A CALIBRATED LEARNING APPROACH TO DISTRIBUTED POWER ALLOCATION IN SMALL CELL NETWORKS

Xinruo Zhang† *Mohammad Reza Nakhai*⋆ *Gan Zheng*† *Sangarapillai Lambotharan*† *Björn Ottersten*‡

† Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, UK
⋆ Department of Informatics, Centre for Telecommunications Research, King's College London, UK
‡Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg

ABSTRACT

This paper studies the problem of max-min fairness power allocation in distributed small cell networks operated under the same frequency bandwidth. We introduce a calibrated learning enhanced time division multiple access scheme to optimize the transmit power decisions at the small base stations (SBSs) and achieve max-min user fairness in the long run. Provided that the SBSs are autonomous decision makers, the aim of the proposed algorithm is to allow SBSs to gradually improve their forecast of the possible transmit power levels of the other SBSs and react with the best response based on the predicted results at individual time slots. Simulation results validate that in terms of achieving max-min signal-to-interference-plus-noise ratio, the proposed distributed design outperforms two benchmark schemes and achieves a similar performance as compared to the optimal centralized design.

***Index Terms*—** small cell networks, distributed power allocation, online learning, calibration

1. INTRODUCTION

Cellular networks have experienced an explosive increase in the number of mobile subscribers in the last decade, and the global mobile data traffic is predicted to reach 48.3 Exabytes per month by 2021 [1]. Hyper-dense deployment of small cell base stations (SBSs) underlaying the existing macrocell cellular networks is considered a promising technique to meet the requirements of the mounting growth of mobile data traffic [2]. However, such network densification in cellular networks with limited licensed spectrum will result in increasing intercell interference (ICI). Cloud radio access networks that allow centralized radio resource coordination across multiple cells, have been widely investigated in the literature for ICI management and resource allocation [3]. However, the requirement of high-capacity fronthaul links to support the immense traffic demands in such a centralized implementation necessitates the distributed operation of SBSs that can be organized and coordinated autonomously. Conventional literature either models the distributed resource allocation problem as a noncooperative game and solves it via iterative processes [4–7], or uses iterative inter-base station (BS) fronthaul information exchange such as the subgradient method in [8–10] and the alternating direction method of multipliers technique in [10, 11] to schedule transmit power and manage ICI among BSs. These designs, nevertheless, assume that the channel remains invariant until the iterations are completed or the convergence is achieved, which is not very practical. Furthermore, some types of noncooperative games, e.g., the Stackelberg game in [6], require the players to alternately make moves, which may not be suitable for the practical scenario where BSs perform simultaneous transmissions.

This paper focuses on the design of an intelligent distributed power allocation mechanism in small cell networks operated under the same spectrum that maximizes the average minimum signal-to-interference-plus-noise ratio (SINR) in the long run. A calibrated learning based distributed multi-armed bandit (MAB) approach is developed for the individual SBSs to learn during the operation, i.e., find a trade-off between exploring the environment and exploiting the current knowledge of the environment. In contrast to the existing distributed power allocation designs that iteratively exchange side information among SBSs within a channel coherence time, this paper considers a more realistic scenario, where SBSs serve their own users simultaneously and only exchange the past power information at the end of each time instance. The proposed algorithm is designed for the SBSs to progressively and distributively calibrate their anticipations of the possible transmit power levels of the other SBSs, and react with the best response to the predicted results.

2. SYSTEM MODEL AND PROBLEM FORMULATION

This paper considers a distributed downlink small cell network consists of N_b SBSs, indexed by $\mathcal{L}_b = \{1, \cdots, N_b\}$, that simultaneously serve their own N_u users, indexed by $\mathcal{L}_u = \{1, \cdots, N_u\}$, over a shared spectrum. Let the time horizon T be divided into discrete time slots and indexed as $\mathcal{T} = \{1, \cdots, T\}$. Each time slot corresponds to a channel coherence time and the channel is assumed to vary across time slots but remains constant within each time slot. The time-division multiple access technology is employed to ensure user fairness via serving the users within the same small cell one after the other in their respective time slots. Each individual SBS is fully synchronized and has to make its decision, i.e., selects a transmit power level from an index set $\mathcal{A} = \{1, \cdots, A\}$ of discrete transmit power levels, independently, to serve its scheduled user during time slot t, $t \in \mathcal{T}$. Furthermore, the SBSs can communicate with each other via capacity-constrained inter-SBS communication links at the end of each time slot.

