Algorithms for Base Station Assignment in Heterogeneous Cellular Networks

by

Kaiming Shen

A thesis submitted in conformity with the requirements for the degree of Master of Applied Science
Graduate Department of Electrical and Computer Engineering
University of Toronto

Copyright © 2013 by Kaiming Shen
Abstract

Algorithms for Base Station Assignment in Heterogeneous Cellular Networks

Kaiming Shen
Master of Applied Science
Graduate Department of Electrical and Computer Engineering
University of Toronto
2013

This thesis aims to find out the base station (BS) assignment strategy which achieves load balancing in heterogeneous networks from the proportional fairness perspective. There are mainly three contributions in this thesis. First, for the BS assignment problem under fixed transmit power, a dual pricing approach called the dual coordinate descent is proposed, which achieves near-optimal solution and is suitable for distributed implementation. Further, this method is modified to apply to the orthogonal frequency-division multiplexing (OFDM) case. Second, for the joint BS assignment and power control problem, this thesis offers the outside incorporation method. This method is shown to be more practical than another novel approach in this thesis. Third, for the joint BS assignment and beamforming problem, this thesis proposes a two-stage method. This approach has comparable performance and low computational complexity as compared to the algorithm in the existing literature.
Acknowledgements

I am indebted to Professor W. Yu, who supervised the thesis, for his advice and encouragement.

I would like to thank Dr. H. Dahrouj, with whom I have had the good fortune to collaborate. I also want to thank G. Sridharan, Y. Lin, Y. Zhou, B. Dai, A. Ayoughi, P. Patil for many helpful discussions. I have a deep gratitude towards all my colleagues and friends at U of T.

Finally, I would like to thank my parents for their supports.
# Contents

1 Introduction 1
   1.1 General Introduction to Assignment Problem 1
   1.2 Motivation 3
   1.3 Overview of the Thesis 5
   1.4 Notations 6

2 Base Station Assignment 7
   2.1 Problem Formulation 8
   2.2 Some Heuristics 10
   2.3 Lagrangian Dual Analysis 11
   2.4 Dual Pricing Methods 13
      2.4.1 Subgradient Method 13
      2.4.2 Dual Coordinate Descent 16
      2.4.3 Low-cost Dual Coordinate Descent 18
   2.5 Duality Gap 21
   2.6 OFDM Case 22
      2.6.1 Averaging Dual Coordinate Descent 23
      2.6.2 Weighted-rate Relaxation Method 25
   2.7 Numerical Evaluation 28
3 Joint Base Station Assignment & Power Control

3.1 Problem Formulation ............................................. 34
3.2 Power Control .................................................... 35
3.3 Max-SINR Rule plus Power Control .............................. 36
3.4 Outside Incorporation Method .................................... 37
3.5 Inside Incorporation Method ..................................... 38
3.6 Numerical Evaluation ............................................. 43

4 Joint Base Station Assignment & Beamforming .................. 45

4.1 Problem Formulation ............................................. 46
4.2 Beamforming Problem ........................................... 47
4.3 Approximate WMMSE Method .................................... 49
4.4 Two-stage Method ................................................ 52
    4.4.1 Stage One: BS Assignment in MIMO Network .......... 52
    4.4.2 Stage Two: Beamforming Update ......................... 53
4.5 Complexity Analysis ............................................. 55
4.6 Numerical Evaluation ............................................. 55

5 Conclusions ......................................................... 59

Bibliography ............................................................ 61
List of Tables

1.1 Payment information .................................................. 2

2.1 Information of $a_{ij}$ in the two-BS model .......................... 11

2.2 System parameters ..................................................... 29

2.3 Multi-path environment which generates frequency-selectiveness .... 31
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A heterogeneous network example</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>Oscillatory behavior of greedy method</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Two derivatives of ( f ) at the same position</td>
<td>14</td>
</tr>
<tr>
<td>2.3</td>
<td>Two cases of updating ( \mu_j^* ) in dual coordinate descent</td>
<td>17</td>
</tr>
<tr>
<td>2.4</td>
<td>Updating ( \mu_j ) to ( \mu_j^* )</td>
<td>19</td>
</tr>
<tr>
<td>2.5</td>
<td>Information exchange for the price update of BS A</td>
<td>19</td>
</tr>
<tr>
<td>2.6</td>
<td>An example of dual coordinate descent getting stuck at a suboptimal position</td>
<td>21</td>
</tr>
<tr>
<td>2.7</td>
<td>A wrap-around HetNet topology</td>
<td>30</td>
</tr>
<tr>
<td>2.8</td>
<td>Dual error convergence: the dual coordinate descent vs the subgradient method</td>
<td>30</td>
</tr>
<tr>
<td>2.9</td>
<td>User rate CDF of the various BS assignment methods after 56 iterations</td>
<td>31</td>
</tr>
<tr>
<td>2.10</td>
<td>User rate CDF of the different BS assignment algorithms in OFDM environment</td>
<td>32</td>
</tr>
<tr>
<td>3.1</td>
<td>The initial BS assignment situation</td>
<td>37</td>
</tr>
<tr>
<td>3.2</td>
<td>An example of Claim 5</td>
<td>40</td>
</tr>
<tr>
<td>3.3</td>
<td>Data rate CDF of the various algorithms for the joint BS assignment and power control problem</td>
<td>43</td>
</tr>
<tr>
<td>4.1</td>
<td>Beamforming updates</td>
<td>51</td>
</tr>
<tr>
<td>4.2</td>
<td>Flow chart of the two-stage method</td>
<td>54</td>
</tr>
</tbody>
</table>
4.3 User rate CDF: the two-stage method vs the max-SINR BS rule plus the WMMSE algorithm .......................... 57
4.4 A small HetNet with 3 macro BSs, 3 pico BSs, and 60 users .......... 57
4.5 User rate CDF: the two-stage method vs the approximate WMMSE method 58
Chapter 1

Introduction

1.1 General Introduction to Assignment Problem

Loosely speaking, the assignment problem is about finding out an object-assigning strategy that maximizes (or minimizes) the predefined objective function.

One of the earliest assignment problems may be traced back to a Chinese anecdote. Sun Bin, a famous military strategist in ancient times, once took part in a horse-racing competition with a king. Each side had three horses assigned to three rounds of races. According to the ability, the three horses for each side were divided into the superior, the regular and the inferior level. The situation was, each horse of the king ran faster than Sun’s horse of the same ability level, but not as fast as Sun’s higher level horse if any. Sun came up with this clever strategy: his inferior horse was assigned to compete with the king’s superior, and then his superior was assigned to compete with the king’s regular, and at last his regular is assigned to compete with the king’s inferior. Sun therefore won the competition by the final score of 2 to 1, even though his opponent had better horses on the whole.

Another classic example of assignment problem is the following task assignment problem. Three tasks (which are denoted by A, B and C respectively) need to be assigned to
three employees: Jack, Tom and Rose. With the payment information shown in Table 1.1, the problem objective is to find the task assignment scheme that costs the least. Note that in this case the number of tasks and the number of employees are the same, hence this problem scenario belongs to the 1-to-1 assignment, to which some classic algorithms apply, e.g., auction algorithm [2] and Hungarian algorithm [3].

Albeit there are a variety of assignment problems, two common features of them can be found. First, since the assignment strategy is mathematically expressed by a series of binary variables (0 or 1), all these assignment problems can be categorized into the discrete 0-1 optimization class. Second, the assignment of each object adds a gain value to the objective function, while the target is to maximize (or minimize) the objective function (i.e., the total gain). On this point, if the gain values are predefined like in the above task assignment problem (e.g., task A being assigned to Jack has the gain value of 30), some existing methods (e.g., Hungarian algorithm) can be adopted to achieve the optimal solution. In contrast, if the gain values are not fixed beforehand, and are impacted by how other objects are assigned, the assignment problem is usually difficult to solve.
1.2 Motivation

Modern wireless networks adopt a cellular structure. Throughout all the cells, access nodes which are usually base stations (BSs), are deployed to provide connection services to user terminals. From the user terminal’s perspective, it needs to associate with one of these access nodes for communication. In some proposals, e.g., the coordinated multi-point (CoMP) technology in [4], a user terminal can even associate with more than one BS at the same time.

The user-BS association can be viewed from an assignment problem perspective. If user $i$ is associated with BS $j$, it can be claimed that user $i$ is assigned to BS $j$. It is therefore “user assignment” rather than “BS assignment”. However, in order to highlight the cellular networks, we still use the latter terminology in this thesis.

![A heterogeneous network example](image)

Figure 1.1: A heterogeneous network example

We look for the BS assignment strategy that benefits heterogeneous networks (HetNets) the most. HetNets have an emerging cellular architecture where access nodes of different types are deployed throughout the geographical area to off-load traffic from macro cells, as shown in Figure 1.1. By splitting the conventional cellular structure into small cells (e.g., femto/pico cells), HetNets architecture allows for more aggressive reuse of frequencies as well as improved coverage and higher overall throughput for the entire
network.

One of the main challenges in the design and implementation of HetNets is the appropriate settings of the transmit power levels of different BSs and the definition of each cell’s coverage area, and the beamforming vectors also need to be taken into account if the multiple-input and multiple-output (MIMO) technology is adopted in the network. Conventionally, the downlink coverage areas of cells are defined according to the received signal-to-interference-plus-noise ratio (SINR). For each user terminal, it simply associates with the BS with the highest SINR. This scheme of BS assignment is often referred to as the max-SINR rule. A key problem with the max-SINR BS assignment is that it does not account for the varied data traffic pattern in HetNets, hence it fails to account for load balancing effectively. Load balancing is essential for a small-cell environment, because femto/pico BSs are often deployed to alleviate traffic “hot-spots” with higher-than-average user density. In these situations, a heuristic that adds a bias term to the reference SINR [5] often needs to be used in practice to address the load imbalance.

There are two mainstream formulations of the BS assignment problem. The first one is power-based for the purpose of interference minimization. Its problem objective is to minimize the total transmit power under the predefined minimum SINR constraint for each user. This power-based BS assignment problem has been extensively analyzed in [6–12]. The other common formulation is rate-based. Specifically, the gain value of each user-BS association is given by a function of user rate, and the objective is to maximize the total gain. It is popular to interpret such a gain value as the utility which is a concept in microeconomics [13], and accordingly that gain function of user rate is referred to as the utility function. The utility function should satisfy the law of diminishing marginal utility [13]: 1) being nondecreasing; 2) being convex. Among all utility functions, the log utility function of the $\alpha$-fairness family [14] is a notable one, whose optimization result is known as proportional fairness. The log-utility maximization target is pursued throughout this thesis.
Chapter 1. Introduction

The BS assignment problem differs from the formerly introduced assignment problems in two aspects. First, due to the fact that a BS usually serves more than one user terminal at the same time, the BS assignment is actually an $N$-to-1 ($N > 1$) assignment. Second, the data rate of a user is impacted by how other users are assigned because of the resource allocation in the cell, and therefore its utility value is unknown until all users in the network have been assigned.

In addition to the BS assignment, two other variables of the network model, which are the transmit power levels and the beamforming vectors, play an important role in the network optimization. These three variables interact with each other, and we look for a joint optimization of them.

1.3 Overview of the Thesis

The ultimate objective in this thesis is to jointly optimize the BS assignment, transmit power levels, and beamforming vectors for the downlink cellular network. This overall problem is decomposed into three following tasks.

First, we study the BS assignment problem under fixed transmit power in a single-input and single-output (SISO) network. Our exploration starts from a simplified network with flat-fading channels. The BS assignment problem is analyzed in the dual domain, where the dual variables have a pricing interpretation. We propose the dual coordinate descent algorithm to update the dual variables for the BS assignment. As compared to the existing method in the literature, our proposed method is shown to achieve near-optimal solution, and has a structure suitable for distributed implementation. A variant of the dual coordinate descent is further developed to reduce the information exchange required by the algorithm. In addition, we extend the dual coordinate descent to the OFDM case.

