femtocells in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 30, pp. 664–672, April 2012.

[4] X. Chen *et al.*, "Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, pp. 627–640, March 2015.

[5] O.-C. Iacoboaiea *et al.*, "SON Coordination in Heterogeneous Networks: A Reinforcement Learning Framework." *IEEE Trans. Wireless Commun.*, vol. 15, pp. 5835–5847, May 2016.

[6] M. Simsek, M. Bennis, and I. Güvenç, "Learning based frequency-and time-domain inter-cell interference coordination in hetnets," *IEEE Trans. Veh. Technol.*, vol. 64, pp. 4589–4602, October 2015.

[7] S. Deb *et al.*, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets," *IEEE/ACM Trans. Netw.*, vol. 22, pp. 137–150, February 2014.

[8] W. Wang *et al.*, "A survey on applications of model-free strategy learning in cognitive wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 18, pp. 1717–1757, March 2016.

[9] 3GPP TR 36.814, "Evolved Universal Terrestrial Radio Access (EUTRA); Further advancements for E-UTRA Physical Layer Aspects," 3rd Generation Partnership Project (3GPP), Tech. Rep., 2010.

[10] 3GPP R1-100142, "System performance of heterogeneous networks with range expansion," 3rd Generation Partnership Project (3GPP), Tech. Rep., 2010.

[11] 3GPP TR 36.887, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Study on energy saving enhancement for E-UTRAN (Release 12)," 3rd Generation Partnership Project (3GPP), Tech. Rep., 2014.

[12] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, December 2014.

[13] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, pp. 235–256, 2002.

[14] R. Allesiardo, R. Féraud, and D. Bouneffouf, "A neural networks committee for the contextual bandit problem," in *International Conference on Neural Information Processing*. Springer, November 2014, pp. 374–381.