2.1. Downlink Transmission Model

Let the channel gain between SBS b, $b \in \mathcal{L}_b$ and its user u, $u \in \mathcal{L}_u$ at time slot t, $t \in \mathcal{T}$, be denoted as $\Psi_{bu}^{[t]}$. Also let $P_{bu}^{[t]} \in \mathcal{A}$ be the transmit power from SBS b to user u. The SINR for user u served by SBS b at time slot t is given by

$$\mathrm{SINR}_{bu}^{[t]} = \frac{P_{bu}^{[t]} |\Psi_{bu}^{[t]}|^2}{\sum\limits_{b' \in \mathcal{L}_b, b' \neq b} P_{b'u'}^{[t]} |\Psi_{b'u}^{[t]}|^2 + \sigma_u^2}, \tag{1}$$

where the numerator of (1) denotes the desired signal for the user, the terms in the denominator of (1) denote, respectively, the ICI caused by other SBSs and the additive white Gaussian noise with noise variance of σ_u^2.

2.2. Problem Formulation

Due to the fact that the individual SBSs are autonomous decision makers in the considered scenario, it is evident that the presence of interference, coupling with the decision of the individual SBSs, raises a conflict: the individual SBSs tries to transmit at its maximum possible transmit power to improve its utility, while the ICI incurred by doing so may degrade the performance of other SBSs. Thus, we formulate the power allocation for distributed small cell networks as a long-term max-min user fairness problem to optimize the power allocation decisions of the SBSs, as

$$\max_{\{P_{bu}^{[t]}\}} \left\{ \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \min_{b} \left(\mathrm{SINR}_{bu}^{[t]} \right) \right\}. \tag{2}$$

Since the system is operated in a distributed manner and the SBSs have to make decisions simultaneously, the key issue to be addressed for the individual SBSs to ensure the system-level performance is the acquisition or the reliable prediction of its opponents' decisions. In the next section, the ϵ-calibrated forecaster [12] will be introduced for each SBS to gradually improve its accuracy in predicting the opponents' decisions. It will be followed by the proposed calibrated learning algorithm that allows the individual SBSs to select the most appropriate transmit power level based on the output of the forecaster, such that the minimum SINR can be maximized in the long run.

3. CALIBRATED LEARNING ALGORITHM

3.1. ϵ-Calibrated Forecaster for the Opponents' Actions

Having a calibrated forecasting scheme not only allows the players to act with the best response to the predicted future events, but also enables multiple players to converge to a reasonable joint play in some cases [12]. The definition of ϵ-calibration can be explained as follows. Let us consider a forecaster b playing a game against the opponent b', where each SBS has a finite set $\mathcal{A}$ of possible outcomes. Define $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$ as the N_ϵ set of candidate probability values over $\mathcal{A}$. Suppose that at each time slot t, $t \in \mathcal{T}$, the forecaster b outputs its forecasts of the opponent b', given by one of the candidate probability values over each SBS's outcome, i.e., $\mathbf{p}_{bb'}^t = \{p_{bb',1}^t, \ldots, p_{bb',A}^t\} \in \mathcal{P}$ for SBS b at time t, whilst the opponent b' simultaneously selects an outcome $a_{b'}^t \in \mathcal{A}$. Let us denote by $\mathcal{K} = \{1, \ldots, N_\epsilon\}$ the indexes of $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$ and $\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}$ the center of N_ϵ balls with radius of ϵ, respectively. Given $\epsilon > 0$, the sequence of forecasts is called ϵ-calibrated if

$$\limsup_{T \to \infty} \sum_{k=1}^{N_\epsilon} \| \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}_{\{k_{bb'}^t = k\}} (\mathbf{p}_k - \boldsymbol{\delta}_{a_{b'}^t}) \| \leq \epsilon, \tag{3}$$

where $\boldsymbol{\delta}_{a_{b'}^t} \in \{0, 1\}$ is the dirac distribution on the opponent b''s outcome $a_{b'}^t \in \mathcal{A}$, and $\mathbf{1}_{k_{bb'}^t}$ is an indicator function that returns one if the prediction of the probability of power levels for a particular SBS at time t is $\mathbf{p}_k$ and zero otherwise. Following the similar procedure as in [12], let us define the target set C as a subset of the ϵ-ball around $(\mathbf{0}, \ldots, \mathbf{0})$ for the calibration norm, as