The second task is to solve the joint BS assignment and power control problem. We
first use Newton’s method to optimize the transmit power when the BS assignment is fixed, and then propose two different methods of combining this power control method with the dual coordinate descent, which are: 1) the outside incorporation method (which implements the power control algorithm and the dual coordinate descent iteratively); 2) the inside incorporation method (which incorporates the power update into the dual coordinate descent based on duality theory). By comparison, the outside incorporation method is more practical because it guarantees the convergence and has higher efficiency.

Finally, we consider the BS assignment problem jointly with beamforming (power control is included in beamforming) for multi-antenna networks. In the existing literature, the WMMSE algorithm [1] is an effective approach to the beamforming problem when the BS assignment is fixed. We then propose a two-stage method which combines the WMMSE algorithm and the previously proposed outside incorporation method. As compared to the existing approach for the joint BS assignment and beamforming problem — the approximate WMMSE method [15], our proposed method leads to comparable performance and has a much lower computational complexity.

1.4 Notations

Throughout this thesis, bold letters are used to denote matrices or vectors (upper case for matrices and lower case for vectors). For sets, \( \mathbb{Z} \) denotes the set of integers; \( \mathbb{R} \) denotes the set of real numbers; \( \mathbb{C} \) denotes the set of complex numbers; \(|\cdot|\) denotes the cardinality operation. For matrices or vectors, \( \|\cdot\| \) denotes the Euclidean norm operation; \(|\cdot|\) denotes the determinant operation; \((\cdot)^T\) denotes the transpose operation; \((\cdot)^H\) denotes the conjugate transpose operation; \( \mathbf{I}_{n \times n} \) denotes the \( n \)-dimensional identity matrix.
Chapter 2

Base Station Assignment

This chapter considers the BS assignment problem in a downlink SISO network model with fixed transmit power. We want the BS assignment to achieve load balancing so that the traffic “hot-spots” with higher-than-average user density can be alleviated in macro cells. The conventional method for the BS assignment is the max-SINR rule, i.e., each user terminal chooses the BS with the highest SINR. However, this rule does not account for load balancing, because the low-power pico/femto BSs can hardly attract users under the max-SINR rule so as to off-load the traffic in macro cells.

We address the BS assignment problem in the dual domain, using a pricing approach to find the optimal dual variables. Our main contribution is a distributed dual pricing method based on the coordinate descent, which is guaranteed to converge and is shown to achieve near-optimal performance in practice. Some contributions of this chapter have been published in [16].

The BS assignment problem has been considered extensively in the existing literature. For example, in order to off-load the traffic in macro cells, [5] expands the coverage area of small cells by adding constant bias terms to the SINR values, but it does not analyze what is the optimal bias term. Some other methods come from intuition, like the probabilistic method in [17] and the greedy method in [18–20]. The idea in [21] is to solve the BS
assignment problem using game theory, since BS assignment is a kind of decision making. The Nash equilibrium of the problem is found in that work. Some special BS assignment scenarios are also studied in the literature. For instance, [22] considers a simple model of only a single pair of macro and pico BSs; [23] considers the situation that user terminals may not report their channel state information (CSI) truthfully out of selfish reasons.

This chapter is organized as follows. Section 2.1 formulates the BS assignment problem for a downlink SISO network with flat-fading channels. This problem model is analyzed in the following Section 2.2 to 2.5. Section 2.2 introduces some common heuristics for the BS assignment problem. Section 2.3 analyzes the problem in the dual domain, and gives a pricing interpretation for the dual variables. Section 2.4 presents three pricing approaches based on the dual analysis results, two of which are innovations in this thesis. Section 2.5 studies the duality gap which gives the error bounds of these three pricing approaches. Section 2.6 extends the BS assignment problem to the OFDM case, for which two original methods are offered. Finally, we present the numerical results in Section 2.7.

2.1 Problem Formulation

Consider the downlink SISO network consisting of $L$ BSs with fixed transmit power levels (which may differ from BSs to BSs), and $K$ active user terminals across the geographic area covered by the network. Let $i$ index each user terminal (i.e., $i \in \{1, 2, \ldots, K\}$), and let $j$ index each BS (i.e., $j \in \{1, 2, \ldots, L\}$). The total frequency bandwidth in this cellular system is $W$, used by all BSs with reuse factor of one. To simplify the problem, we assume flat-fading channels and frequency-flat power spectral densities (PSDs) for the model, thus the SINR values are constants across the bandwidth. Let $h_{ij} \in \mathbb{C}$ be the channel between user $i$ and BS $j$, and let $p_j$ be the transmit PSD at BS $j$. The SINR
value can be calculated by
\[ \text{SINR}_{ij} = \frac{|h_{ij}|^2 p_j}{\sum_{j' \neq j} |h_{ij'}|^2 p_{j'} + \sigma_i^2} \] (2.1)
where \( \sigma_i^2 \) is the PSD of the additive white Gaussian noise (AWGN) of user \( i \).

As mentioned earlier, we adopt the sum log-utility maximization objective to realize load balancing in the network. It is shown in [24] that Round-robin scheduling is the optimal scheme for frequency resource scheduling if channels and transmit PSDs are both frequency-flat. Hence, if a total of \( k_j \) users are associated with BS \( j \), each of them will share \( 1/k_j \) of the frequency resource. For user \( i \) who is associated with BS \( j \), its rate is calculated by
\[ R_{ij} = \frac{W}{k_j} \log \left( 1 + \frac{\text{SINR}_{ij}}{\Gamma} \right) \] (2.2)
where \( \Gamma \) is the SNR gap of user \( i \), which is determined by the coding and modulation scheme. Without loss of generality, \( \Gamma \) is assumed to be the same for all the users.

Let \( x_{ij} \) be a binary variable (1 or 0) denoting whether or not user \( i \) is associated with BS \( j \). The objective function of the BS assignment can be written as:
\[ f_o(X, R) = \sum_{i,j} x_{ij} \log (R_{ij}) \] (2.3)
Furthermore, we introduce parameter \( a_{ij} \) as:
\[ a_{ij} = \log \left( \frac{W}{k_j} \log \left( 1 + \frac{\text{SINR}_{ij}}{\Gamma} \right) \right) \] (2.4)
Substituting \( a_{ij} \) back into (2.3), we can formulate the BS assignment problem as
\[
\begin{align*}
\text{maximize} & \quad \sum_{i,j} a_{ij} x_{ij} - \sum_j k_j \log (k_j) \\
\text{subject to} & \quad \sum_j x_{ij} = 1, \quad \forall i \\
& \quad \sum_i x_{ij} = k_j, \quad \forall j \\
& \quad \sum_j k_j = K \\
& \quad x_{ij} \in \{0, 1\}, \quad \forall i, \forall j
\end{align*}
\] (2.5a)
where constraint (2.5b) states that each user can only associate with one BS, and constraint (2.5d) states that all users in the network will be served.

2.2 Some Heuristics

The greedy method [18–20] is a simple way to improve the total utility. Its basic idea is to have one or a few user terminals switch their BS association each time to increase their own utility values. However, the performance of the greedy method is not easy to control. The convergence has slow speed if too few users are chosen to switch their BS association each time. On the contrary, the algorithm may exhibit an oscillatory behavior [19] if too many users switch BSs at the same time. The following instance illustrates how the oscillatory behavior happens. Consider the model displayed in Figure 2.1 (a) with the $a_{ij}$ given in Table 2.1 ($a_{ij}$ is defined in (2.4)). In this case, all the users can switch their BSs at the beginning of each time-slot. Initially all the users are associated with BS A as shown in Figure 2.1 (a), and in the next time-slot all of them will switch to BS B by the greedy method as shown in Figure 2.1 (b), but in the third time-slot all of them will return to BS A by the greedy method, and so forth.

![Figure 2.1: Oscillatory behavior of greedy method](image)

A probabilistic heuristic is proposed in [17] to decide the BS assignment. In this
approach, the BS assignment is determined in random fashion. The probability for a user to associate with BS $j$ is proportional to its estimated spectral efficiency with this BS. This method is quite intuitive, but is lacking in rigorous support.

Another heuristic idea is to relax the 0-1 discrete variable $x_{ij}$ to the real-number variable in the closed interval of $[0, 1]$. Therefore, constraint (2.5e) in problem (2.5) is converted into: $0 \leq x_{ij} \leq 1$, $\forall i, \forall j$. After this relaxation problem is solved, we round its solution to binary values which determine the BS assignment.

\section*{2.3 Lagrangian Dual Analysis}

In this section, we formulate the dual problem for the primal problem (2.5), and also show how the primal solution is recovered from the dual solution. One important idea in this part of work is that the dual variables can be interpreted as the BS-specific prices, which gives rise to the dual pricing approaches described in the next section.

Introduce dual variables $\mu = (\mu_1, \cdots, \mu_L)^T$ for constraint (2.5c), and $\nu$ for constraint (2.5d). The Lagrangian function with respect to these two constraints is

$$L(X, k, \mu, \nu) = \sum_{i,j} a_{ij} x_{ij} - \sum_j k_j \log(k_j) - \sum_j \mu_j \left(\sum_i x_{ij} - k_j\right) - \nu \left(\sum_j k_j - K\right).$$

(2.6)
The dual function $g(\cdot)$ can then be written as

$$
g(\mu, \nu) = \begin{cases} 
\max_{X,k} & L(X,k,\mu,\nu) \\
\text{s.t.} & \sum_j x_{ij} = 1, \ i = 1, \ldots, K \\
& x_{ij} \in \{0, 1\}, \ \forall i, \forall j
\end{cases} \tag{2.7}
$$

The above problem has the following explicit analytic solution:

$$
x_{ij}^* = \begin{cases} 
1, & \text{if } j = j^{(i)} \\
0, & \text{if } j \neq j^{(i)}
\end{cases} \quad \text{where } j^{(i)} = \arg \max_{j'} (a_{ij'} - \mu_{j'}) \tag{2.8}
$$

and

$$
k_j^* = e^{\mu_j - \nu - 1} \tag{2.9}
$$

Note that if $j^{(i)}$ in (2.8) is not unique, $x_{ij}$ can be assigned value 1 for any of the BSs with maximum $(a_{ij} - \mu_j)$ without affecting the value of dual function.

Note that the solution of $x_{ij}$ in (2.8) is quite intuitive. The dual variable $\mu_j$ is the price of BS $j$, while $a_{ij}$ is the utility of the user $i$ if it associates with BS $j$. Each user maximizes its utility $a_{ij}$ minus the price among all possible BSs, while the BSs choose their prices to balance their loads.

Substituting (2.8) and (2.9) back into (2.7), we obtain the closed form for the dual objective as:

$$
g(\mu, \nu) = \sum_i \max_j (a_{ij} - \mu_j) + \sum_j (e^{\mu_j - \nu - 1}) + \nu K. \tag{2.10}
$$

The dual problem is just to minimize $g(\cdot)$ over $\mu$ and $\nu$:

$$
\min_{\mu,\nu} g(\mu, \nu) \tag{2.11}
$$

One of the main contributions in this thesis is that the optimization of $g(\mu, \nu)$ can be done via coordinate descent. This approach is inspired by the development of auction algorithm in [2]. The BS assignment problem in this section can be thought of as a generalization of the 1-to-1 assignment problem solved by the auction algorithm [2] to an $N$-to-1 ($N \geq 1$) case.
Finally, we describe how to recover the primal variable $x_{ij}$ from the dual solution. This can be done through (2.8), but there are possibilities that a user has more than one BS with the same maximal value for $(a_{ij} - \mu_j)$. Resolving such ties may call for many common heuristics. In our simulations, only a very small number of users are typically involved in ties, so tie-breaking via exhaustive search is feasible.

## 2.4 Dual Pricing Methods

This section looks into the dual problem (2.11). Three different approaches are presented here, two of which are the innovations of this thesis.

### 2.4.1 Subgradient Method

The subgradient method is adopted in [24] to solve the dual problem for the BS assignment. In what follows, we first briefly introduce about the subgradient method, and then describe how it is used to solve (2.11).