$$C = \{\underline{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_{N_\epsilon}) \in \mathbb{R}^{AN_\epsilon} : \sum_{i=1}^{N_\epsilon} \|\mathbf{x}_k\| \leq \epsilon, \mathbf{x}_k \in \mathbb{R}^A\} \tag{4}$$

and let us define the vector-valued pay-off function as

$$\mathbf{m}(k, a) = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{p}_k - \boldsymbol{\delta}_a, \mathbf{0}, \ldots, \mathbf{0}), k \in \mathcal{K}, a \in \mathcal{A}, \tag{5}$$

where $\mathbf{m}(k, a) \in \mathbb{R}^{AN_\epsilon}$ contains one non-zero vector element of $\mathbb{R}^A$ located at the k-th position given by the difference between the predicted probability value $\mathbf{p}_k$ and the dirac distribution of the opponent's true action $\boldsymbol{\delta}_a$, and $N_\epsilon - 1$ zero vector elements elsewhere. According to Blackwell's approachability theorem [13], the condition of ϵ-calibration in (3) is equivalent to the following statement: the closed convex set C is approachable by the vector-valued regrets, i.e.,

$$\overline{\mathbf{m}}_{bb'}^T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{m}(k_{bb'}^t, a_{b'}^t) = \frac{1}{T} (\sum_{t=1}^{T} \mathbf{1}_{\{k_{bb'}^t = 1\}} (\mathbf{p}_1 - \boldsymbol{\delta}_{a_{b'}^t}), \ldots, \sum_{t=1}^{T} \mathbf{1}_{\{k_{bb'}^t = N_\epsilon\}} (\mathbf{p}_{N_\epsilon} - \boldsymbol{\delta}_{a_{b'}^t})), \tag{6}$$

and there exists a prediction distribution $\varphi_{bb'}^t$ over $\mathcal{P}$, such that for $\forall a \in \mathcal{A}$,

$$(\overline{\mathbf{m}}_{bb'}^{t-1} - \prod_C(\overline{\mathbf{m}}_{bb'}^{t-1})) \cdot (\mathbf{m}(\varphi_{bb'}^t, a) - \prod_C(\overline{\mathbf{m}}_{bb'}^{t-1})) \leq 0, \quad (7)$$

where $\cdot$ denotes the inner product and $\prod_C(\overline{\mathbf{m}}_{bb'}^{t-1})$ is the projection of $\overline{\mathbf{m}}_{bb'}^{t-1}$ in l_2-norm onto C. Recall from Section 2 that we denote by $\mathcal{A} = \{1, \cdots, A\}$ the finite index set of discrete transmit power levels or actions of SBSs with cardinality of A in the considered scenario. Each SBS b acts as a forecaster, the other SBSs $\{b'\}_{b' \neq b, b' \in \mathcal{L}_b}$ correspond to the opponents and the actions of the opponents can be regarded as the other SBSs transmitting to their respective users at certain discrete power levels. The procedure of the ϵ-calibrated forecaster at each SBS b to predict each of the other SBSs' actions is summarized in Algorithm 1 below.

Algorithm 1 *ϵ-calibrated forecaster at SBS b*

Input: current time-slot t, ϵ^t, discount factor β, actual arm chosen by the opponent at time $t-1$, i.e., $\{a_{b'}^{t-1}\}$.
Update the average vector-valued regrets up to time $t-1$ as per (6), as $\overline{\mathbf{m}}_{bb'}^{t-1} = \frac{1}{t-1}\sum_{t'=1}^{t-1} \mathbf{m}(k_{bb'}^{t'}, a_{b'}^{t'})\beta^{(t-t')}$.
Calculate the projection $\prod_C(\overline{\mathbf{m}}_{bb'}^{t-1})$, by obtaining the optimal solution $\underline{\mathbf{x}}^* \in \mathbb{R}^{AN_\epsilon}$ of the following problem
$\min_{\underline{\mathbf{x}}} ||\underline{\mathbf{x}} - \overline{\mathbf{m}}_{bb'}^{t-1}||_2^2$
s.t. $x_k \geq 0,\ 1 \leq k \leq AN_\epsilon$ and $\sum_{k=1}^{AN_\epsilon} x_k \leq \epsilon^t$.
Optimize the prediction distributions $\{\varphi_{bb'}^t\}$, i.e.,
$\underset{\varphi_{bb'}^t}{\operatorname{argmin}} \max_{a \in \mathcal{A}} (\overline{\mathbf{m}}_{bb'}^{t-1} - \prod_C(\overline{\mathbf{m}}_{bb'}^{t-1})) \cdot \mathbf{m}(\varphi_{bb'}^t, a)$,
based on the multiplicative weights algorithm [14], such that for $\forall a \in \mathcal{A}$, (7) is satisfied.
Output the forecast result $\{\mathbf{p}_{bb'}^t\}$ according to the optimal prediction distribution $\{\varphi_{bb'}^t *\}$ over $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$.