For a convex function $f: \mathbb{R}^n \to \mathbb{R}$, we claim that $c \in \mathbb{R}^n$ is the subgradient of $f$ at $x \in \mathbb{R}^n$ if it satisfies the condition below:

$$f(y) \geq f(x) + c^T(y - x), \quad \forall y \in \mathbb{R}^n. \quad (2.12)$$

Note that more than one subgradient of $f$ may exist at $x$. For instance, Figure 2.2 displays two subgradients of $f(x)$ at $x = 8$ (which are the slopes of the arrows). To minimize the value of function $f$, the subgradient method updates variable $x$ by

$$x^{(t+1)} = x^{(t)} - \alpha^{(t)} c^{(t)}, \quad (2.13)$$

where $t$ is the iteration count; $c^{(t)}$ denotes the subgradient of $f$ at $x^{(t)}$; $\alpha^{(t)}$ denotes the step size for the update. Let $f^*$ be the globally minimal value of function $f$, and let $f_{\text{best}}^{(t)}$ be the minimal value of function $f$ in the past $t$ iterations. If $f \neq -\infty$, we have the
Figure 2.2: Two derivatives of $f$ at the same position

following conclusion as proved in [25]:

$$f^* - \lim_{t \to \infty} f^{(t)}_{\text{best}} < \epsilon.$$  \hspace{1cm} (2.14)

where $\epsilon$ is the error bound determined by the step size choice. Particularly, it is shown in [25] that $\epsilon$ approaches 0 if the step size meets some conditions, so the subgradient method can achieve the optimal solution.

We now use the subgradient method to solve problem (2.11). It is observed that if dual variable $\mu$ is fixed at $\mu^{(t)}$, $g(\cdot)$ is a differentiable convex function of $\nu$, and the optimal $\nu$ can be found as

$$\nu^{(t+1)} = \log \frac{\sum_j e^{\mu^{(t)}_j}}{K}$$  \hspace{1cm} (2.15)

However, $g(\cdot)$ is not a differentiable function of $\mu_j$, so instead of taking the gradient with respect to $\mu_j$, we use the subgradient to update $\mu$. Note that more than one subgradient may exist, and for simplicity we always take the left partial gradient as the
Chapter 2. Base Station Assignment

partial subgradient, which is

\[ e^{\mu_j - \nu - 1} - |\mathcal{U}_j|, \quad (2.16) \]

where \( \mathcal{U}_j \) is defined by

\[ \mathcal{U}_j = \left\{ i \mid a_{ij} - \mu_j = \max_{j'}(a_{ij'} - \mu_{j'}) \right\} \quad (2.17) \]

Then the subgradient method updates \( \mu \) in each step according to

\[ \mu_j^{(t+1)} = \mu_j^{(t)} - \alpha^{(t)} \left( e^{\mu_j^{(t)} - \nu^{(t)} - 1} - |\mathcal{U}_j| \right), \quad j = 1, \ldots, L \quad (2.18) \]

One problem with the subgradient method is that its convergence rate depends heavily on the choice of step size \( \alpha^{(t)} \). Possible choices of \( \alpha^{(t)} \) include a constant (but the constant is difficult to choose) or diminishing step sizes (which guarantee convergence but can be quite slow in practice). As a baseline for comparison, we adopt the self-adaptive scheme of [26] as suggested in [24]. In this scheme, the step size is calculated by

\[ \alpha^{(t)} = \gamma_t \frac{g(x^{(t)}) - g_t}{\|c^{(t)}\|}, \quad 0 < \gamma_t \leq \gamma \leq \overline{\gamma} < 2, \quad (2.19) \]

where \( \gamma \) and \( \overline{\gamma} \) are two predefined boundaries for the parameter \( \gamma_t \); \( g_t \) is the estimated value for \( g^* \), expressed as

\[ g_t = \min_{0 \leq t' \leq t} f(x^{(t')}) - \delta_t. \quad (2.20) \]

In (2.20), \( \delta_t \) is updated by

\[ \delta_{t+1} = \begin{cases} \rho \delta_t, & \text{if } g(x^{(t)}) \leq g_t; \\ \max\{\beta \delta_t, \delta\}, & \text{if } g(x^{(t)}) > g_t; \end{cases} \quad (2.21) \]

where \( \rho \geq 1, 0 < \beta < 1, \delta_1 > 0 \) and \( \delta > 0 \) are the predefined parameters in this self-adaptive subgradient method. It is shown in [26] that:

\[ \lim_{t \to \infty} g_{\text{best}}^{(t)} \geq g^* - \delta. \quad (2.22) \]

where \( g_{\text{best}}^{(t)} \) is the minimal value of function \( g \) in the past \( t \) iterations. Note that because all the \( \mu_j \)'s need to be updated at the same time using the same step size (in order
to ensure convergence), the distributed implementation of the subgradient method requires synchronized price updates across the BSs. The dual coordinate descent algorithm proposed in the next subsection removes such a requirement.

### 2.4.2 Dual Coordinate Descent

The main contribution of this chapter of the thesis is a dual coordinate descent algorithm for the price update. The basic idea of the proposed approach is to recognize that the dual function (2.10) is in a closed form, and thus it can be optimized in a coordinate descent fashion in closed form. Fixing all the $\mu_j$’s, we see again that the optimal $\nu$ can be updated by (2.15). Fixing $\nu$ and all $\mu_j$’s except one of them, we observe that $g(\cdot)$ is in fact the sum of a continuous piece-wise linear function and an exponential function. So we can take its left or right partial derivatives and choose $\mu_j$ to be such that the left partial derivative at $\mu_j$ is less than or equal to zero, and the right partial derivative is greater than or equal to zero. Mathematically, define functions

$$f_1(\mu_j) = |U_j|, \text{ where } U_j = \left\{ i \left| a_{ij} - \mu_j = \max_j a_{ij'} - \mu_{j'} \right. \right\}$$

and

$$f_2(\mu_j) = e^{\mu_j - \nu - 1}.$$  

(2.23) 

(2.24)

It is easy to see that the left partial derivative of $g(\cdot)$ with respect to $\mu_j$ is exactly $f_2(\mu_j) - f_1(\mu_j)$. Hence, fixing all other dual variables, the $\mu_j$ that minimizes $g(\cdot)$ is just

$$\mu_j^{(t+1)} = \sup \{ \mu_j | f_2(\mu_j) - f_1(\mu_j) \leq 0 \}$$

(2.25)

Figure 2.3 illustrates the price update condition, which seeks $\mu_j^*$ to match $f_1(\mu_j^*)$ and $f_2(\mu_j^*)$. Here, $f_1(\cdot)$ is a step function. The functions $f_1(\cdot)$ and $f_2(\cdot)$ may not intersect, but the optimal $\mu_j^*$ can always be determined uniquely. A complete description of dual coordinate descent algorithm is presented in Algorithm 1.

The dual coordinate descent algorithm is quite intuitive. The dual variable $\mu_j$ is the price of BS $j$, while $a_{ij}$ is the utility of the user $i$ if it associates with BS $j$. Each user
Chapter 2. Base Station Assignment

\[ \mu_j^* = f_1(\cdot) \cdot f_2(\cdot) \]

(a) \( f_1 \) and \( f_2 \) have intersection

(b) no intersection

Figure 2.3: Two cases of updating \( \mu_j^* \) in dual coordinate descent

Algorithm 1 Dual Coordinate Descent

Initialization: Set \( \mu_j = 0 \), \( \forall j \); set \( \nu = \log \frac{\sum_j e^{\mu_j-1}}{K} \);

repeat

\[ \text{for each } j \in \{1, \cdots, L\} \text{ do} \]

1) Update \( \mu_j \) according to (2.25);

\[ \text{end for} \]

2) Update \( \nu \) according to (2.15);

until \( (\mu, \nu) \) converges;

3) Set user-BS association according to (2.8). Resolve ties if necessary.
maximizes its utility minus the price among all possible BSs, while the BSs choose their prices in an iterative fashion to balance their loads.

As compared to the subgradient method, the main advantage of the dual coordinate descent algorithm is that BSs do not need to synchronize in their price updates. In fact, the order of price updates in the dual coordinate descent can be arbitrary. Since each dual update step always decreases the dual objective, the iterative algorithm is always guaranteed to converge.

### 2.4.3 Low-cost Dual Coordinate Descent

In this subsection, we develop a variant of the dual coordinate descent which requires less communication overhead than the dual coordinate descent does. In the dual coordinate descent, in order to construct function $f_1$ (e.g., Figure 2.4 (a)), the $a_{ij}$ information needs to be collected from all the users in the network (e.g., Figure 2.5 (a)). However, it is observed that only a part of $f_1$ is actually needed in the price update. When $\mu_j$ is updated for an overloaded cell (which satisfies $k_j > e^{\mu_j - \nu - 1}$), in order to raise the current $\mu_j$ to $\mu_j^*$, only the right-hand part of $f_1$ on the current $\mu_j$ needs to be known at the BS (e.g., Figure 2.4 (b)). To construct this part of $f_1$, the $a_{ij}$ information of the local users (those who are currently associated with BS $j$) is sufficient (e.g., Figure 2.5 (b)), and hence lots of information exchange on $a_{ij}$ is saved. Therefore, in the low-cost dual coordinate descent, we only implement the dual coordinate descent update for the overloaded cells. This approach is stated in Algorithm 2.
Chapter 2. Base Station Assignment

(a) the dual coordinate descent

(b) the low-cost dual coordinate descent

Figure 2.4: Updating $\mu_j$ to $\mu_j^*$

(a) the dual coordinate descent

(b) the low-cost dual coordinate descent

Figure 2.5: Information exchange for the price update of BS A
Chapter 2. Base Station Assignment

Algorithm 2 Low-cost Dual Coordinate Descent

**Initialization:** Set $\mu_j = 0$, $\forall j$; set $\nu = \log \frac{\sum_j e^{\mu_j - 1}}{K}$;

repeat

for each $j \in \{1, \cdots, L\}$ do

if $K_j > e^{\mu_j - \nu - 1}$ then

1) Update $\mu_j$ according to (2.25);

end if

end for

2) Update $\nu$ according to (2.15);

until $(\mu, \nu)$ converges;

3) Set user-BS association according to (2.8). Resolve ties if necessary.
2.5 Duality Gap

Because of the fact that $g(\cdot)$ is not differentiable everywhere, the coordinate descent algorithm is not guaranteed to converge to the optimal dual solution. Figure 2.6 shows an example: the displayed function cannot get increased at the corner position in either coordinate direction, and obviously the corner position is not the globally optimal point. Furthermore, because of the integer constraints, there may be a non-zero duality gap between the primal and the dual problems. For the BS assignment problem considered in this section, the duality gap can actually be characterized in closed form.

**Claim 1.** The difference between the objective function $f_o(X, R)$ optimized by the dual coordinate descent and the global optimum is bounded by $\sum_j k_j \log \left( k_j/e^{\mu_j-\nu-1} \right)$.

**Proof.** Let $(X, k)$ be the primal solution recovered from the dual solution $(\mu, \nu)$, and let $R$ be the corresponding user rate result calculated by (2.2). We have this result:

$$f_o(X, R) = \sum_{i,j} a_{ij} x_{ij} - \sum_j k_j \log(k_j)$$  \hspace{1cm} (2.26a)
\[= \sum_{i,j} a_{ij} x_{ij} - \sum_j k_j \log \left( e^{\mu_j - \nu} - 1 \right) - \sum_j k_j \log \left( \frac{k_j}{e^{\mu_j - \nu} - 1} \right) \quad (2.26b)\]

\[= \sum_i (a_{ij} - \mu_j) x_{ij} + \sum_j k_j + \sum_j \nu k_j - \sum_j k_j \log \left( \frac{k_j}{e^{\mu_j - \nu} - 1} \right) \quad (2.26c)\]

\[= \sum_i \max_j (a_{ij} - \mu_j) + K + \nu K - \sum_j k_j \log \left( \frac{k_j}{e^{\mu_j - \nu} - 1} \right) \quad (2.26d)\]

\[= \sum_i \max_j (a_{ij} - \mu_j) + \sum_j e^{\mu_j - \nu} - 1 + \nu K - \sum_j k_j \log \left( \frac{k_j}{e^{\mu_j - \nu} - 1} \right) \quad (2.26e)\]

\[= g(\mu, \nu) - \sum_j k_j \log \left( \frac{k_j}{e^{\mu_j - \nu} - 1} \right) \quad (2.26f)\]

where (2.15) accounts for the conversion of (2.26d) into (2.26e).

Let \((X^*, k^*)\) be the optimal solution to problem (2.5), and let \(R^*\) be the resulting user rate. By weak duality, it always holds that \(g(\mu, \nu) \geq f_o(X^*, R^*)\). Combining this result with (2.26f), we prove the claim:

\[f_o(X, R) \geq f_o(X^*, R^*) - \sum_j k_j \log \left( \frac{k_j}{e^{\mu_j - \nu} - 1} \right) \quad (2.27)\]

Note that whenever \(k_j = e^{\mu_j - \nu} - 1\) for a BS \(j\), as in Figure 2.3 (a), it does not contribute to the duality gap. When a BS is involved in ties, the duality gap is minimized when \(k_j\) is made as close to \(e^{\mu_j - \nu} - 1\) as possible. This duality gap also applies to the optimization results of the low-cost dual coordinate descent and the subgradient method.

### 2.6 OFDM Case

This section addresses the BS assignment problem in the OFDM case. We assume that the bandwidth \(W\) is divided into \(N\) tones (also known as subcarriers), and the channel varies from tone to tone due to frequency-selective fading. The transmit PSD is fixed and frequency-flat across the bandwidth at each BS. Let \(h_{ij}^{n} \in \mathbb{C}\) be the channel between user \(i\) and BS \(j\) in tone \(n\). If tone \(n\) of BS \(j\) is allocated to user \(i\), its data rate in this
tone is calculated by

$$R_{ij}^n = \frac{W}{N} \log \left( 1 + \frac{|h_{ij}^n|^2 p_j}{\Gamma (\sum_{j' \neq j} |h_{ij'}^n|^2 p_{j'} + \sigma_i^2)} \right)$$  \hspace{1cm} (2.28)

Let $d_{ij}^n$ be the fraction of tone $n$ of BS $j$ allocated to user $i$ in the long run. The long-term average rate of user $i$ is (if user $i$ is associated with BS $j$)

$$R_{ij} = \sum_n d_{ij}^n R_{ij}^n$$  \hspace{1cm} (2.29)

The BS assignment in the OFDM case can then be written as:

$$\begin{align*}
\text{maximize} & \quad \sum_{i,j} x_{ij} \log \left( R_{ij} \right) \\
\text{subject to} & \quad \sum_j x_{ij} = 1, \forall i \\
& \quad \sum_i d_{ij}^n = 1, \forall j, \forall n \\
& \quad 0 \leq d_{ij}^n \leq x_{ij}, \forall i, \forall j, \forall n \\
& \quad x_{ij} \in \{0, 1\}, \forall i, \forall j
\end{align*}$$  \hspace{1cm} (2.30a-e)

where constraint (2.30d) states that user $i$ can share the tone resource of BS $j$ only if it is associated with this BS. Note that if all the tones between the user and the BS have the same channel parameter (i.e., flat-fading case), the solution of $d_{ij}^n \ast$ is simply $1/k_j$ (where $k_j = \sum_i x_{ij}$), as mentioned previously in Section 2.1.

### 2.6.1 Averaging Dual Coordinate Descent

This method extends the dual coordinate descent to the OFDM case by averaging the estimated tone rates across the bandwidth. This average estimated tone rate is computed by

$$\bar{R}_{ij}^{\text{tone}} = \frac{1}{N} \sum_n R_{ij}^n$$  \hspace{1cm} (2.31)

We then just consider a flat-fading model with the spectral efficiency of $\left(N \bar{R}_{ij}^{\text{tone}}\right)/W$, in which the BS assignment can be determined by the dual coordinate descent. This approach is stated in Algorithm 3.
Algorithm 3 Averaging Dual Coordinate Descent

1) Compute the average tone rate $R_{ij}^{\text{tone}}$ by (2.31), $\forall i, \forall j$;
2) Compute $a_{ij} = \log \left( \frac{N R_{ij}^{\text{tone}}}{R_{ij}} \right)$, $\forall i, \forall j$.
3) Implement the dual coordinate descent (see Algorithm 1) with the $a_{ij}$’s in step 2.

We next need to solve the optimal $d_{ij}^n$ in problem (2.30) for the $x_{ij}$ given by the averaging dual coordinate descent. Although numerical method (e.g., Newton’s method) can be applied to draw the solution of $d_{ij}^n$, we suggest another approach via proportionally fair tone scheduling, which is a part of work in the next subsection.

Claim 2. The difference between the objective function $f_o(X, R)$ optimized by the averaging dual coordinate descent and the global optimum is bounded by

$$
\sum_i \max_j \left( \log \left( \frac{b R_{ij}}{R_{ij}^{\text{tone}}} \right) \right) + \sum_j k_j \log \left( \frac{k_j}{e^{k_j} - 1} \right),
$$

where $R(D)$ is defined in (2.29); variable $k_j$ equals $\sum_i x_{ij}$; $\epsilon$ is the error caused by the dual coordinate descent. The inequality between (2.32b) and (2.32c) derives from the error
bound result in Section 2.5 in the assumed flat-fading model whose spectral efficiency is
\[ \log \left( N \tilde{R}_{ij}^{\text{one}} \right). \]

It has been shown in Section 2.5 that \( \epsilon \leq \sum_j k_j \log (k_j / e^{\mu_j - \nu - 1}) \), therefore the overall error bound of the averaging dual coordinate descent is:
\[ \sum_i \max_j \left( \log \left( \tilde{R}_{ij} / \tilde{R}_{ij}^{\text{one}} \right) \right) + \sum_j k_j \log (k_j / e^{\mu_j - \nu - 1}). \]

Note that if the channels are flat-fading across the bandwidth (hence \( \tilde{R}_{ij} = \tilde{R}_{ij}^{\text{one}} \)), the error bound is then just \( \sum_j k_j \log (k_j / e^{\mu_j - \nu - 1}) \), which is exactly the error bound result in Section 2.5 for the flat-fading case. From Claim 2, we learn that the error bound of the averaging dual coordinate descent becomes loose when the tone variance is intensified.

### 2.6.2 Weighted-rate Relaxation Method

This subsection proposes a new method addressing the BS assignment problem from a weighted-rate maximization perspective. As shown in [27], the log-utility objective and the weighted-rate objective are equivalent in the long term as long as the weight of each user is the reciprocal of its long-term average rate. We can formulate the weighted-rate maximization problem as

\[
\max_{\mathbf{x}, \mathbf{y}} \sum_{i,j,n} \alpha_i R_{ij}^n x_{ij} y_{ij}^n \tag{2.33a}
\]

subject to
\[
\sum_j x_{ij} = 1, \quad \forall i \tag{2.33b}
\]

\[
\sum_i y_{ij}^n = 1, \quad \forall j, \forall n \tag{2.33c}
\]

\[
y_{ij}^n \leq x_{ij}, \quad \forall i, \forall j, \forall n \tag{2.33d}
\]

\[
x_{ij}, y_{ij}^n \in \{0,1\}, \quad \forall i, \forall j, \forall n \tag{2.33e}
\]

where \( \alpha_i \) is the weight of user \( i \); \( y_{ij}^n \) being 0 or 1 indicates whether or not tone \( n \) of BS \( j \) is allocated to user \( i \). Constraint (2.33c) states that each tone can be allocated to only user within the same time-slot; constraint (2.33d) states that user \( i \) can be allocated with the tone \( n \) of BS \( j \) only if it is associated with the BS, i.e., \( y_{ij}^n \) must be 0 if \( x_{ij} \) is 0. We
need to remark that \( x_{ij} \) in this problem formulation is actually redundant because it is dependent on \( y_{nij} \). Variable \( x_{ij} \) is reserved for the ease of following discussion. In order to realize proportional fairness, the weight \( \alpha_i \) should be the reciprocal of the long-term average rate \( \overline{R}_i \), but \( \overline{R}_i \) is unknown beforehand. In practice, \( \overline{R}_i \) is updated every time-slot by

\[
\overline{R}_i^{(t+1)} = \varphi \overline{R}_i^{(t)} + (1 - \varphi) R_{ij}^{(t)}, \quad 0 < \varphi < 1
\]  

(2.34)

where \( t \) is the time index; \( R_{ij}^{(t)} \) is calculated by (if user \( i \) is associated with BS \( j \) in the current time-slot)

\[
R_{ij} = \sum_{i,j,n} x_{ij} y_{nij}^n R_{ij}^n.
\]  

(2.35)

Here \( R_{ij} \) is interpreted as the instantaneous rate, because the weight \( \alpha_i \) is updated every time-slot and accordingly the solution of \( (X, Y) \) is also changed every time-slot.

For problem (2.33), the solution of \( Y \) can be easily obtained as

\[
y_{nij}^* = \begin{cases} 
1, & \text{if } i = i^{(j,n)} \\
0, & \text{otherwise} 
\end{cases}
\]  

where \( i^{(j,n)} = \arg \max_{i'} \left( \alpha_{i'} R_{n i'j}^n x_{i'j} \right) \)  

(2.36)

Note that possibly more than one user satisfy the condition of having \( y_{nij}^* \) of 1 in (2.36). To resolve these ties, \( y_{nij}^n \) can be assigned value 1 for any of these users without affecting the optimal objective function value. This tone-allocation procedure defined by (2.36) is referred to as the proportionally fair tone scheduling [28], which is also adopted in the averaging dual coordinate descent to determine the resource allocation in each cell.

We continue to solve for \( X^* \) in problem (2.33). One contribution in this part of work is that variable \( Y \) can be eliminated from problem (2.33).

**Claim 3.** Variable \( Y \) in problem (2.33) can be replaced by a function of \( X \) as

\[
y_{nij}^n = \lim_{z \to \infty} \frac{\left( \alpha_i R_{ij}^n \right)^z x_{ij}}{\sum_{i'} \left( \alpha_i R_{i'j}^n \right)^z x_{i'j}}
\]  

(2.37)

**Proof.** We first consider the case that the following condition is satisfied:

\[
|B^n_j| = 1, \quad \text{where } B^n_j = \{i = \arg \max_{i'} (\alpha_{i'} R_{n i'j}^n x_{i'j})\}
\]  

(2.38)
Then the value of the left-hand side of (2.37) equals the result in (2.36), i.e., the solution of \( Y \). When \( X^* \) is available, \( Y^* \) can be obtained by (2.37). Hence, Claim 3 is proved for this special case.

We then consider a more general situation where \( |B^n_j| = l^n_j \in \mathbb{Z}^+ \). In this case, \( y^n_{ij} = 1/l^n_j \), \( \forall i \in B^n_j \) according to (2.37), which does not satisfy constraint (2.33e), hence the solution of \( Y^* \) cannot be obtained by (2.37). Our approach is to pick one user (e.g., user \( i \)) in \( B^n_j \) and set its \( y^n_{ij} \) to 1, while setting \( y^n_{i'j} = 0 \), \( \forall i' \neq i \). It can be observed that this approach is exactly how we resolve ties for (2.36). Let \( y^n_{ij}' \) be the result given by (2.38), and let \( y^n_{ij}^{\ast} \) be the solution given by our approach. It can be shown that:

\[
\sum_{i,j,n} \alpha_i R^n_{ij} x_{ij} y^n_{ij}' = \sum_{i,j,n} \alpha_i R^n_{ij} x_{ij} y^n_{ij}^{\ast}
\]

Therefore, substituting (2.37) into problem (2.33) does not effect the solution of \( X \). When \( X^* \) is obtained, \( Y^* \) can be found by (2.36).