3.2. Calibrated Learning Algorithm for Power Allocation

In order to maximize the long-term system-level performance over all SBSs, the impact of the uncertain factors such as wireless channel conditions need to be learned over time. Here we propose a distributed bandit approach to account for the impact and to avoid local optimum of the ϵ-calibrated forecaster. Let us consider an MAB problem that models a system of A arms whose expected rewards are i.i.d. over time with unknown means. The objective of an MAB problem is to maximize the accumulated reward over time through a trade-off between exploring the environment to find profitable actions, while exploiting current knowledge to make the empirically best decisions among a set of actions [15]. The max-min SINR problem investigated in this paper can be regarded as a distributed MAB problem, where each SBS acts as an agent and a forecaster, A transmit power levels correspond to A actions or arms, and the instantaneous reward of individual power levels chosen by SBS b can be defined as $\mathrm{SINR}_{bu}^{[t]}$. The details of the proposed calibrated learning

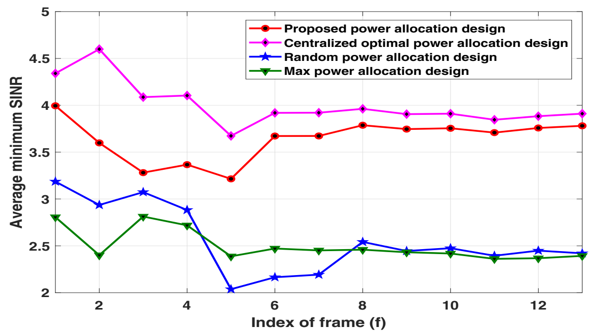
Algorithm 2 *Distributed Calibrated Learning Main Algorithm at SBS b*

Initialize: $t = 1$; total no. of time slots T, ϵ^t, $\gamma_{t=1} = 1$.
While $t \neq T$ **do**
with probability of γ_t: **Exploration Stage**
-The SBS transmits to its user at a random power level.
with probability of $1 - \gamma_t$: **Exploitation Stage** :
-Receive the information of actual arm chosen by the opponents, i.e., $\{a_{b'}^{t-1}\}$, and the average mean rewards, i.e., $\bar{\boldsymbol{\mu}}_{b'b}^{[t-1]} = \{\bar{\mu}_{b'b,1}, \ldots, \bar{\mu}_{b'b,A}\},\ b' \neq b, b' \in \mathcal{L}_b$ of previous time slot from the other SBSs.
-Receive predictions of other SBSs for the current time slot, i.e., $\{\mathbf{p}_{bb'}^t = \{p_{bb',a'}^t\}_{a' \in \mathcal{A}}\}$, from Algorithm 1.
-Calculate the estimated mean reward as
$\hat{\boldsymbol{\mu}}_b^{[t]} = \{\sum_{a'=1}^{A} p_{bb',a'}^t \mathrm{SINR}_{bu}^{[t]}(a)\}_{a \in \mathcal{A}}$.
-Transmit at a power level associated with the highest max-min reward, as $k_{bb'}^t = \underset{a_b^t \in \mathcal{A}}{\operatorname{argmax}} \min(\hat{\boldsymbol{\mu}}_b^{[t]}, \{\bar{\boldsymbol{\mu}}_{b'b}^{[t-1]}\})$.
Observe the associated true reward $\mathrm{SINR}_{bu}^{[t]}$.
Average $\bar{\boldsymbol{\mu}}_{bb'}^{[t]}$ over past time slots, as
$\bar{\boldsymbol{\mu}}_{bb'}^{[t]} = \frac{\sum_{t'=1}^{t-1} \bar{\boldsymbol{\mu}}_{bb'}^{[t']}\beta^{(t-t')} + \mathbf{1}_{k_{bb'}^t}\mathrm{SINR}_{bu}^{[t]}}{t}, \forall b' \neq b, b' \in \mathcal{L}_b$.
increment the time slot count $t = t + 1$.
end while