According to Claim 3, the objective function (2.33a) is equivalent to

\[
\max_X \lim_{z \to \infty} \sum_{i,j,n} \left( \frac{\alpha_i R^n_{ij}}{\sum_{i'} \left( \frac{\alpha_{i'} R^n_{i'j}}{z} \right)^2} \right)^{z+1} x_{ij} \sum_{i'} \left( \frac{\alpha_{i'} R^n_{i'j}}{z} \right)^z x_{i'j}
\]

Because \( x_{ij} \) is binary, (2.40) can be simplified to

\[
\max_X \lim_{z \to \infty} \sum_{i,j,n} \left( \frac{\alpha_i R^n_{ij}}{\sum_{i'} \left( \frac{\alpha_{i'} R^n_{i'j}}{z} \right)^2} \right)^{z+1} x_{ij} \sum_{i'} \left( \frac{\alpha_{i'} R^n_{i'j}}{z} \right)^z x_{i'j}
\]

We suggest using the relaxation heuristic to solve problem (2.33) with the objective function replaced by (2.41). After \( x_{ij} \)'s are relaxed to real numbers, the relaxation problem can be written as:

\[
\max_X \lim_{z \to \infty} \sum_{i,j,n} \left( \frac{\alpha_i R^n_{ij}}{\sum_{i'} \left( \frac{\alpha_{i'} R^n_{i'j}}{z} \right)^2} \right)^{z+1} x_{ij} \sum_{i'} \left( \frac{\alpha_{i'} R^n_{i'j}}{z} \right)^z x_{i'j}
\]

subject to

\[
\sum_j x_{ij} = 1, \ \forall i
\]

\[
0 \leq x_{ij} \leq 1, \ \forall i, \forall j
\]

After the above problem is solved, we round its solution \( X^* \) to binary values, and then recover \( Y^* \) according to (2.36).
2.7 Numerical Evaluation

The simulation experiments of this section are in a wrap-around downlink SISO HetNet which contains 7 macro BSs, 21 pico BSs and 210 user terminals as shown in Figure 2.7. The system parameters are listed in Table 2.2. The transmit PSD is fixed at the peak value for each BS. For simplicity, we abbreviate “the dual coordinate decent” to “DCD” in the legends of plots.

Figure 2.8 compares the convergence behavior of the dual coordinate descent with that of the adaptive subgradient method. Here each iteration refers to either a single update of $\mu_j$ in (low-cost) dual coordinate descent or a subgradient update of all $\mu_j$’s. We see that dual coordinate descent converges to within $10^{-1}$ of the optimum with only two rounds of iterations per BS (i.e., 56 iterations), while the convergence of subgradient method is very sensitive to its parameters. Here, we set $\rho = 1.2$, $\beta = 0.9$, and $\delta = 0.002$ in the adaptive subgradient method [26] and see that different settings of $\delta_1$ and $\gamma_k$ can result in quite different convergence behaviors. Note that in Figure 2.8 the dual coordinate descent algorithm does not converge to the optimum. This is due to the fact that it is possible for coordinate descent to get stuck in a suboptimal point. This gap is quite small according to the simulation result, however. It is also observed that the penalty of using the low-cost dual coordinate descent is that the dual convergence is not as good as the other proposed approaches, but it brings the benefit of low communication cost.

Figure 2.9 displays the cumulative distribution of data rates after 56 iterations for the various BS assignment algorithms. We see that the subgradient method, the dual coordinate descent and the low-cost dual coordinate descent all offer substantial rate improvement to low-rate users (who are normally edge users) as compared to the max-SINR BS assignment rule. For instance, the 50th-percentile rate is increased by about 33%, which is a consequence of off-loading traffic from the macro BSs to the pico BSs.

Figure 2.10 shows the cumulative distribution of user rates for the various algorithms
in the OFDM case. We divide the total bandwidth into 64 tones, and use Channel B of the vehicular test environment (see Table 2.3) in [29] to simulate the frequency selectiveness. For simplicity, constraint (2.42b) is ignored when numerical approach is used to solve problem (2.42). In the simulation, \( z \) being 10 is found large enough to improve the performance of the weighted-rate relaxation method, i.e., setting greater value for \( z \) can hardly improve the total utility. In the plot, it is observed that the averaging dual coordinate descent and the weighted-rate relaxation method achieve almost the same performance, both offering substantial rate improvement as compared to the max-SINR rule. However, the weighted-rate relaxation method has two defects. First, it cannot be implemented in distributed fashion. Second, lots of BS switching may occur in this method because its BS assignment strategy updates every other time-slot. In contrast, the averaging dual coordinate descent does not have these problems, which makes it preferable in practice.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel Bandwidth</td>
<td>10 MHz</td>
</tr>
<tr>
<td>Frequency Reuse Factor</td>
<td>1</td>
</tr>
<tr>
<td>Duplex Mode</td>
<td>TDD</td>
</tr>
<tr>
<td>Macro BS Peak Transmit PSD</td>
<td>-27 dBm/Hz</td>
</tr>
<tr>
<td>Pico BS Peak Transmit PSD</td>
<td>-47 dBm/Hz</td>
</tr>
<tr>
<td>Antenna Gain</td>
<td>15 dBi</td>
</tr>
<tr>
<td>SNR Gap</td>
<td>0 dB</td>
</tr>
<tr>
<td>Background Noise</td>
<td>-169 dBm/Hz</td>
</tr>
<tr>
<td>Distance-dependent Attenuation</td>
<td>128.1 + 37.6 \log_{10}(d), \ d \text{ is in km}</td>
</tr>
<tr>
<td>Shadowing</td>
<td>Log normal as ( \mathcal{N}(0, 8^2) )</td>
</tr>
</tbody>
</table>

Table 2.2: System parameters
Figure 2.7: A wrap-around HetNet topology

Figure 2.8: Dual error convergence: the dual coordinate descent vs the subgradient method
Chapter 2. Base Station Assignment

Figure 2.9: User rate CDF of the various BS assignment methods after 56 iterations

<table>
<thead>
<tr>
<th>Tap</th>
<th>Relative delay (ns)</th>
<th>Average power (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>-2.5</td>
</tr>
<tr>
<td>2</td>
<td>300</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>8900</td>
<td>-12.8</td>
</tr>
<tr>
<td>4</td>
<td>12900</td>
<td>-10.0</td>
</tr>
<tr>
<td>5</td>
<td>17100</td>
<td>-25.2</td>
</tr>
<tr>
<td>6</td>
<td>20000</td>
<td>-16.0</td>
</tr>
</tbody>
</table>

Table 2.3: Multi-path environment which generates frequency-selectiveness
Figure 2.10: User rate CDF of the different BS assignment algorithms in OFDM environment
Chapter 3

Joint Base Station Assignment & Power Control

This chapter addresses the joint BS assignment and power control problem. Load imbalance in the HetNet is usually caused by the situation that the transmit power of femto/pico BSs is much lower than that of macro BSs. Transmit power setting can also determine the interference in the network. Therefore, power control plays a key role in the network optimization. We tackle the power control problem jointly with BS assignment in this chapter.

For the joint BS assignment and power control problem, two novel approaches are proposed in this thesis. These two methods both combine the dual coordinate descent and the power control method, but reflect two divergent ways of combination. One of them optimizes the power and the BS assignment iteratively, while the other optimizes the power based on the duality theory.

Early works such as [11, 12, 30] reveal that the joint optimization of BS assignment and transmit power improves the network performance significantly. Some works such as [31] and [32] specialize in the problem for the code division multiple access (CDMA) system. The contribution on the joint optimization in a recent work [33] is based on Benders’
decomposition [34]. However, all these works adopt the power-based problem formulation, in which the objective is to minimize the total transmit power while ensuring that the minimum SINR constraint is satisfied for each user, and therefore their optimization results do not account for load balancing. In contrast, many recent works investigate the joint optimization from a rate-based viewpoint, as we do in this thesis. For example, [35] finds the BS assignment and power settings that maximize the sum throughput of the network, but only under certain restricted conditions and only for the case where the number of users and the number of BSs are the same. The algorithm proposed in [10] is based on so-called “two-sided scalable function” idea, which is shown to improve the total throughput. The approach in [36] is basically a kind of relaxation heuristic.

This chapter is organized as follows. Section 3.1 formulates the joint BS assignment and power control problem in a downlink SISO network with flat-fading channels. Section 3.2 studies the power control problem when the BS assignment is fixed. Section 3.3 shows that combining power control with the max-SINR BS assignment does not account for load balancing. Section 3.4 presents the outside incorporation method, and Section 3.5 presents the inside incorporation method, which are two innovations in this thesis. Section 3.6 compares the performance of the proposed methods.

### 3.1 Problem Formulation

The network model in this chapter is the same as in Section 2.1 except that the transmit power here can be changed. Using the same variable symbols as in last chapter, we
formulate the joint BS assignment and power control problem as:

\[
\begin{align*}
\text{maximize} & \quad \sum_{i,j} x_{ij} \log (R_{ij}(k_j, p)) \\
\text{subject to} & \quad 0 \leq p_j \leq \bar{p}_j, \ \forall j \\
& \quad \sum_j x_{ij} = 1, \ \forall i \\
& \quad \sum_i x_{ij} = k_j, \ \forall j \\
& \quad \sum_j k_j = K \\
& \quad x_{ij} \in \{0, 1\}, \ \forall i, \forall j
\end{align*}
\] (3.1a)

where \( R_{ij}(k_j, p) \) is computed by (2.2); \( \bar{p}_j \) is the peak PSD constraint of BS \( j \).

### 3.2 Power Control

In this section, we focus on the optimization of \( p \) in problem (3.1) with the BS assignment fixed. This power control problem can be written as:

\[
\begin{align*}
\text{maximize} & \quad \sum_{i,j} x_{ij} \log (R_{ij}(k_j, p)) \\
\text{subject to} & \quad 0 \leq p_j \leq \bar{p}_j, \ \forall j
\end{align*}
\] (3.2a)

where \( x_{ij} \) is given beforehand, and \( k_j \) is determined by (3.1d). We use Newton’s method to arrive a locally optimal solution to this problem. Let \( f_{\text{power}}(p) \) be the object function (3.2a), and introduce parameter \( r_{ij} \) as

\[
r_{ij} = \log(1 + \text{SINR}_{ij})
\] (3.3)

We then can write the first-order and the second-order partial derivatives of \( f_{\text{power}}(p) \) with respect to \( p_j \) as:

\[
\begin{align*}
\partial_{p_j} f_{\text{power}} &= \sum_i \frac{\text{SINR}_{ij} x_{ij}}{r_{ij}(1 + \text{SINR}_{ij})} - \sum_{i,j' \neq j} \frac{\Gamma |h_{ij'}|^2 \text{SINR}^2_{ij'}}{|h_{ij'}|^2 r_{ij'} (1 + \text{SINR}_{ij'})} x_{ij'} p_j.
\end{align*}
\] (3.4)
and
\[
\partial_{p_j, f_{\text{power}}} = -\sum_i \left( \frac{1}{r_{ij}^2} + \frac{1}{r_j} \right) \frac{\text{SINR}_{ij}^2}{(1 + \text{SINR}_{ij})^2} \frac{x_{ij}}{p_j^2} \\
+ \sum_{i,j \neq j} \frac{\Gamma^2 |h_{ij}|^4 \text{SINR}_{ij}^3 (2r_{ij'} + \text{SINR}_{ij'}(r_{ij} - 1)) x_{ij'}}{|h_{ij'}|^4 r_{ij'}^2 (1 + \text{SINR}_{ij'})^2} p_j^2.
\]
(3.5)

Following the heuristic in [28], we only use the diagonal entries of Hessian matrix in Newton’s method so as to reduce the computational complexity. Then Newton step is computed by:
\[
\Delta p_j = \frac{\partial_{p_j, f_{\text{power}}}}{\partial_{p_{ij}, f_{\text{power}}}}.
\]
(3.6)

Newton’s method updates \( p_j \)'s through:
\[
p_{j(t+1)} = \left[ p_{j(t)} + \alpha_{nt} \Delta p_{j} \right]_{0}.
\]
(3.7)

where \( \alpha_{nt} \) is the step size, which can be determined by backtracking line search [37].

The above formula indicates that: 1) \( p_{j(t+1)} = \bar{p}_{j} \) if \( p_{j(t)} + \alpha_{nt} \Delta p_{j} > \bar{p}_{j} \); 2) \( p_{j(t+1)} = 0 \) if \( p_{j(t)} + \alpha_{nt} \Delta p_{j} < 0 \).

### 3.3 Max-SINR Rule plus Power Control

This section analyzes an unsuccessful approach, in order to show that combining power control with the max-SINR BS assignment cannot achieve proportional fairness.

With the power control achieved in last section, a naive idea is that the power control method should be implemented iteratively with the conventional BS assignment scheme (i.e., the max-SINR rule), which gives rise to the iterative max-SINR and power control method. However, in the later part of this chapter our simulation results show that this iterative approach actually fails to improve the performance of the network.

The following simple instance gives an intuitive explanation for the failure of the iterative max-SINR rule and power control method. Consider a two-BS scenario with the initial BS assignment strategy as displayed in Figure 3.1. During the next power
control phase, BS A has its transmit PSD level raised due to the fact that it serves the majority of users, and in contrast BS B has its transmit PSD level decreased because most users suffer the interference from it. During the max-SINR assignment phase of a second round, those users who are already associated with BS A would stick with BS A even more because its transmit PSD level has been increased. The power control phase of the second round will increase BS A’s PSD level and decrease BS B’s PSD level further for the same reason in the last round, and so forth. This cycle cannot help off-load the traffic in cell A, and on the contrary it makes the load imbalance situation even worse.

![Figure 3.1: The initial BS assignment situation](image)

### 3.4 Outside Incorporation Method

The main contribution of this chapter of the thesis is the observation that if power control is done jointly with the BS assignment method proposed in Chapter 2, much better overall performance can be obtained. This method is referred to as the outside incorporation method. It is an intuitive approach where the BS assignment and the transmit power levels are optimized in an iterative fashion. When the transmit power is fixed in (3.1), we get the same BS assignment problem as in Chapter 2, to which the dual coordinate descent applies. When the BS assignment is fixed in (3.1), we get
the same power control problem as in Section 3.2, to which Newton’s method applies. This incorporation method is regarded as “outside” because transmit PSD levels are not optimized in the dual domain.

The outside incorporation method does not necessarily achieve the global optimum, but the convergence can be guaranteed. This is because, in this method, both the dual coordinate descent and Newton’s method for power control are implemented to raise the total log-utility, i.e., the objective function is nondecreasing during the iteration process of this approach. This approach is stated in Algorithm 4.

**Algorithm 4 Outside Incorporation Method**

**Initialization:** Set $p_j$’s to feasible values;

repeat

1) Implement the dual coordinate descent (see Algorithm 1);

repeat

2) Compute Newton step $\Delta p_j$ according to (3.6), $\forall j$;

3) Update $p_j$ according to (3.7), $\forall j$;

until Newton’s method converges;

until both the dual coordinate descent and Newton’s method converge.

---

### 3.5 Inside Incorporation Method

In contrast to the outside incorporation method, it is also possible to combine the power control method and the dual coordinate descent from the duality theory perspective. We call the resulting method the inside incorporation method. We first give the Lagrangian dual analysis to the joint optimization problem (3.1). The Lagrangian function and the
dual function can be written as:

$$L(X, k, p, \mu, \nu) = \sum_{i,j} x_{ij} \log (R_{ij}(k_j, p)) - \sum_j \mu_j \left( \sum_i x_{ij} - k_j \right) - \nu \left( \sum_j k_j - K \right)$$

(3.8)

and

$$g(\mu, \nu) = \begin{cases} \max_{X,k,p} L(X, k, p, \mu, \nu) \\ s.t. 0 \leq p_j \leq \bar{p}_j, \; \forall j \\ \sum_j x_{ij} = 1, \; \forall i \\ x_{ij} \in \{0,1\} \end{cases}$$

(3.9)

For the above problem (3.9), the solution of $k_j$ can be obtained analytically by (2.9), i.e., $k_j^* = e^{\mu_j - \nu - 1}$. But solving for $X^*$ and $p^*$ is hard because these two variables interact with each other. To approximate the solution of $X$ and $p$, we propose the idea of optimizing $X$ by (2.8) and optimizing $p$ by Newton’s method in iterative fashion. Obviously, the solution by this approach is usually suboptimal. In order to get the solution as close to the real one as possible, we choose multiple starting points of $(X, p)$ for this iterative approach.

The subgradient of $g(\cdot)$ with respect to $\mu$ cannot be obtained from the $g(\cdot)$ expression because the dual function in closed form is unknown (i.e., problem (3.9) cannot be solved analytically). However, we are able to show that $e^{\mu_j - \nu - 1} - |U_j|$ is the subderivative of $g(\cdot)$ with respect to $\mu_j$ in the following claim.

**Claim 4.** If $(X^*, k^*, p^*)$ is the global optimum solution to (3.9) for fixed $(\mu, \nu)$, then $e^{\mu_j - \nu - 1} - |U_j|$ is the subderivative of $g(\cdot)$ with respect to $\mu_j$ at $(\mu, \nu)$, where $|U_j|$ is defined in (2.17).

**Proof.** Let $c$ be the subderivative of $L(\cdot)$ with respect to $\mu_j$ at $(X^*, k^*, p^*, \mu, \nu)$. Define vector $\hat{\mu}$ as $(\mu_1, \cdots, \mu_{j-1}, \tilde{\mu}_j, \mu_{j+1}, \cdots, \mu_L)^T$, then we need to have:

$$L(X^*, k^*, p^*, \hat{\mu}, \nu) - L(X^*, k^*, p^*, \mu, \nu) \geq c(\hat{\mu}_j - \mu_j), \; \forall \hat{\mu}_j \in \mathbb{R}$$

(3.10)
Because \((X^*, k^*, p^*)\) is the solution to (3.9) when the dual variable is \((\mu, \nu)\), the above inequality can be rewritten as:

\[
L(X^*, k^*, p^*, \mu, \nu) - g(\mu, \nu) \geq c(\hat{\mu}_j - \mu_j), \quad \forall \mu_j \in \mathbb{R}
\]  

(3.11)

Let \((\hat{X}, \hat{k}, \hat{p})\) be the solution to (3.9) when the dual variable is \((\hat{\mu}, \nu)\), and then we have:

\[
g(\hat{\mu}, \nu) = L(\hat{X}, \hat{k}, \hat{p}, \hat{\mu}, \nu)
\geq L(X^*, k^*, p^*, \hat{\mu}, \nu), \quad \forall \hat{\mu}_j \in \mathbb{R}
\]  

(3.12)

(3.13)

Combining (3.11) and (3.13) together, we obtain:

\[
g(\hat{\mu}, \nu) - g(\mu, \nu) \geq c(\hat{\mu}_j - \mu_j), \quad \forall \hat{\mu}_j \in \mathbb{R}
\]  

(3.14)

Therefore, \(c\) is the subderivative of \(g(\cdot)\) with respect to \(\mu_j\) at \((\mu, \nu)\). □

![Figure 3.2: An example of Claim 5](image)

Figure 3.2 illustrates the proof of Claim 5 in a special case of \(g(\cdot)\) having just one variable \(\mu_j\). Let \((X^*, k^*, p^*)\) be the primal variable that maximizes the Lagrangian function...
$L(X, k, p, \mu_j)$ when $\mu_j = 8$, so $g(\mu_j)$ and $L(X^*, k^*, p^*, \mu_j)$ have the same value at $\mu_j = 8$. According to the duality theory, $L(X^*, k^*, p^*, \mu_j)$ is always less or equal to $g(\mu_j)$ at the same $\mu_j$, as shown in the figure. Therefore, for any subderivative of $L(X, k, p, \mu_j)$ with respect to $\mu_j$ at $\mu_j = 8$ (denoted by the arrow), we observe that it is also the subgradient of $g(\mu_j)$ with respect to $\mu_j$ at $\mu_j = 8$.

Because the dual function $g(\cdot)$ is convex, its subderivative with respect to $\mu_j$ is negative if $\mu_j < \mu_j^*$, and is positive if $\mu_j > \mu_j^*$. We suggest the bisection search for seeking $\mu_j^*$. We relegate the detailed description of the inside incorporation method to Algorithm 5.

Please note that the subderivative given by Claim 4 is true only if $(X^*, k^*, p^*)$ is the globally optimal solution to problem (3.9). However, as mentioned previously, the solution of $(X^*, k^*, p^*)$ can only be approximated by the iterative approach in practice, hence the coordinate descent update in the inside incorporation method is not rigorous and it is not guaranteed to achieve the optimal solution. Actually, we cannot even guarantee the convergence of the inside incorporation method. In practice, we try multiple starting points of $(X, p)$ to approximate the solution as close to the real one as possible, as shown in Algorithm 5.
Algorithm 5 Inside Incorporation Method

**Initialization:** Set $\mu_j = 0, \forall j$; set $\nu = \log \frac{\sum_j e^{\nu_j - 1}}{K}$; choose $\epsilon > 0$;

repeat

for each $j \in \{1, \cdots, L\}$ do

1) Initialize $a$ and $b$, e.g., $a = 0, b = \nu + 1 + \log K$;

repeat

2) $\mu_j \leftarrow \frac{a + b}{2}$;

3) Optimize $X$ and $p$ iteratively starting from multiple points (which can be generated randomly), and then update $(X^*, p^*)$ to the optimized variable with the maximum $L(\cdot)$;

4) Calculate the subderivative by $e^{\nu_j - 1} - |U_j|$;

if the subderivative is negative then

5) $a \leftarrow \mu_j$;

else if the subderivative is positive then

5') $b \leftarrow \mu_j$;

end if

until $b - a \leq \epsilon$, or the subderivative equals 0;

end for

6) $\nu \leftarrow \log \frac{\sum_j e^{\nu_j - 1}}{K}$

until $(\mu, \nu)$ converges;

7) Set user-BS association according to (2.8). Resolve ties if necessary.
3.6 Numerical Evaluation

The simulation experiment of this chapter is done in the flat-fading network model as described in Section 2.7 (the topology is displayed in Figure 2.7, and the system parameters are shown in Table 2.2).

![Figure 3.3: Data rate CDF of the various algorithms for the joint BS assignment and power control problem](image)

Figure 3.3 compares the cumulative distribution of data rates for the various methods proposed in this chapter. Ten starting points of $(X, p)$ are generated randomly for the inside incorporation method (see step 2 in Algorithm 5). As the figure shows, the iterative max-SINR and power control method does not address load imbalance effectively, whose performance is similar to that of the max-SINR rule under the peak transmit power. We also implement the max-SINR method under the transmit power optimized by the inside incorporation method, and it is almost as good as the inside incorporation method. We can also observe that the performance gap is quite narrow between the inside incor-
poration method and the outside incorporation method, but the outside incorporation method is more useful in practice due to the fact that it has easy implementation and is guaranteed to converge.
Chapter 4

Joint Base Station Assignment &
Beamforming

This chapter looks into the joint BS assignment, power control and beamforming problem. Because power control is included in beamforming, this joint optimization problem is simply referred to as the joint BS assignment and beamforming problem.

The core idea of our proposed method is to decouple the overall problem into two subproblems where the BS assignment and the beamforming vectors are optimized separately. This method has lower computational complexity and comparable performance as compared to the baseline algorithm. Besides, our proposed method does not cause the excessive handover problem.

In the existing literature, [38] addresses the joint BS assignment and beamforming problem from the total transmit power minimization perspective; [39] proposes a method based on convex-concave procedure [40]. Actually, there are many works on the beamforming problem with the BS assignment fixed, e.g., [1, 41–44], among which the WMMSE algorithm in [1] is of great interest because its sum weighted-rate maximization objective is equivalent to the proportional fairness objective under the certain condition. This WMMSE algorithm is modified in [15] to optimize the BS assignment and beamforming
vectors jointly.