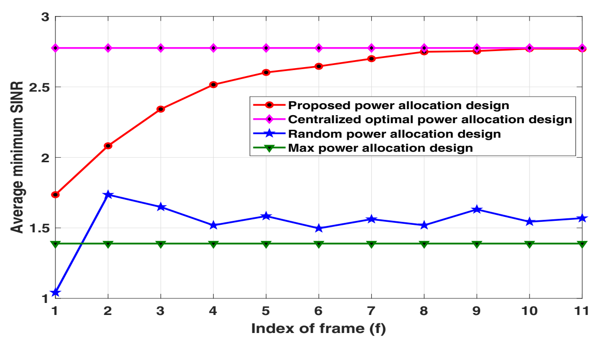
algorithm to be executed at the individual SBSs are described in Algorithm 2. Note that the only information an SBS needs to share with other SBSs at the end of time slot t for the proposed design is its actual chosen arm, i.e., a_b^t, and a vector of its average mean reward, i.e., $\bar{\boldsymbol{\mu}}_{bb'}^{[t]}$, which is a light overhead as compared to that of the conventional iterative distributed power allocation designs [8–11]. Furthermore, the practical scenario with a large number of SBSs and users does not affect the scalability of the proposed algorithm, as it may only decelerate the convergence speed and/or increase the computational burden.

4. SIMULATION RESULTS

Consider a distributed small cell network consisting 7 SBSs, $N_u = 2$ users are randomly scheduled in each small cell. Here we only consider the top $N_b = 2$ neighboring interferers among all the SBSs to account for the worst-case ICI. Each SBS can transmit at $\{1, 2\}$ W discrete transmit power levels. The channel gain $\Psi_{bu}^{[t]}$ is scaled by $G_a L_{bu} \sigma_F^2 e^{-0.5\frac{(\sigma_s \ln 10)^2}{100}}$ [8], where $L_{bu} = 125.2 + 36.3\log_{10}(d)$ denotes the path loss model over a distance of d km between SBS b and its user u, and the antenna gain is $G_a = 15$ dBi. The candidate probability values are set to be $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\} = \{(0,1), \ldots, (1,0)\}$ with $N_\epsilon = 21$ over $A = 2$ actions. The performance of all power allocation designs is evaluated versus frames, where each frame contains 100 time slots and the performance at each frame is averaged over the current frame and the one before. In order to demonstrate the advantages of our proposed power allocation design, a random power allocation design that randomly selects transmit power level from $\mathcal{A}$ at individual time slots as well as a max power allocation design that always chooses the maximum transmit power level, have been set as the benchmark schemes. Furthermore, a centralized optimal max-min SINR design that exhaustively searches for the optimal transmit power levels for SBSs at individual time slots, is employed as a performance upper bound.



(a)



(b)

Fig. 1. Min SINR of various power allocation designs for (a) rapid channel variations, (b) near-static channel conditions.

Fig. 1 compares the average minimum SINR for various power allocation designs for highly uncertain wireless channel variations, i.e., Fig. 1(a), and near-static channel conditions, i.e., Fig. 1(b), respectively. As can be seen from Fig. 1, the proposed design progressively converges to the centralized design for the case of static environment and has a slightly degraded performance when the channel varies rapidly, whilst the performance gap between the proposed design and the optimal centralized design is narrowed with increasing number of frames. In addition, the proposed design outperforms two benchmark schemes in both cases. This is due to the fact that in the benchmark schemes, the individual SBS autonomously selects its transmit power level without taking into account the impact it may have on its counterparts as well as the variations in wireless channel conditions.

Table 1. Reward table for SBSs at the final time slot

SBS1\ SBS2	Arm1	Arm2
Arm1	2.4286/ 9.6637	1.2146/ 20.3274
Arm2	4.8572/ 6.8347	2.4291/ 13.6694

Table I provides the possible SINR values that could be achieved for various combinations of arms chosen by the two SBSs at the final time slot. One may observe from the table that the maximization of the minimum SINR of 4.8572 is achieved when SBS1 and SBS2 transmit, respectively, at power level 2 and power level 1. Fig. 2 presents the individual SBS's prediction distribution, i.e., $\varphi^t_{bb'}$, of candidate probability values $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$ over its opponent's arm chosen at the final time slot. It is obvious that SBS1 and SBS2 are highly likely to be expected by their opponents to choose arm 2 and arm 1, respectively. This conclusion is in agreement with Table I, where SBS1 choosing arm 2 whilst SBS2 selecting arm 1 leads to the maximization of minimum SINR.