This chapter is organized as follows. Section 4.1 formulates the joint BS assignment and beamforming problem for a downlink MIMO network. Section 4.2 introduces the WMMSE algorithm [1] which is designed to solve the beamforming problem with the BS assignment fixed. Section 4.3 proposes the two-stage method for the joint BS assignment and beamforming problem. This method is one of the main contributions of this thesis. Section 4.4 introduces a baseline algorithm in [15] for the joint BS assignment and beamforming problem, which is referred to as the approximate WMMSE method. Section 4.5 analyzes the computational complexity for all these algorithms. Section 4.6 presents the numerical results.

4.1 Problem Formulation

This chapter considers a downlink MIMO network model with flat-fading channels. Let $N_i$ be the number of antennas deployed at user $i$, and let $M_j$ be the number of antennas deployed at BS $j$. The channel between user $i$ and BS $j$ can be denoted by matrix $H_{ij} \in \mathbb{C}^{N_i \times M_j}$. The transmit PSD of each BS is assumed to be frequency-flat across the bandwidth $W$. The problem objective is to maximize the total log-utility on the long-term average rates in the network:

$$\max \sum_i \log (R_i) \quad (4.1)$$

Suppose that each user supports one data stream for communication. Let $v_{ij} \in \mathbb{C}^{M_j}$ be the current beamforming vector of BS $j$ intended for user $i$. In the current time-slot, the achievable rate of user $i$ who is associated with BS $j$ can be calculated by

$$R_{ij} = W \log \det \left( I_{N_i \times N_i} + H_{ij} v_{ij} v_{ij}^H H_{ij}^H \left( \sum_{(i',j') \neq (i,j)} H_{i'j'} v_{i'j'} v_{i'j'}^H H_{i'j'}^H + \sigma_i^2 I_{N_i \times N_i} \right)^{-1} \right) \quad (4.2)$$
where $\sigma_i^2$ is the PSD of AWGN for user $i$. We need to remark that here $R_{ij}$ is the instantaneous rate, because $v_{ij}$ can be changed every other time-slot. The long-term average rate $\bar{R}_i$ in (4.1) is updated exponentially by ($t$ is the time index)

$$\bar{R}_i^{(t+1)} = \varphi \bar{R}_i^{(t)} + (1 - \varphi) R_{ij}^{(t)}, \ 0 < \varphi < 1$$ (4.3)

As mentioned previously in Section 2.6.2, the sum log-utility objective and the sum weighted-rate objective are equivalent in the long term, as long as the weight of each user is the reciprocal of its long-term average rate. This sum weighted-rate maximization problem can be written as:

$$\max_{\mathbf{v}} \sum_{i,j} \alpha_i R_{ij}(\mathbf{v}) \quad (4.4a)$$

subject to

$$\sum_{i \in C_j} \|v_{ij}\|^2 \leq \bar{p}_j, \ \forall j \quad (4.4b)$$

$$\|v_{ij}\| \cdot \|v_{ij'}\| = 0, \ \forall i, \forall j, \forall j' \neq j \quad (4.4c)$$

where $\alpha_i = 1/\bar{R}_i$; $R_{ij}(\mathbf{v})$ is calculated by (4.2). In the above problem, constraint (4.4b) states that the total transmit PSD of BS $j$ cannot exceed the peak value $\bar{p}_j$; constraint (4.4c) states that each user can only be served by one BS at the same time (if user $i$ is served by BS $j$, $\|v_{ij}\| > 0$).

### 4.2 Beamforming Problem

In Section 4.2 and 4.3, we first describe an approach in the existing literature for solving problem (4.4). We begin by considering the beamforming problem when the BS assignment is fixed. Defining $C_j = \{i | x_{ij} = 1\}$ for each BS $j$, we formulate the beamforming problem with predefined $X$ as:

$$\max_{\mathbf{v}} \sum_j \sum_{i \in C_j} \alpha_i R_{ij}(\mathbf{v}) \quad (4.5a)$$

subject to

$$\sum_{i \in C_j} \|v_{ij}\|^2 \leq \bar{p}_j, \ \forall j \quad (4.5b)$$
As proposed in [44] and [1], the above problem is equivalent to the mean-square error (MSE) optimization problem:

\[
\begin{align*}
\text{minimize} & \quad \sum_{j} \sum_{i \in C_j} \alpha_i (w_{ij} e_{ij} - \log w_{ij}) \\
\text{subject to} & \quad \sum_{i \in C_j} \|v_{ij}\|^2 \leq p_j, \quad \forall j
\end{align*}
\]

where \(u_{ij}\) is the receive vector of user \(i\) intended for BS \(j\); \(w_{ij}\) is a newly introduced variable interpreted as the weight in the MSE problem; \(e_{ij}\) is the MSE of user \(i\) calculated by (if user \(i\) is associated with BS \(j\))

\[
e_{ij} = |1 - u_{ij}^H H_{ij} v_{ij}|^2 + \sum_{(i',j') \neq (i,j)} |u_{ij}^H H_{ij'} v_{i'j'}|^2 + \sigma_i^2 \|u_{ij}\|^2
\]

The WMMSE algorithm solves problem (4.6) through the coordinate descent. With \(w_{ij}\) and \(v_{ij}\) fixed, the optimal \(u_{ij}\) is just the solution to \(\nabla u_{ij} \sum_{j} \sum_{i \in C_j} \alpha_i R_{ij}(v) = 0\), which is:

\[
u_{ij}^* = \left( \sum_{j'} \sum_{i' \in C_{j'}} H_{ij'} v_{i'j'} H_{ij'}^H + \sigma_i^2 I_{N_i \times N_i} \right)^{-1} H_{ij} v_{ij}
\]

This \(u_{ij}^*\) result is also known as the minimum mean-square error (MMSE) receiver. Likewise, with \(u_{ij}\) and \(v_{ij}\) fixed, the optimal \(w_{ij}\) is

\[
w_{ij}^* = \frac{1}{e_{ij}}
\]

Specifically, when the \(u_{ij}\) in \(E_{ij}\) is the MMSE receiver (4.8), \(w_{ij}^*\) can be simplified to

\[
w_{ij}^* = \frac{1}{1 - u_{ij}^H H_{ij} v_{ij}}
\]

With \(u_{ij}\) and \(w_{ij}\) fixed, the optimal \(v_{ij}\) is

\[
v_{ij}^*(\lambda_j^*) = \alpha_i \left( \sum_{j'} \sum_{i' \in C_{j'}} \alpha_{i'} w_{i'j'} H_{ij'}^H u_{i'j'} u_{ij}^H H_{ij'} + \lambda_j^* I_{M_j \times M_j} \right)^{-1} H_{ij}^H u_{ij} w_{ij}
\]

where \(\lambda_j^*\) is the optimal dual variable for constraint (4.6b). The following formulation shows how to obtain \(\lambda_j^*\). According to duality theory, \(\lambda_j^*\) should satisfy the complementary
slackness \[37\]:

\[
\lambda_j^* \left( \sum_{i \in C_j} \|v_{ij}(\lambda_j)\|^2 - \bar{p}_j \right) = 0 \tag{4.12}
\]

Therefore, if the matrix \(\sum_{i' \neq j} \alpha_{i'j'} w_{i'j'} H_{i'j'} u_{i'j'} H_{i'j'}^H u_{i'j'} H_{i'j'}^H\) is invertible and \(\sum_{i \in C_j} \|v_{ij}(0)\|^2 \leq \bar{p}_j\), we have \(\lambda_j^* = 0\). Otherwise, \(\lambda_j^*\) is the solution to the equation below:

\[
\sum_{i \in C_j} \|v_{ij}(\lambda_j)\|^2 = \bar{p}_j \tag{4.13}
\]

The left-hand side of equation (4.13) is a decreasing function of \(\lambda_j\), so the solution of \(\lambda_j^*\) can be easily found through a bisection search. The WMMSE algorithm is described in Algorithm 6.

**Algorithm 6 WMMSE Algorithm**

**Initialization:** Initialize \(v_{ij}\)'s to feasible values; choose \(\epsilon > 0\);

repeat

1) Update \(u_{ij}\) according to (4.8), \(\forall j, \forall i \in C_j\);

2) Update \(w_{ij}\) according to (4.10), \(\forall j, \forall i \in C_j\);

3) Update \(v_{ij}\) according to (4.11), \(\forall j, \forall i \in C_j\);

until \(|\sum_{i,j} \log w_{ij} - \sum_{i,j} \log w_{ij}'| \leq \epsilon\).

### 4.3 Approximate WMMSE Method

The approximate WMMSE method is proposed in [15] to address the joint BS assignment and beamforming problem. The core idea is to incorporate the BS assignment optimization into the WMMSE algorithm [24] by adding penalty terms to the MSE values. In the approximate WMMSE method, the initial \(v_{ij}\) does not have to meet with constraint (4.4c), but the feasible solution of \(v_{ij}\) can be obtained in the end. The MSE value with the penalty addition is written as:

\[
e_{ij}^c = e_{ij} + c \left( \sum_{j' \neq j} \|v_{ij'}\|^2 \|u_{ij'}\|^2 \right) \tag{4.14}
\]
where \( c \) is a positive factor.

Here is the explanation for the MSE penalty term \( c \left( \sum_{j' \neq j} \|v_{ij'}\|^2 \|u_{ij}\|^2 \right) \) in the formula above: if user \( i \) is served by at most one BS (i.e., constraint (4.4c) is satisfied), the penalty term equals 0; otherwise, the penalty term is positive. Hence, when \( c \) approaches infinity, the situation that a user is served by multiple BSs will not happen otherwise the objective function also approaches infinity. This is how the BS assignment and the beamforming vectors are optimized jointly. The MSE problem with the penalty terms can be written as:

\[
\begin{align*}
\text{minimize} & \quad \sum_{i,j} \alpha_i \left( w_{ij}c_i - \log w_{ij} \right) \\
\text{subject to} & \quad \sum_{i,j} \|v_{ij}\|^2 \leq \bar{p}_j, \forall j
\end{align*}
\]

The approach to the above problem is similar to the WMMSE algorithm. Optimal \( u_{ij}, v_{ij} \) and \( w_{ij} \) are updated iteratively by taking the partial gradient and setting it to zero\(^1\). Specifically, with \( v_{ij} \) and \( w_{ij} \) fixed, the optimal \( u_{ij} \) is

\[
u_{ij}^* = \left( \sum_{i',j'} H_{ij'}v_{ij'}H_{ij'}^H + \sigma_i^2 I_{N_i \times N_i} + c \sum_{j' \neq j} \|v_{ij'}\|^2 I_{N_i \times N_i} \right)^{-1} H_{ij}v_{ij}
\]

With \( u_{ij} \) and \( v_{ij} \) fixed, if the current \( u_{ij} \) satisfies (4.16), the optimal \( w_{ij} \) is

\[
w_{ij}^* = \frac{1}{1 - u_{ij}^H H_{ij} v_{ij}}
\]

With \( w_{ij} \) and \( u_{ij} \) fixed, the optimal \( v_{ij} \) is

\[
v_{ij}^* = \left( \sum_{i',j'} \alpha_i w_{i'j'} H_{i'j'}^H u_{i'j'} v_{i'j'} H_{i'j'} + \lambda_j^* I_{M_j \times M_j} + \alpha_i c \sum_{j' \neq j} w_{ij'} \|u_{ij'}\|^2 I_{M_j \times M_j} \right)^{-1} \alpha_i w_{ij} H_{ij}^H u_{ij}
\]

The way of determining the optimal dual variable \( \lambda_j^* \) is similar to that in the WMMSE algorithm. If the matrix \( \sum_{i',j'} \alpha_i w_{i'j'} H_{i'j'}^H u_{i'j'} v_{i'j'} H_{i'j'} + \alpha_i c \sum_{j' \neq j} w_{ij'} \|u_{ij'}\|^2 I_{M_j \times M_j} \) is invertible and \( \sum_{i \in U_j} \|v_{ij}\|^2 \leq \bar{p}_j, \lambda_j^* = 0 \). Otherwise, \( \lambda_j^* \) is the solution to (4.19), which

\(^1\)The updating formulas for \( u_{ij}, v_{ij} \) and \( w_{ij} \) in [15] are incorrect.
can be found via bisection search.

\[
\sum_{i,j} \|v_{ij}(\lambda_j)\|^2 = \bar{p}_j
\]  

(4.19)

The approximate WMMSE method is stated in Algorithm 7.