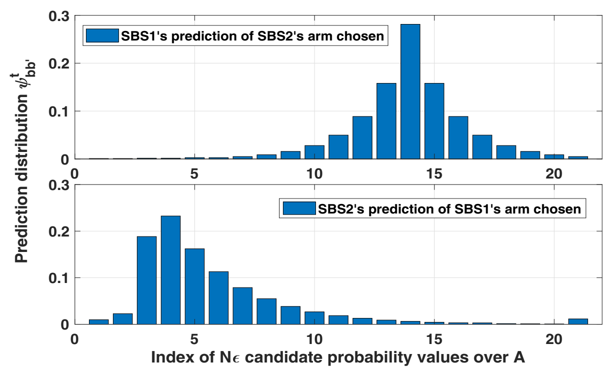


Fig. 2. Prediction distribution $\varphi^t_{bb'}$ of N_ϵ candidate probability values at final time slot.

5. CONCLUSION

This paper proposes a calibrated learning based power allocation algorithm in distributed small cell networks, which adapts to the actions of the SBSs and achieves max-min user fairness in the long run. The proposed design allows the SBSs to gradually improve their prediction on the opponents' behaviour and react with the best response based on the forecast results at individual time slots. Simulation results validate that the proposed distributed design outperforms two benchmark schemes and closely follows the optimal centralized design with limited amount of side information exchange.

6. REFERENCES

[1] Cisco, "Cisco visual networking index: Forecast and methodology, 2016-2021," *White Paper*, Sept 2017.

[2] F. Al-Turjman, E. Ever, and H. Zahmatkesh, "Small cells in the forthcoming 5g/iot: Traffic modelling and deployment overview," in *Proceedings Title*. IEEE, Aug 2018.

[3] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wubben, "Cloud technologies for flexible 5g radio access networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, May 2014.

[4] Z. Han, D. Niyato, W. Saad, T. Baar, and A. Hjrungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*, Cambridge University Press, New York, USA, 2012.

[5] Z. Han and K. J. R. Liu, "Noncooperative power-control game and throughput game over wireless networks," *IEEE Transactions on Communications*, vol. 53, no. 10, pp. 1625–1629, Oct. 2005.

[6] Z. Wang, B. Hu, X. Wang, and S. Chen, "Interference pricing in 5g ultra-dense small cell networks: a stackelberg game approach," *IET Communications*, vol. 10, no. 15, pp. 1865–1872, Oct. 2016.

[7] G. Bacci, E. V. Belmega, P. Mertikopoulos, and L. Sanguinetti, "Energy-aware competitive power allocation for heterogeneous networks under qos constraints," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 4728–4742, 2015.

[8] Z. Xiang, M. Tao, and X. Wang, "Coordinated multicast beamforming in multicell networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 12–21, Jan. 2013.

[9] X. Zhang and M. R. Nakhai, "Robust chance-constrained distributed beamforming for multicell interference networks," *IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2016.

[10] O. Tervo, H. Pennanen, D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Distributed optimization for coordinated beamforming in multicell multigroup multicast systems power minimization and sinr balancing," *IEEE Transactions on Signal Processing*, vol. 66, no. 1, pp. 171–185, Jan. 2018.

[11] C. Shen, T. H. Chang, K. Y. Wang, Z. Qiu, and C. Y. Chi, "Distributed robust multicell coordinated beamforming with imperfect csi: An admm approach," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2988–3003, Feb. 2012.

[12] S. Mannor and G. Stoltz, "A geometric proof of calibration," *Mathematics of Operations Research*, vol. 35, no. 4, pp. 721–727, Nov. 2010.

[13] D. Blackwell, "An analog of the minimax theorem for vector payoffs," *Pacific Journal of Mathematics*, vol. 6, no. 1, pp. 1–8, 1956.

[14] Y. Freund and R. E. Schapire, "Adaptive game playing using multiplicative weights," *Games and Economic Behavior*, vol. 29, pp. 79–103, 1999.

[15] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multi-armed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, May 2002.