**Algorithm 7 Approximate WMMSE Method**

**Initialization:** Initialize \(c\) to a positive number; initialize \(v_{ij}\)'s to any value; choose \(\epsilon > 0, q > 1\);

repeat

1) \(w'_{ij} \leftarrow w_{ij}, \forall i, \forall j\);

2) Update \(u_{ij}\) according to (4.16), \(\forall i, \forall j\);

3) Update \(w_{ij}\) according to (4.17), \(\forall i, \forall j\);

4) Update \(v_{ij}\) according to (4.18), \(\forall i, \forall j\);

5) \(c \leftarrow qc\);

until \(|\sum_{i,j} \log w_{ij} - \sum_{i,j} \log w'_{ij}| \leq \epsilon\) and (4.4c) is satisfied;

Figure 4.1: Beamforming updates

The problems with the approximate WMMSE method are three fold. First, the algorithm has an extremely high computational complexity, because the updates of \(u_{ij}\), \(v_{ij}\) and \(w_{ij}\) need to be done between any BS and any user in the entire network, e.g., Figure 4.1 (b). In contrast, the WMMSE algorithm only requires the beamforming-vector
updates within each cell, e.g., Figure 4.1 (a). Second, its performance and convergence speed depend heavily on its parameter $c$ and $q$, but these parameters are hard to decide. Third, the approximate WMMSE method changes the BS assignment strategy every other time-slot, and thus generates a lot of BS switches (also known as handovers), which is unacceptable to the network system in practice.

### 4.4 Two-stage Method

The main contribution of this chapter of the thesis is an original two-stage method that addresses the BS assignment and beamforming problem jointly. As the name indicates, the two-stage method consists of two stages, which are respectively designed for the BS assignment and the beamforming optimization. The first stage adopts the outside incorporation method (see Section 3.5) to determine the BS assignment. The second stage optimizes the beamforming vectors using the WMMSE algorithm. Specifically, we schedule only limited number of users in the WMMSE algorithm in order to reduce the computational complexity.

#### 4.4.1 Stage One: BS Assignment in MIMO Network

The target of the first stage is to determine the BS assignment. We first need to evaluate the multiplexing gain, i.e., how much improvement can be achieved by the spatial multiplexing using beamforming. The maximum multiplexing gain can be represented by the degrees of freedom (DoF), as suggested in [45]. The DoF of user $i$ is defined as (if user $i$ is associated with BS $j$)

$$
\text{DoF}_{ij} = \lim_{\text{SNR}_{ij} \to \infty} \frac{R_{ij}}{\log(\text{SNR}_{ij})}
$$

(4.20)

The value of $\text{DoF}_{ij}$ is evaluated as $\min(M_j, K_j)$. Since there are usually many more users than the transmit antennas at the BS in a cell, the approximate value of $\text{DoF}_{ij}$ can be further simplified to $M_j$. 
Chapter 4. Joint Base Station Assignment & Beamforming

We assume a SISO network model where the channel $h_{ij}$ is evaluated based on the distance-dependent attenuation. SINR value can be calculated for this SISO model according to (2.1). Then the utility parameter $a_{ij}$ is redefined as:

$$a_{ij} = \log \left( M_j W \log \left( 1 + \frac{\text{SINR}_{ij}}{\Gamma} \right) \right) \quad (4.21)$$

The main part of stage one is to implement the outside incorporation method (see Section 3.5) taking the above redefined $a_{ij}$. We need to remark that the PSD variable $p_j$ in the outside incorporation method is just used to assist in the BS assignment decision, but is not adopted as the power control result. The actual transmit PSD levels are optimized in the next stage.

4.4.2 Stage Two: Beamforming Update

This stage updates the beamforming vectors every other time-slot based on the WMMSE algorithm. After the BS assignment is determined by stage one, we can formulate the beamforming problem exactly as (4.5), which is solved by the WMMSE algorithm.

In the WMMSE algorithm, the beamforming vectors are updated for all the users in each cell. However, because at most $M_j$ parallel transmissions are supported by BS $j$, no more than $M_j$ users in the cell can be served at the same time. We can take advantage of this fact to reduce the computational complexity of the WMMSE algorithm. Basically, we only choose those users who are most likely to get served after beamforming to take part in the WMMSE algorithm, while for the rest the transmit/receive vectors are fixed at the null vector. This procedure is referred to as the WMMSE scheduling. In our approach, the users are chosen according to the estimated weighted rate $\alpha_i R_{ij}$ from the highest to the lowest, where $R_{ij}$ is calculated by (2.2) in the SISO model assumed in stage one. The number of the users chosen by the WMMSE scheduling in the cell of BS $j$ is a parameter $S_j$ of the two-stage method, which should be greater than $M_j$. We relegate the description of the two-stage method to Algorithm 8.
Algorithm 8 Two-stage Method

**Initialization:** Choose $S_j \geq M_j, \forall j$;

1) Implement the outside incorporation method (see Algorithm 4) with the $a_{ij}$’s calculated by (4.21), and thereafter obtain the optimization result $(X, p)$; determine the BS assignment according to $X$;

**repeat**

2) Update $R_i$ according to (2.34), $\forall i$; $\alpha_i \leftarrow 1/R_i$, $\forall i$;

3) Choose $S_j$ users who are associated with BS $j$ according to $\alpha_i R_{ij}'$ from the highest to the lowest, $\forall j$, where $R_{ij}$ is calculated by (2.2) under the $p$ in step 1;

4) Implement the WMMSE algorithm [1] for the chosen users in each cell over the entire network;

**until** $R_i$ converges, $\forall i$;

Note that the BS assignment decision and the beamforming decision are made in different time scale in the two-stage method, as shown in Figure 4.2. The BS assignment decision is long-term so as to avoid the excessive handover problem.

![Figure 4.2: Flow chart of the two-stage method](image-url)
Chapter 4. Joint Base Station Assignment & Beamforming

4.5 Complexity Analysis

This section analyzes the computational complexity of the various algorithms. We focus on the part of algorithm which is implemented most frequently (e.g., every time-slot), hence stage one of the two-stage method can be ignored for the complexity analysis.

We generally assume that: 1) the matrix multiplication operation of $A^{m \times n}B^{n \times p}$ yields a complexity of $O(mnp)$; 2) the matrix inversion operation (by Gaussian elimination) of $(A^{m \times m})^{-1}$ yields a complexity of $O(m^3)$.

Recall that $K$ is the total number of users in the network, and $L$ is the total number of BSs. For simplicity, let $M$ and $N$ be the number of antennas at each BS and each user terminal respectively, and let $J = \sum_j S_j$, which is the total number of users chosen by the WMMSE scheduling (normally $J \ll K$). We skip the bisection search for $\lambda_j^*$ in the complexity analysis. It can be shown that the WMMSE algorithm yields a complexity of $O(K^2MN^2 + K^2M^2N + KM^3 + KN^3)$ for each round of iteration. Similarly, the two-stage method has a per-iteration complexity of $O(J^2MN^2 + J^2M^2N + JM^3 + JN^3)$.

Note that without WMMSE scheduling, the two-stage method yields the same complexity as the WMMSE algorithm. The per-iteration complexity of the approximate WMMSE method is the highest, which is $O(L^2K^2MN^2 + L^2K^2M^2N + LKM^3 + LKN^3)$.

4.6 Numerical Evaluation

We first compare the two-stage method with the max-SINR rule plus the WMMSE algorithm. The max-SINR rule is implemented under the peak transmit PSD levels. The network model is the same as before (see Section 2.1) except that: 1) each BS has 4 antennas and each user terminal has 2 antennas; 2) Rayleigh fading is introduced.

Let $S$ be the number of users scheduled in each cell (i.e., $S_j = S, \forall j$, see Algorithm 8). As shown in Figure 4.3, the overall user throughput have substantial improvement by the two-stage method as compared to the max-SINR BS assignment under the peak
transmit power, e.g., the 50th-percentile rate is almost doubled when $S = 8$. We can also observe that the performance of the two-stage method improves when $S$ is raised, but the improvement is small when $S$ is large enough. Therefore, the two-stage method can still have comparable performance even if the scheduling number is not large.

We then compare the cumulative distribution of data rates for the two-stage method and the approximate WMMSE method. Because the computational complexity of the approximate WMMSE method is extremely high, it is too time-consuming to test this method in the previous network model. Instead, we compare these two algorithms in a miniature model whose topology is shown in Figure 4.4, while the other settings (e.g., channel parameters, antenna numbers at each BS, etc.) are the same as that of the first network model in this section. In Figure 4.5, it is observed that the approximate WMMSE method has better performance than the two-stage method. However the gap is distinct for only a few high-rate users but rather small for the others, so the two-stage method has comparable performance as compared to the approximate WMMSE method.
Figure 4.3: User rate CDF: the two-stage method vs the max-SINR BS rule plus the WMMSE algorithm

Figure 4.4: A small HetNet with 3 macro BSs, 3 pico BSs, and 60 users
Figure 4.5: User rate CDF: the two-stage method vs the approximate WMMSE method
Chapter 5

Conclusions

Load balancing is crucial to heterogeneous networks, however the conventional max-
SINR BS assignment rule does not address it effectively. This thesis investigates BS
assignment optimization jointly with power control and beamforming, in order to achieve
load balancing from a total log-utility maximization perspective.

For a SISO network with flat-fading channels and fixed transmit power, the BS as-
signment problem can be solved in the dual domain where the dual variables perform
as the BS-specific prices. The approach in [24] is basically updating the dual variables
by the subgradient method, but its step size is hard to decide (which impacts the con-
vergence speed). Our proposed method updates the dual variables based on coordinate
descent, which is referred to as the dual coordinate descent algorithm. As compared to
the subgradient method, the dual coordinate descent is suitable for distributed imple-
mentation, and achieves near-optimal solution while having fast convergence speed. We
then develop a variant of the dual coordinate descent aimed to reduce the information
exchange between BSs and users, which is referred to as the low-cost dual coordinate
descent. As the numerical results indicate, the low-cost dual coordinate descent does not
have its performance as good as the dual coordinate descent, but it still has practical val-
ues in some cases for its advantage of requiring little communication overhead. Next, the
averaging dual coordinate descent is proposed to solve the BS assignment problem in the OFDM case. As compared to the baseline algorithm (i.e., the weighted-rate relaxation method), the averaging dual coordinate descent exhibits equally good performance, while being fast and suitable for distributed implementation (these advantages are inherited from the dual coordinate descent).

For the joint BS assignment and power control problem, this thesis offers two divergent approaches, both combining the dual coordinate descent and the existing power control method. The first one is to optimize the power and the BS assignment in iterative fashion, which is referred to as the outside incorporation method. The other approach, known as the inside incorporation method, incorporates the power update into the dual coordinate descent based on the duality theory. The problems with the inside incorporation method are two fold: 1) this algorithm is not guaranteed to converge; 2) the implementation is complicated. In contrast, these two problems do not exist in the outside incorporation method. Therefore, the outside incorporation method is more useful in practice, although the numerical results show that the inside incorporation method leads to slightly better performance.

For the joint BS assignment, power control and beamforming problem, this thesis proposes the two-stage method that combines the outside incorporation method and the WMMSE algorithm [1]. An important difference between the two-stage method and the baseline algorithm (i.e., the approximate WMMSE method [15]) is that the former gives the long-term BS assignment while the latter gives the BS assignment changing every other time-slot. For practical reasons, such frequent change of the BS assignment strategy given by the approximate WMMSE method is unacceptable to the network system. In addition, the two-stage method has a much lower computational complexity as compared to the approximate WMMSE method, and also has comparable performance as verified in the simulation experiment.
Bibliography


[41] F. Rashid-Rarrokhi, K. J. R. Liu, and L. Tassiulas, “Transmit beamforming and


